Debiasing Without Protected Attributes: Latent Concept Erasure from Textual Profiles

Anonymous ACL submission

Abstract

Most fairness research in NLP assumes direct access to protected attributes such as gender, race, or nationality. In practice, however, such information is often unavailable due to privacy constraints, missing metadata, or legal restrictions, even though models may infer it from indirect textual cues. This raises a key question: can debiasing succeed without direct access to sensitive attributes? We propose H-SAL, which performs post-hoc concept and attribute erasure using *self-description text* as an implicit debiasing signal. To support this setting, we introduce a multi-domain Stack Exchange-based fairness benchmark for helpfulness prediction that includes both *explicit* and *implicit* signals, enabling comparison between standard debiasing with protected labels and debiasing without access to sensitive information. Across encoder and decoder-only language models, we find that implicit self-description often matches or outperforms explicit-label-based debiasing. Our results broaden representation-level fairness research and provide a new benchmark for studying debiasing under realistic data constraints.

1 Introduction

Fairness research in natural language processing (NLP) has largely assumed that protected attributes are explicitly available. Under this setup, many methods use demographic labels such as gender, race, or nationality to debias models and measure fairness through group performance gaps. In practice, however, these attributes are often missing or unavailable due to privacy, legal, or data-collection constraints (Ashurst and Weller, 2023). Yet their absence from the database does not imply their absence from model representations: Neplenbroek et al. (2025) show that LLMs can infer demographics from stereotypical dialogue cues, encode them in latent representations, and let them shape responses even without explicit disclosure.

Input (X): *Q: What is the C# version of VB.NET's InputBox?*
A: You mean InputBox? Just look in the Microsoft.VisualBasic namespace. C# and VB.Net share a common library. If one language can use it...
Answer Helpfulness (Y): Answer score = 3
About Me ($\mathcal{H}$): *Director of Information Technology, York University Nebraska. ...The avatar is both because I play counter strike and a nod to lambda expressions in C#...*
Relevant Attributes (Z): Reputation = 418,535; Location = York, Nebraska, USA

Input (X): *Q: What is the closest alternative to "rubbish" in American English?*
A: First of all, the word rubbish is pretty well understood by American English speakers, and although it does have a British flavor it is used occasionally by Americans...
Answer Helpfulness (Y): Answer score = 8
About Me ($\mathcal{H}$): *I have a degree in linguistics and am partial to descriptivist approaches to usage...I am a native speaker of American English ...*
Relevant Attributes (Z): Reputation = 69,358; Location = San Jose, CA, United States

Figure 1: **Self-description cues used for textual erasure.** Real examples from Stack Exchange user profiles, where $\mathcal{H}$ denotes the user's self-description. The highlighted spans mark words or phrases that may implicitly reveal attributes such as education, reputation or location, and are therefore used to learn directions for textual erasure.

A prominent line of work addresses bias through post-hoc representation erasure under explicit supervision. Methods such as INLP (Ravfogel et al., 2020), SAL (Shao et al., 2023b), RLACE (Ravfogel et al., 2022a,b), and LEACE (Belrose et al., 2023) remove demographic information while largely preserving task utility. However, they rely on direct access to demographic labels, which are often unavailable or unreliable in deployment. To relax this requirement, Shao et al. (2023a) propose

zero-shot debiasing by alternating between inferring attribute assignments and removing directions correlated with bias, while Shao et al. (2025) show that cross-lingual transfer can identify shared bias subspaces for low-resource settings.

These works raise a broader question: can concept and attribute erasure succeed using only implicit textual signals rather than explicit protected-attribute labels? More fundamentally, are explicit demographic labels really the strongest supervision for debiasing, or can indirect cues be equally effective? Although prior work has explored proxies such as name embeddings for race and gender (Romanov et al., 2019), direct protected-attribute supervision remains the dominant paradigm.

We study this question through concept and attribute erasure from user self-description text in Stack Exchange (SE) communities (Figure 1). Our implicit signal comes only from free-form self-description, not from structured metadata such as reputation or engagement. We propose H-SAL, a variant of SAL (Shao et al., 2023b) that uses profile-text representations in place of explicit attribute labels. Because the same self-description signal captures multiple attribute-related cues, a single erasure step can mitigate several forms of user-attribute bias rather than requiring a separate run per attribute. We compare H-SAL against standard debiasing with explicit protected attributes, allowing us to test both whether implicit supervision is viable and whether it can surpass direct demographic supervision.

Across a new fairness benchmark for helpfulness prediction, we find that self-description text provides a surprisingly strong debiasing signal. In both encoder and decoder-only models, implicit post-hoc erasure often matches or outperforms explicit-label-based debiasing. More broadly, our results show that effective representation-level debiasing need not depend on explicitly observed protected attributes, but can instead emerge from indirect textual cues through which social information is encoded in language.

2 Related Work

Much fairness research in NLP assumes access to explicit protected attributes and removes them from learned representations through post-hoc erasure. INLP iteratively projects representations into the null space of attribute classifiers (Ravfogel et al., 2020), while RLACE, kernelized concept erasure, SAL, and LEACE extend this paradigm through adversarial optimization, nonlinear removal, spectral methods, and closed-form linear erasure (Ravfogel et al., 2022a,b; Shao et al., 2023b; Belrose et al., 2023). Earlier adversarial approaches also suppressed demographic information during training (Edwards and Storkey, 2016; Elazar and Goldberg, 2018; Barrett et al., 2019). Together, these methods form a strong debiasing toolkit, but they mostly rely on explicit protected-attribute supervision.

A smaller line of work considers weaker or indirect supervision. Romanov et al. (2019) use name embeddings as proxies for race and gender; AM-SAL (Shao et al., 2023a) removes unaligned attributes without paired instance-level labels, and IMSAE (Shao et al., 2025) transfers erasure subspaces across languages in zero-shot settings. Related work includes debiasing without demographics (Orgad and Belinkov, 2023), proxy-based fairness formulations (Gupta et al., 2018), and evidence that LLMs infer demographic traits from subtle cues and encode them in latent representations even without explicit disclosure (Neplenbroek et al., 2025). Similar post-hoc editing ideas also appear in text-to-image debiasing, though usually with explicit demographic annotations (Fu et al., 2025). Our work is closest to this second line, but uses only free-form self-description, rather than names or structured metadata, as the implicit signal and compares it against explicit-attribute debiasing.

3 Problem Formulation and Notation

We use bold lowercase letters for vectors, bold uppercase letters for matrices, and uppercase letters for random vectors. For random variables, $\mathbf{S}$ and $\mathbf{T}$, we denote by $\rho(\mathbf{S}, \mathbf{T})$ the correlation between them. When they are standardized (mean 0, variance 1), their correlation is also their covariance.

Each example consists of four variables:

$$\mathbf{X} \in \mathbb{R}^d, \qquad \mathbf{Y} \in \mathbb{R}, \qquad \mathbf{Z} \in \mathcal{Z}, \qquad \mathcal{H},$$

where $\mathbf{X}$ is the representation of the main-task input, $\mathbf{Y}$ is the task label, $\mathbf{Z}$ is the guarded attribute ($\mathcal{Z}$ denotes the set of possible attribute values), and $\mathcal{H}$ is the raw self-description text. We encode $\mathcal{H}$ with a text encoder ϕ into a continuous representation $\mathbf{H} = \phi(\mathcal{H}) \in \mathbb{R}^{d'}$.

Our goal is to predict $\mathbf{Y}$ from $\mathbf{X}$ while reducing the information about $\mathbf{Z}$ contained in $\mathbf{X}$. In the explicit setting, debiasing methods have direct access to $\mathbf{Z}$. In our implicit setting, $\mathbf{Z}$ is not used

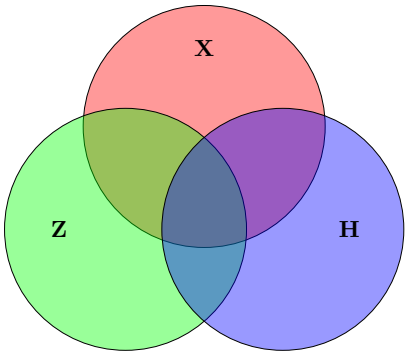


Figure 2: A Venn diagram showing the information content overlap between the different r.v.s. A success of H-SAL depends on how much the part of $\mathbf{Z}$ that overlaps with $\mathbf{X}$ also overlaps with $\mathbf{H}$.

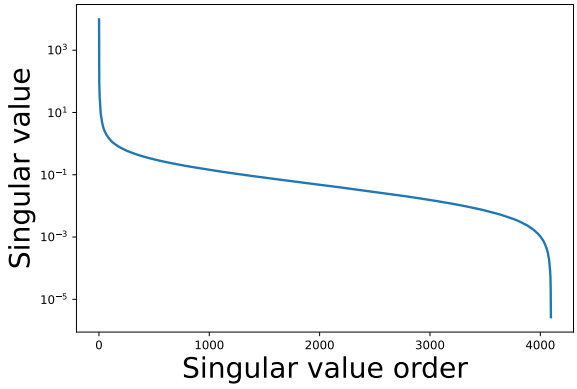


Figure 3: Singular values of the cross-covariance matrix between input representations $\mathbf{X}$ and self-description representations $\mathbf{H}$ for Llama-3.1-8B on the Mathematics dataset. The gradual decay suggests that user-related information is distributed across multiple latent directions, helping explain why H-SAL captures richer bias-related features and achieves stronger debiasing than SAL.

by the debiasing method; instead, we use the self-description text $\mathcal{H}$, through its representation $\mathbf{H}$, as an indirect signal for erasure.

We assume n samples $\left\{\left(\mathbf{x}^{(i)}, y^{(i)}, \mathcal{H}^{(i)}\right)\right\}_{i=1}^{n}$, with corresponding guarded attributes $\left\{z^{(i)}\right\}_{i=1}^{n}$. Let $\mathbf{h}^{(i)} = \phi\left(\mathcal{H}^{(i)}\right) \in \mathbb{R}^{d'}$ denote the representation of the i-th self-description. Our method estimates an erasure subspace from paired samples $(\mathbf{x}^{(i)}, \mathbf{h}^{(i)})$.

4 Methodology

Building on Spectral Attribute removaL (SAL; Shao et al. 2023b), we refer to our approach as H-SAL (hidden SAL). The algorithm performs the erasure by projecting the input representations from $\mathbf{X}$ into a subspace that is least covarying with $\mathbf{Z}$ (represented as a one-hot vector). Rather than working directly with $\mathbf{Z}$, we perform our erasure implicitly, by building the cross-covariance matrix that relies on $\mathbf{H}$:

$$\mathbf{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^{(i)}\left(\mathbf{h}^{(i)}\right)^{\top} \in \mathbb{R}^{d\times d'}.$$

Once $\mathbf{\Omega}$ is calculated, we perform singular value decomposition (SVD) on it, and then use the projection on samples from $\mathbf{X}$ based on $\hat{\mathbf{X}}$:

$$\hat{\mathbf{X}} = \left(\boldsymbol{I} - \boldsymbol{U}_{1:k}\boldsymbol{U}_{1:k}^{\top}\right)\mathbf{X}.$$

where $\boldsymbol{U}_{1:k}$ are the k columns of the left singular vector matrix derived by SVD on $\mathbf{\Omega}$ associated with the largest singular vectors. One of the reasons that SAL is ideal for the implicit approach is that it allows us to use continuous non-categorical variables as protected attributes, allowing us to replace a categorical $\mathbf{Z}$ with a vector $\mathbf{H}$.

It is beyond the scope of this paper to motivate this approach, as it relies on previous known work which uses projection methods to debias or erase information from embeddings (Mu and Viswanath, 2018; Liang et al., 2020; Shao et al., 2023b; Ravfogel et al., 2022a, *inter alia*).

Discussion and Justification (This can be skimmed in a first reading.) Replacing $\mathbf{Z}$ with $\mathbf{H}$ as a mediator has several key advantages:

- It does not require direct access to the protected attribute, but rather allows accessing and erasing it through indirect evidence and information;
- It allows a more gradual control over the erasure level, because k can vary and the rank of $\mathbf{\Omega}$ is larger than typically with a binary or multiclass attribute $\boldsymbol{Z}$ (if $\mathbf{\Omega}$ was computed directly using $\boldsymbol{Z}$); As shown in Figure 3, the singular values decay much more gradually when $\mathbf{\Omega}$ is derived from $\mathbf{H}$, indicating a higher effective rank than when using $\mathbf{Z}$ directly.
- It may capture more variability and subtle cues that are not directly captured by a discrete predetermined $\mathbf{Z}$, but are captured by $\mathbf{H}$.

We note that a central premise of this approach, when the goal is to erase the $\mathbf{Z}$ concept is that if $\mathbf{X}$ and $\mathbf{Z}$ covary, and if $\mathbf{H}$ and $\mathbf{Z}$ covary, then $\mathbf{X}$ and $\mathbf{H}$ covary in similar overlapping way. Hence, it is possible to use $\mathbf{H}$ as an intermediate variable.

Covariance levels are not necessarily retained through intermediate variables. For example, $\mathbf{H}$ and $\mathbf{Z}$ could be independent of each other (hence with covariance 0), but if $\mathbf{X} = \boldsymbol{A}\mathbf{Z} + \boldsymbol{B}\mathbf{H}$ for two

matrices $\boldsymbol{A} \in \mathbb{R}^{d\times m}, \boldsymbol{B} \in \mathbb{R}^{d\times d'}$ then $\mathbf{X}$ could covary with both. Or, alternatively, if all variables are Boolean and $\mathbf{X} = \mathbf{Z}$ *xor* $\mathbf{H}$, the covariance between $\mathbf{X}$ and either of the other $\mathbf{H}$ and $\mathbf{Z}$ could be 0 while $\mathbf{Z}$ and $\mathbf{H}$ could covary.

However, the intuition that covariance is transferable through an intermediate variable is to an extent true, and for simplicity, assume that the mean of $\mathbf{X}$, $\mathbf{H}$ and $\mathbf{Z}$ is 0 and that the random vectors are normalized by the square-root inverse of their (full-rank) covariance matrices (in which case, their covariance becomes the identity matrix). Define S to be a function that takes a matrix and returns the largest singular value of the matrix together with the left singular value and right singular value associated with them. Then, define:

$$(\rho_1, \mathbf{u}_1, \mathbf{v}_1) = S(\mathbb{E}[\mathbf{X}\mathbf{Z}^\top]), \quad (1)$$

$$(\rho_2, \mathbf{u}_2, \mathbf{v}_2) = S(\mathbb{E}[\mathbf{X}\mathbf{H}^\top]), \quad (2)$$

$$(\rho_3, \mathbf{u}_3, \mathbf{v}_3) = S(\mathbb{E}[\mathbf{H}\mathbf{Z}^\top]). \quad (3)$$

It holds that ρ_i corresponds to the correlation (and covariance, due to the normalization above) between the $\mathbf{u}_i$, $\mathbf{v}_i$ projection of the corresponding r.v.s in the covariance matrix. The question one might ask is: *can we express (or lower-bound) ρ_1 in terms of ρ_2 and ρ_3 (and all $\mathbf{u}_i$ and $\mathbf{v}_i$)?* In cases where $\mathbf{Z}$ is available in only small amounts, this means it could be labeled through $\mathbf{H}$ on a small dataset, and then the concept erasure could be applied to $\mathbf{X}$ through $\mathbf{H}$ with provided guarantees on the amount removed with respect to $\mathbf{Z}$. This idea can be formalized as follows. Define the function for $0 \leq a, b \leq 1$:

$$f(a, b) = \max\{0, ab - \sqrt{(1-a^2)(1-b^2)}\}.$$

In Appendix A, we show that under simple conditions on the singular vectors, it holds that:

$$\rho_1 \geq f(\rho_2, f(\rho_3, \mathbf{u}_3^\top \mathbf{v}_2)).$$

Note that the bound can be symmetrically flipped to have ρ_2 on the left, dependent on ρ_1. Figure 4 provides a plot of the bound on ρ_1 for two cases: one in which there is a 45 degree angle between $\mathbf{u}_3$ and $\mathbf{v}_2$ and one in which they are identical.

While we aim to minimize covariance between the projected $\mathbf{X}$ and $\mathbf{Z}$ (or $\mathbf{H}$), we do so by finding the maximally covarying directions, and then project to their complement space. Figure 2 provides the intuition in terms of a Venn diagram.

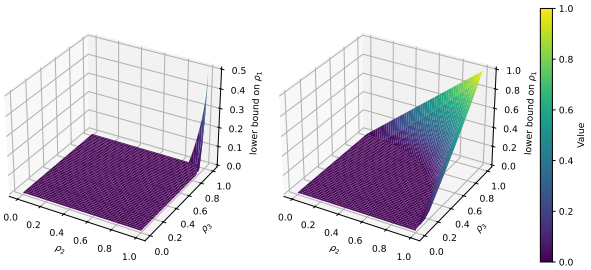


Figure 4: Plots for the bound for ρ_1 where $\mathbf{u}_3^\top \mathbf{v}_2 = 0.5$ (left) and $= 1$ (right).

5 The Stack Exchange Fairness Benchmark

We study fairness in a setting where socially salient attributes are not given as structured labels, but are implicitly recoverable from text. Unlike benchmarks that assume explicit protected attributes, real platforms often expose user-authored metadata, such as biographies, that indirectly reveal identity or background.

To study this setting, we construct a new benchmark from SE. Each example consists of a main-task input $\mathbf{X}$, profile text $\mathcal{H}$, and user attribute labels $\mathbf{Z}$. Here, $\mathbf{X}$ is the question–answer context for helpfulness prediction ($\mathbf{Y}$), and $\mathcal{H}$ is user-provided self-description. The attribute $\mathbf{Z}$ is used only for evaluation and explicit-label baselines; in the implicit setting, the model must identify and remove the bias signal from profile text alone. Figure 1 shows such signals in practice.

The benchmark is designed to test: (1) implicit bias detection; (2) generalization beyond memorized user cues through *user-disjoint* splits; and (3) the separation of model bias from distributional artifacts through *natural* and *balanced* settings.

5.1 Source Communities and Example Construction

Domains. We construct four subsets from the SE network: **StackOverflow**, **Mathematics**, **SuperUser**, and **English**. Together, they span diverse content and discourse styles, from technical problem solving (StackOverflow, SuperUser) to formula-heavy reasoning (Mathematics) and language-focused explanations (English), enabling evaluation across different notions of helpfulness.

Instances. Each datapoint is an answer-level example consisting of: (i) the question–answer context as the main-task input $\mathbf{X}$; (ii) user-authored profile text $\mathcal{H}$ (primarily `AboutMe`); and (iii) user

metadata needed to derive evaluation attributes $\mathbf{Z}$ (e.g., location, reputation, engagement). We start from public SE data dumps and construct examples by joining post- and user-level tables, then apply a shared text cleaning and filtering pipeline.[1]

5.2 User Attributes and Implicit Bias Sources

We study five user attributes as sources of bias in helpfulness prediction: three profile attributes (**DisplayName**, **gender**, **country**) and two platform-derived status signals (**reputation**, **engagement**). For evaluation, the main task label **helpfulness** and the user attributes **reputation**, **gender** and **engagement** are binarized, while **country** is represented as a five-class attribute across all datasets and **DisplayName** is represented through clustered name groups.

DisplayName, gender, and country. DisplayName, gender, and country act as identity-related proxy signals. DisplayName groups are obtained by clustering user display names in embedding space, with the number of clusters selected by silhouette score. Gender labels are also derived from users' `DisplayName`: we match extracted names against the Baby Names in England and Wales dataset provided by the Office for National Statistics, using the records from 1996 onward to construct male- and female-name lexicons (Office for National Statistics, 2025). Country labels are derived from normalized free-form location strings and restricted to the top five most frequent countries in each dataset; the exact country sets are reported in Table 3. These attributes may capture cultural, linguistic, gendered, or regional cues that affect perceived credibility beyond answer content.

Reputation and engagement. Reputation and engagement are platform-native status signals. Reputation summarizes prior community approval, while engagement reflects participation intensity. Both can induce halo effects, causing models to favor users with signals of status or embeddedness rather than judging the answer itself.

Explicit vs. implicit settings. The benchmark supports both *explicit* and *implicit* fairness settings. In the explicit setting, the attribute label $\mathbf{Z}$ is available to the debiasing method. In the implicit setting, only user profile text $\mathcal{H}$ is available, while $\mathbf{Z}$ is used solely for evaluation. Full attribute construction details are deferred to Appendix B.2.

[1]We drop entries with missing or insufficient main text or profile text. Full extraction and preprocessing details are provided in Appendix B, with summary statistics reported in Table 2 in the appendix.

5.3 Binarizing Attributes and Labels

User status attributes and post helpfulness are not binary on SE. We therefore derive binary labels by thresholding platform scores:

$$\begin{aligned}\mathbf{Y} &= \mathbb{1}[\texttt{AnswerScore} \geq \tau_Y],\\ \mathbf{Z}_{\text{rep}} &= \mathbb{1}[\texttt{Reputation} \geq \tau_Z],\\ \mathbf{Z}_{\text{eng}} &= \mathbb{1}[\texttt{Engagement} \geq \tau_Z].\end{aligned}$$

Here, $\mathbf{Y}$ is the main-task helpfulness label, while $\mathbf{Z}_{\text{rep}}$ and $\mathbf{Z}_{\text{eng}}$ are binary status attributes derived from user reputation and engagement. Country and display name labels are derived via country normalization and display name clustering (§5.2).

Choosing thresholds. Thresholds are selected separately for each Stack Exchange community based on the empirical score distributions. For helpfulness, we choose dataset-specific answer-score thresholds so that the downstream label $\mathbf{Y}$ can be balanced within each split, avoiding majority-class prediction. For reputation and engagement, we choose dataset-specific thresholds that yield non-degenerate status groups while preserving sufficient data for user-disjoint evaluation. The selected thresholds are reported in Table 4 in the appendix.

For balanced settings, we additionally construct the largest feasible subset in which each joint group $(\mathbf{Z}, \mathbf{Y}) \in \{0, 1\}^2$ occupies 25% of the data. This allows us to separate model-internal bias from simple label–attribute imbalance. Full threshold grids and selected values are reported in Appendix B.3.

5.4 Data Splits

We use a **User-Disjoint** setting. Splits are constructed by post counts with a 6:1:3 train–validation–test ratio, while enforcing that no `UserId` appears in more than one split. This avoids train–test leakage through repeated users.

The helpfulness label $\mathbf{Y}$ is balanced in each split to prevent majority-class prediction. For the attribute label $\mathbf{Z}$, reputation, engagement, and gender are approximately balanced, while display name clusters and country retain their natural distributions because these attributes are highly imbalanced and long-tailed; fully balancing them would discard most examples and yield very small subsets. Further construction details are provided in Appendix B.3.

5.5 Relation to Other SE Datasets

Prior work has used SE as a source of natural language data with associated user attributes (Rao and Daumé III, 2018; Brooke, 2019; Hazoom et al., 2021, *inter alia*). The closest to our setting is Shao et al. (2025), who introduce the multilingual SE-FAIR dataset. Our benchmark differs in two ways: it studies bias sources available through *textual user metadata*, enabling implicit debiasing from free-form profiles rather than only structured labels, and it includes a true **user-disjoint** regime, preventing methods from exploiting recurring user-specific cues across data splits.

6 Experimental Setup

We evaluate whether latent concept erasure reduces bias in helpfulness prediction while preserving prediction accuracy. Each experiment instantiates: (i) $\mathbf{Y}$, corresponding to answer helpfulness; (ii) $\mathbf{Z}$, corresponding to one of the five bias sources described in §5; and (iii) profile text $\mathcal{H}$ for the implicit signal.

Representations and encoders. We instantiate the framework with three open-source encoders: BERT (Devlin et al., 2019), Llama-3.1-8B (Grattafiori et al., 2024), and Mistral-7B (Jiang et al., 2023). For each input text, we extract representations from the final hidden layer. For BERT, we use the first-token (`[CLS]`) representation; for decoder-only models, we use the final non-padding token representation. These representations are used both for helpfulness prediction and for estimating the erasure subspace.

Debiasing variants. We compare four conditions: **Baseline** (no debiasing), **SAL (random)** (removing random directions as a control), **SAL** (estimating the erasure subspace from the explicit attribute label $\mathbf{Z}$), and **H-SAL** (estimating the erasure subspace from the implicit profile representation $\mathbf{H}$ without directly using $\mathbf{Z}$). After projection, we train a lightweight classifier to predict helpfulness. The random control tests whether any observed change comes from removing an arbitrary subspace rather than one aligned with the target bias. Full classifier and implementation details are provided in Appendix C.1.

Metrics. We report main-task accuracy for helpfulness prediction and **TPR-Gap** as the fairness metric. Following prior work on equal opportunity (De-Arteaga et al., 2019), TPR-Gap measures disparities in true positive rates across groups defined by the target attribute $\mathbf{Z}$. Lower values indicate fairer behavior. Full details in Appendix D.1.

Erasure strength. Erasure requires choosing the number of removed directions, k. We use $k = 50$ in the main results and examine smaller values in §7.2; additional details are provided in Appendix C.2.

Predictability of Z from H. As discussed in §4, our method assumes that the representation $\mathbf{H}$ contains information about the protected attribute $\mathbf{Z}$; otherwise, there would be little signal to remove. Table 5 confirms this assumption, shows that profile-text embeddings encode attribute-related information, and provides a justification for studying debiasing in the representation space.

7 Experiments

We evaluate three models and compare **implicit** and **explicit** textual erasure against a **random** baseline. Our main results are averaged over the three models, providing a compact view of the overall fairness–utility trade-off across datasets and attributes. Appendix E reports per-model results and sensitivity to the number of removed directions.

7.1 Main Results

Table 1 reports model-averaged results in the natural user-disjoint setting across BERT, Llama-3.1-8B, and Mistral-7B. We focus on model averages to highlight the overall pattern across domains and bias types; full per-model results appear in Appendix E.1. Across all five attributes, the random baseline stays close to the original model, with only small changes in TPR-Gap, indicating that the fairness gains of SAL and H-SAL do not come from removing arbitrary directions. Reductions in TPR-Gap are sometimes accompanied by modest drops in main-task accuracy; because H-SAL removes a larger subspace ($k = 50$) than SAL (bias-specific rank), it tends to trade slightly more accuracy for fairness, a trade-off we revisit in §7.2.

Overall, *H-SAL is competitive with, and often outperforms, SAL*, even though SAL uses the protected attribute directly. The advantage is clearest for display name, country and reputation in the technical domains, while gender and engagement show a mixed picture. An explanation is that the explicit attribute is often a coarse summary, whereas profile text encodes a richer set of correlated cues, such as place names, affiliations, seniority markers, and stylistic signals. Consistent with this interpretation, the ablation in §7.2 shows that H-SAL

Bias Type	Method	Stackoverflow		Mathematics		SuperUser		English	
		Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
DisplayName	Baseline	69.2	4.9	64.0	4.8	71.1	4.1	63.5	9.9
	SAL (random)	↓0.4 68.8	↓0.7 4.2	↓0.2 63.8	↓0.4 4.4	71.1	↑0.1 4.3	↓1.0 62.5	↓1.5 8.4
	SAL	↓0.8 68.4	↓2.9 2.0	↓1.1 62.8	↓2.0 2.8	↓0.3 70.9	↑0.3 4.4	↓0.8 62.7	↓0.8 9.1
	H-SAL	↓1.7 67.5	↓3.6 1.3	↓1.9 62.1	↓3.2 1.6	↓1.0 70.1	↓1.2 2.9	↓2.1 61.5	↓3.7 6.2
Gender	Baseline	56.7	0.3	54.7	2.8	58.8	11.6	54.3	9.9
	SAL (random)	↓0.8 55.9	↑0.5 0.8	↑0.3 55.0	↓0.1 2.8	↓1.6 57.2	↓0.8 10.7	↓1.3 52.9	↑1.8 11.7
	SAL	↓0.5 56.1	↑0.5 0.8	↑0.4 55.1	↓0.1 2.8	↑0.4 59.1	11.6	↑0.1 54.4	↓0.1 9.8
	H-SAL	↓2.4 54.3	↑1.2 1.5	↓1.4 53.3	↓1.5 1.4	↓0.7 58.1	↓1.2 10.4	↓1.8 52.4	↓0.4 9.5
Country	Baseline	67.4	11.5	61.0	9.7	63.4	7.0	59.3	10.1
	SAL (random)	↓0.3 67.0	↓0.7 10.8	↓0.1 60.9	↓0.5 9.3	↓0.2 63.2	↓0.1 6.9	59.3	↓0.8 9.2
	SAL	↓0.6 66.8	↓8.1 3.4	↓0.1 60.9	↓2.7 7.1	↓0.3 63.1	↓3.9 3.0	↑0.1 59.5	↓0.3 9.8
	H-SAL	↓2.3 65.1	↓9.5 2.0	↓3.9 57.1	↓5.5 4.2	↓0.9 62.5	↓3.5 3.5	↓1.5 57.8	↓0.6 9.4
Reputation	Baseline	69.2	8.6	64.0	11.8	71.1	8.8	63.5	9.8
	SAL (random)	↓0.4 68.8	↓0.9 7.7	↓0.2 63.8	↓0.8 11.1	71.1	↑0.1 8.9	↓1.0 62.5	↓0.8 9.0
	SAL	↓0.8 68.4	↓8.4 0.2	↓1.2 62.7	↓10.0 1.8	↓0.8 70.3	↓4.9 3.9	↓1.3 62.2	↓4.9 4.9
	H-SAL	↓1.7 67.5	↓8.3 0.3	↓1.9 62.1	↓10.1 1.7	↓1.0 70.1	↓5.5 3.2	↓2.1 61.5	↓4.9 4.9
Engagement	Baseline	68.6	7.7	63.6	8.5	67.2	8.4	63.6	9.6
	SAL (random)	↓0.4 68.2	↓1.0 6.7	↓0.2 63.4	↓0.9 7.6	↓0.1 67.1	↓0.7 7.8	↓0.6 63.0	↓2.0 7.6
	SAL	↓1.1 67.5	↓7.2 0.5	↓1.1 62.5	↓6.4 2.1	↓0.4 66.8	↓5.4 3.1	↓1.3 62.2	↓6.0 3.6
	H-SAL	↓2.0 66.6	↓7.2 0.5	↓2.5 61.2	↓6.7 1.8	↓0.7 66.5	↓4.9 3.6	↓1.9 61.7	↓3.1 6.5

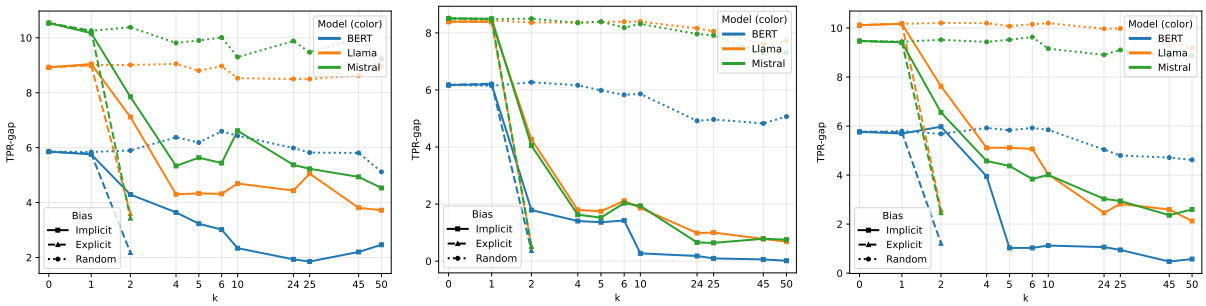
Table 1: **Average results under the user-disjoint split.** Results are averaged across BERT, Llama-3.1-8B, and Mistral-7B over five bias attributes and four Stack Exchange domains. **Main** denotes helpfulness prediction accuracy, and **TPR-Gap** denotes generalized disparity, computed as $2\times$ the standard deviation of true positive rates across groups. SAL uses bias-specific ranks, while SAL (random) and H-SAL use $k = 50$ removed directions for all attributes. The helpfulness label is balanced in each split; reputation, engagement, and gender are approximately balanced, while display-name clusters and country retain their natural distributions. Baseline denotes no erasure.

often continues to improve as more directions are removed, while random removal stays flat.

Display Name. Display name is the strongest case for implicit erasure: H-SAL achieves the lowest TPR-Gap in all four domains. On StackOverflow it reduces the gap from 4.9% to 1.3%, compared with 2.0% for SAL, and Mathematics shows the same ordering (4.8% to 1.6%, versus 2.8%). The contrast is sharpest on SuperUser, where the baseline gap is small (4.1%) and neither random nor SAL reduces it (4.3% and 4.4%, respectively), while H-SAL still lowers it to 2.9%. On English, H-SAL reduces the gap from 9.9% to 6.2%, again ahead of SAL at 9.1%. Profile text therefore provides a consistently useful signal for name-related bias, even where the explicit label does not.

Gender. Gender shows the smallest baseline disparities of all five attributes, and on StackOverflow the gap is already negligible (0.3%), so there is little bias to remove and all methods keep it small. On the remaining domains, H-SAL gives the lowest TPR-Gap: it reduces the gap from 2.8% to 1.4% on Mathematics, and produces smaller reductions on SuperUser (11.6% to 10.4%) and English (9.9% to 9.5%), whereas SAL leaves the gap essentially unchanged. The gender results are thus modest and reflect a weak baseline signal rather than a failure of erasure; main-task accuracy is also lower in the gender setting, indicating an intrinsically harder, attribute-balanced task.

Country. For country, H-SAL produces the largest reductions in the two technical domains: it lowers the gap from 11.5% to 2.0% on StackOverflow and from 9.7% to 4.2% on Mathematics, in both cases well below SAL (3.4% and 7.1%). On SuperUser the two methods are close, with SAL marginally lower (3.0% versus 3.5%). English is a harder case: the baseline gap (10.1%) barely moves under any method, including SAL. This pattern sug-



(a) Superuser (b) StackOverflow (c) Mathematics

Figure 5: **Engagement bias reduction across three Stack Exchange communities under the natural-disjoint split.** Each subfigure plots TPR-Gap against the number of removed directions (k) for `BERT`, `Mistral-7B`, and `Llama-3.1-8B`, comparing random, explicit, and implicit SAL. **TPR-Gap** is the generalized disparity ($2\times$ StdDev) of true positive rates across groups, equivalent to the difference in true positive rates in the binary case. Lower values indicate smaller disparity.

gests that location-related cues are present in user biographies in a sufficiently systematic form on StackOverflow and Mathematics, but that the signal is much weaker, or more entangled with task content, in the English community.

Reputation. Reputation is a case where H-SAL and SAL are essentially on par, and both far outperform random removal. H-SAL gives the lowest or tied-lowest TPR-Gap in three of the four domains: it matches SAL on English (both 4.9%), slightly improves on it on Mathematics (1.7% versus 1.8%), and is clearly better on SuperUser (3.2% versus 3.9%). StackOverflow is the only domain where SAL is marginally ahead (0.2% versus 0.3%). In all cases the gap is reduced substantially from a baseline of roughly 9–12%, showing that profile text recovers status-related bias about as effectively as the explicit reputation label.

Engagement. Engagement gives a more mixed outcome. H-SAL is best or tied on StackOverflow (both methods reach 0.5%, from a baseline of 7.7%) and Mathematics (1.8% versus 2.1%), but SAL is stronger on SuperUser (3.1% versus 3.6%) and also on English, where SAL reaches 3.6% while H-SAL only lowers the gap to 6.5%. Both methods improve substantially over random removal in every domain. Engagement therefore illustrates that the implicit signal is not uniformly superior: when participation-related cues are weakly expressed in self-description, the explicit label can still be the more reliable supervision source.

7.2 Effect of the Number of Removed Directions

Table 1 reports results under a strong erasure setting, $k = 50$. Figure 5 shows TPR-Gap as k varies for engagement bias across communities and backbones; additional model-averaged results for $k \in \{5, 10, 24, 45\}$ are provided in Appendix E.2. As k increases, H-SAL generally reduces TPR-Gap more than the random baseline, indicating that improvements come from removing bias-aligned rather than arbitrary directions. This also helps explain why H-SAL can outperform explicit SAL: profile text spreads bias across many correlated directions (Figure 3), while the explicit label spans only a narrow subspace. Since larger k may also affect helpfulness accuracy, k controls the fairness–utility trade-off.

8 Conclusion

We study fairness in a setting where protected attributes are not directly available, but may still be implicitly encoded in user-authored text. We introduce a multi-domain Stack Exchange-based benchmark that includes both explicit attribute labels and implicit self-description, alongside a method that performs erasure once from implicit self-description rather than requiring a separate run per predefined attribute. We find that implicit self-description often provides a strong debiasing signal, matching or outperforming explicit-label-based erasure across most settings. These results broaden representation-level fairness research and suggest that indirect textual cues can serve as an effective basis for fairness interventions in realistic settings where demographic labels are absent.

Limitations

Our study has several limitations. First, although the empirical results are strong, we evaluate only one family of post-hoc debiasing methods, namely SAL and its implicit variant. Other representation-level debiasing methods, such as INLP, RLACE, LEACE, or newer post-hoc editing approaches, may behave differently in the implicit setting. We note, however, that SAL lends itself very naturally to our setup, which is the reason we use it.

Second, our benchmark focuses on helpfulness prediction in Stack Exchange communities. While this provides a realistic and diverse testbed, the findings may not transfer directly to other tasks, domains, or interaction settings. In particular, the relationship between user self-description and socially salient attributes may vary substantially across platforms. We note, however, that our dataset spans four distinct communities, each with substantially different discourse styles.

Third, our implicit signal is restricted to self-description text. This is a deliberate design choice that isolates the role of free-form textual metadata, but it leaves open how other realistic implicit signals, such as writing style, interaction history, or broader profile context, might affect debiasing.

Ethical Considerations

Our work is motivated by the fact that protected attributes are often unavailable in practice due to privacy, legal, and data-collection constraints. The goal of our method is not to infer or reconstruct sensitive user information, but to reduce the extent to which models rely on such information when making predictions. At the same time, studying implicit bias necessarily involves attributes that may be socially sensitive or correlated with identity. This creates two ethical risks. First, user-authored profile text may contain personal information, and even aggregate analysis of such text should be handled with care. Second, methods for detecting or removing implicit bias signals could be misused to profile users or infer private attributes. We therefore position our benchmark and methods strictly as tools for fairness evaluation and mitigation, not for demographic inference or user profiling.

A further consideration is that the attributes studied in this paper, especially *display name* and *country*, are proxies rather than ground-truth demographic labels. They should not be interpreted as definitive indicators of identity. Likewise, reducing disparity on a benchmark does not guarantee fairness in all downstream settings, since fairness is multi-dimensional and depends on the deployment context. We therefore view this benchmark as a controlled testbed for studying one important aspect of fairness under missing protected attributes, rather than as a complete solution to fairness in real-world systems.

References

Carolyn Ashurst and Adrian Weller. 2023. Fairness without demographic data: A survey of approaches. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *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 6330–6335, Hong Kong, China. Association for Computational Linguistics.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pages 66044–66063. Curran Associates, Inc.

Sian Brooke. 2019. "condescending, rude, assholes": Framing gender and hostility on Stack Overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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. Association for Computational Linguistics.

Harrison Edwards and Amos J. Storkey. 2016. Censoring representations with an adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Zihao Fu, Ryan Brown, Shun Shao, Kai Rawal, Eoin Delaney, and Chris Russell. 2025. Fairimagen: Post-processing for bias mitigation in text-to-image models. In *Advances in Neural Information Processing Systems*, volume 38, pages 29488–29512. Curran Associates, Inc.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness. *Preprint*, arXiv:1806.11212.

Moshe Hazoom, Vibhor Malik, and Ben Bogin. 2021. Text-to-SQL in the wild: A naturally-occurring dataset based on stack exchange data. In *Proceedings of the 1st Workshop on Natural Language Processing for Programming (NLP4Prog 2021)*, pages 77–87, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2025. Reading between the prompts: How stereotypes shape LLM's implicit personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20367–20400, Suzhou, China. Association for Computational Linguistics.

Office for National Statistics. 2025. Baby names in England and Wales: from 1996. Dataset, released 31 July 2025.

Hadas Orgad and Yonatan Belinkov. 2023. BLIND: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821, Toronto, Canada. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. 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, Online. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022a. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022b. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. In *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 4187–4195, Minneapolis, Minnesota. Association for Computational Linguistics.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023a. Erasure of Unaligned Attributes from Neural Representations. *Transactions of the Association for Computational Linguistics*, 11:488–510.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023b. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, Dubrovnik, Croatia. Association for Computational Linguistics.

Shun Shao, Yftah Ziser, Zheng Zhao, Yifu Qiu, Shay B Cohen, and Anna Korhonen. 2025. Iterative multilingual spectral attribute erasure. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29230–29255, Suzhou, China. Association for Computational Linguistics.

A More Details on Justification

Lemma 1. *Let* $\mathbf{R}$*,* $\mathbf{T}$ *and* $\mathbf{S}$ *be three scalar random variables. Define:*

$$f(a,b) = \max\{0, ab - \sqrt{(1-a^2)(1-b^2)}\}.$$

If the correlation between all pairs of the three random variables is non-negative, then:

$$\rho(\mathbf{R},\mathbf{S}) \geq f(\rho(\mathbf{R},\mathbf{T}), \rho(\mathbf{T},\mathbf{S})).$$

There is an extensive discussion of this lemma on Mathematics Stack Exchange at https://tinyurl.com/pzn4uf58, including an explanation of why it is true. We note that f is monotonically non-decreasing in both arguments. Using the definitions in Eq. 1–3 and Lemma 1, together with a composition of f with itself, we obtain:

Proposition 1. *Assume* $\mathbf{u}_3^\top \mathbf{v}_2 \geq 0$ *and that* $\rho(\mathbf{v}_2^\top \mathbf{H}, \mathbf{u}_3^\top \mathbf{Z}) \geq 0$*. It holds that*

$$\rho_1 \geq f(\rho_2, f(\rho_3, \mathbf{u}_3^\top \mathbf{v}_2)).$$

Proof. We know the correlation of projected random vectors is maximized when using the first left/right singular vector, and therefore:

$$\rho_1 \geq \underbrace{\rho(\mathbf{u}_2^\top \mathbf{X}, \mathbf{v}_3^\top \mathbf{Z})}_{\hat{\rho}}.$$

By Lemma 1 it holds that:

$$\hat{\rho} \geq f(\rho_2, \rho(\mathbf{v}_2^\top \mathbf{H}, \mathbf{u}_3^\top \mathbf{Z})).$$

By the monotonicity of f in both arguments over $[0,1]$ and the definition of ρ_3 we can use Lemma 1 again (with $\mathbf{T} = \mathbf{u}_3^\top \mathbf{H}$) and replace the second argument to get:

$$\hat{\rho} \geq f(\rho_2, f(\rho_3, \rho(\mathbf{u}_3^\top \mathbf{H}, \mathbf{v}_2^\top \mathbf{H}))),$$

Since we assume $\mathbf{H}$ is normalized, it holds that the last ρ argument equals $\mathbf{u}_3^\top \mathbf{v}_2$. Putting it all together we obtain that

$$\rho_1 \geq f(\rho_2, f(\rho_3, \mathbf{u}_3^\top \mathbf{v}_2))$$

as required. □

We note that due to the degree of freedom in the sign of the left and right singular vectors in SVD, we can relax the condition that $\mathbf{u}_3^\top \mathbf{v}_2$ and $\rho(\mathbf{v}_2^\top \mathbf{H}, \mathbf{u}_3^\top \mathbf{Z})$ are both non-negative to them having the same sign.

Dataset	Split	Reputation	Gender	Location	Engagement	DisplayName
StackOverflow	Train	69976 / 69976	2894 / 2894	45088 / 16428 / 16767 / 35975 / 13877	104839 / 104839	17087 / 52472 / 27583 / 8349 / 34461
	Validation	10017 / 10017	716 / 716	7456 / 1876 / 2168 / 4969 / 1786	15059 / 15059	2501 / 7799 / 3861 / 1110 / 4763
	Test	19967 / 19967	854 / 854	12027 / 4934 / 4001 / 9896 / 3994	29933 / 29933	4986 / 15058 / 7900 / 2298 / 9692
Mathematics	Train	33273 / 33273	3616 / 3616	34836 / 7755 / 8060 / 12447 / 14074	32913 / 32913	5949 / 4021 / 15362 / 26265 / 14949
	Validation	4770 / 4770	1112 / 1112	5345 / 942 / 1563 / 1544 / 1727	4619 / 4619	797 / 715 / 2250 / 3694 / 2084
	Test	9452 / 9452	1140 / 1140	11788 / 2692 / 2082 / 3762 / 3963	9361 / 9361	1457 / 1193 / 4204 / 7784 / 4266
SuperUser	Train	9601 / 9601	456 / 456	9617 / 5257 / 2354 / 2587 / 2044	9251 / 9251	2213 / 2052 / 4652 / 3486 / 6799
	Validation	1379 / 1379	194 / 194	1121 / 1100 / 377 / 305 / 313	1335 / 1335	279 / 282 / 644 / 501 / 1052
	Test	2746 / 2746	188 / 188	1998 / 1856 / 1258 / 851 / 566	2640 / 2640	581 / 514 / 1409 / 1009 / 1979
English	Train	2882 / 2882	616 / 616	4425 / 4770 / 1276 / 736 / 308	2838 / 2838	1798 / 579 / 682 / 1474 / 1231
	Validation	415 / 415	278 / 278	490 / 766 / 90 / 59 / 6	406 / 406	257 / 80 / 97 / 237 / 159
	Test	812 / 812	266 / 266	1671 / 1107 / 329 / 118 / 161	806 / 806	407 / 230 / 230 / 459 / 298

Table 2: **Number of users in each split by protected-attribute class under the natural disjoint split.** For binary attributes, entries are reported as class counts in the form $Z = 0$ / $Z = 1$. For location, entries are reported as counts for each country class. For DisplayName, entries are reported as counts for each of the display name clusters.

B Dataset Construction Details

B.1 Data Extraction and Preprocessing

We start from the public Stack Exchange data dumps for each community and construct answer-level examples by joining post- and user-level tables. For each answer, we retrieve: (i) the answer text and its score; (ii) the associated question context, such as title and tags when available; and (iii) user metadata including `UserId`, `DisplayName`, `AboutMe`, `Location`, `Reputation`, and interaction statistics used to compute engagement. These fields are aligned into a post-level table where each row corresponds to one answer and contains both the main-task text and the user profile text $\mathcal{H}$.

The main-task text is constructed by concatenating the cleaned question title and cleaned answer body into a single prompt of the form `Question: <title>, Answer: <answer>`. This text is then encoded to obtain the representation $\mathbf{X}$ used for helpfulness prediction and debiasing.

All textual fields, including question and answer content, self-descriptions, and location strings, are processed with a shared cleaning pipeline. We remove HTML markup and normalize whitespace and formatting to obtain cleaned text fields. We discard entries where either the main answer text or the user profile text is empty after cleaning, and filter out very short texts that provide insufficient signal for representation learning or probing.[2]

Table 2 provides statistics on the dataset.

[2]We apply the same validity checks across domains to keep preprocessing consistent.

B.2 Attribute Construction

We study five user attributes as sources of bias in helpfulness prediction: **display name**, **gender**, **country**, **reputation**, and **engagement**. Three of these (**display name**, **gender**, and **country**) are socially salient profile attributes, while the other two (**reputation**, **engagement**) are platform-derived status signals. We do not treat them as definitive ground-truth demographics; instead, they define evaluation groups that may correlate with writing style, topical focus, visibility, and community norms, and may therefore be spuriously predictive of helpfulness even when answer quality is comparable.

Display Name. Display names can encode cultural, linguistic, or stylistic cues that models may exploit. We simulate this source of bias by clustering user names in embedding space and treating the resulting clusters as evaluation groups. We select the number of clusters using the average silhouette score and use five clusters in our experiments.

Country. Country labels are derived from users' free-form location strings. Because these strings are noisy and heterogeneous, we normalize them to standardized country labels using a location-resolution pipeline and retain only cases that resolve unambiguously to a single country. To ensure sufficient support per group, we restrict each domain to its most frequent countries after normalization. We provide the resulting country labels in Table 3.

Gender. Gender is inferred from users' `DisplayName`. We extract likely first names and match them to male and female name lists from the Office for National Statistics' *Baby names in*

Dataset	Retained country labels
Mathematics	USA, FRA, CAN, DEU, GBR
StackOverflow	USA, IND, DEU, GBR, CAN
English	GBR, USA, CAN, IND, ITA
SuperUser	USA, GBR, CAN, DEU, IND

Table 3: **Country labels retained for each Stack Exchange dataset.** Country labels are obtained by normalizing users' free-form location strings and retaining the five most frequent countries in each dataset.

England and Wales: from 1996 dataset (Office for National Statistics, 2025). We keep only entries with an unambiguous match. Gender is a widely studied demographic attribute in fairness and bias evaluation, and display names provide a common proxy signal through which such bias may enter text-based systems.

Reputation and engagement. Reputation is a platform-native status score summarizing prior community feedback, while engagement measures participation intensity based on historical interaction statistics. Although both are correlated with user experience, they may also act as socially salient status signals and induce halo effects, causing models to favor high-status or highly active users independently of answer quality.

Explicit vs. implicit settings. The benchmark supports both *explicit* and *implicit* fairness settings. In the explicit setting, the attribute label $\mathbf{Z}$ is directly available to the debiasing method. In the implicit setting, only the user profile text $\mathcal{H}$ is available, and $\mathbf{Z}$ is used solely for evaluation. Our main experiments focus on the implicit setting, which is the natural target for latent concept and attribute erasure.

Display-name cluster selection. To choose the number of display-name clusters, we compute the average silhouette score for candidate values of k. For each point i, let $a(i)$ be the average distance to points in its own cluster, and let $b(i)$ be the minimum average distance to points in any other cluster. The silhouette score is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

and the average silhouette score for k clusters is

$$\bar{s}(k) = \frac{1}{n} \sum_{i=1}^{n} s(i).$$

We choose

$$k^* = \arg\max_{k \in \mathcal{K}} \bar{s}(k),$$

which favors clusters that are compact and well separated. This procedure selects five display-name clusters in our setting.

B.3 Evaluation Split Construction

All splits in our benchmark are constructed to be user-disjoint: no `UserId` appears in more than one split. This ensures that both probe performance and fairness metrics reflect generalization to unseen users rather than recurring user-specific cues. We use a 60/10/30 train–validation–test ratio by post counts, while enforcing user disjointness to prevent leakage through repeated users.

The helpfulness label $\mathbf{Y}$ is balanced in each split to avoid majority-class prediction. For the attribute label $\mathbf{Z}$, reputation, engagement, and gender are approximately balanced where feasible. Display-Name clusters and country retain their natural distributions because they are highly imbalanced and long-tailed; fully balancing them would discard most examples and yield very small subsets.

Selected thresholds. Table 4 reports the dataset-specific thresholds used to binarize helpfulness, reputation, and engagement. Helpfulness thresholds are chosen according to the answer-score distribution of each dataset so that the downstream task label can be balanced in each split. Reputation and engagement thresholds are selected to form non-degenerate status groups with sufficient support for user-disjoint evaluation.

Dataset	Help. thr.	Rep. thr.	Eng. thr.
English	3	185	6
Mathematics	2	466	22
StackOverflow	2	2498	246
SuperUser	2	121	121

Table 4: **Selected thresholds for binarizing labels and status attributes.** Help., Rep., and Eng. denote helpfulness, reputation, and engagement, respectively. Helpfulness is derived from answer score. Reputation and engagement are user-level status attributes.

B.4 Example Construction

The main task is binary helpfulness prediction from Stack Exchange question–answer data. Each example is constructed at the answer level. The task

input $\mathbf{X}$ is derived from a textual prompt formed by concatenating the question title and the answer text (e.g., "Question: … Answer: …"). The implicit textual signal $\mathcal{H}$ is the user's profile text, taken from the `AboutMe` field. The target attribute $\mathbf{Z}$ is one of *DisplayName*, *Country*, *Gender*, *Reputation*, or *Engagement*, depending on the experiment.

C Experimental Setup Details

C.1 Classifier and Training Details

For downstream helpfulness prediction, we train a logistic regression classifier on top of the projected representation. We use the `scikit-learn` implementation with `max_iter=1000` and `random_state=0`; all other hyperparameters are left at their default values unless otherwise noted.

We use the same classifier architecture and training procedure for all experimental conditions: **Baseline** (no debiasing), **SAL (random)**, **SAL**, and **H-SAL**. Keeping the downstream classifier fixed ensures that observed differences in main-task accuracy and fairness metrics are attributable to the representation intervention itself rather than to differences in model capacity or optimization.

In the **SAL (random)** condition, we remove the same number of directions as in the corresponding debiasing setup, but select those directions at random instead of deriving them from either the explicit attribute label $\mathbf{Z}$ or the implicit profile representation $\mathbf{H}$. This condition serves as a control: if random removal produces little change while explicit or implicit erasure substantially reduces disparity, this supports the interpretation that the learned subspace is specifically aligned with the bias signal rather than reflecting an arbitrary loss of information.

For each condition, the classifier is trained on the transformed training representations and evaluated on the correspondingly transformed test representations. This design isolates the contribution of the erasure step while preserving a common prediction pipeline across all experiments.

C.2 Choice of Erased Dimensionality

A key hyperparameter in concept erasure is the number of removed directions, k, which controls the strength of the intervention. Smaller values of k typically preserve more task-relevant information, while larger values can remove a broader bias-aligned subspace at the cost of some main-task performance. We therefore treat k as a fairness–utility trade-off parameter rather than as a fixed universal setting.

In the main experiments, we report results with $k = 50$, which corresponds to a relatively strong erasure setting. To assess sensitivity to this choice, we additionally evaluate $k \in \{5, 10, 24, 45\}$ and report the resulting trends in §7.2. These results show how the strength of erasure affects both fairness and accuracy, and allow practitioners to choose an operating point appropriate for their application.

C.3 Model Size and Compute

We use three backbone encoders with different model scales: BERT-base-uncased with approximately 110M parameters, Llama-3.1-8B with approximately 8B parameters, and Mistral-7B-v0.3 with approximately 7B parameters. We do not fine-tune these backbone models; they are used for representation extraction, followed by linear probing and erasure-based evaluation. Experiments were run on NVIDIA GH200 GPU nodes.

D Evaluation and Analysis

D.1 Evaluation Metrics

To evaluate the balance between solving $\mathbf{Y}$ (predicting helpfulness) and debiasing $\mathbf{X}$, we use accuracy (for $\mathbf{Y}$) and an equal opportunity gap measure (reduced to TPR-Gap in the case of binary $\mathbf{Z}$; De-Arteaga et al. 2019).

More specifically, for a specific class c in the main task (*Helpful* or *Unhelpful*), and a protected attribute g (a specific location or a reputation group), the disparity is defined as:

$$\text{Gap}_c = 2 \times \sigma(\{\text{TPR}_{c,g} \mid g \in Z\})$$

where $\text{TPR}_{c,g} = P(\hat{\mathbf{Y}} = c \mid \mathbf{Y} = c, \mathbf{Z} = g)$ and σ calculates the standard deviation. The r.v. $\hat{\mathbf{Y}}$ stands for the prediction, and the probability is estimated during testing based on data.[3] In the case of binary $\mathbf{Z}$, it trivially holds $\text{Gap}_c = |\text{TPR}_0 - \text{TPR}_1|$ where TPR_g is the true positive rate for class $g \in \{0, 1\}$.

For multiclass sensitive attributes (for example, five locations), we generalize the metric by calculating the root mean square (RMS) deviation of the true positive rates for each group from the global mean:

$$\text{Gap}_c = 2 \times \sqrt{\frac{1}{|\{g\}|} \sum_g (\text{TPR}_{c,g} - \mu_c)^2}$$

[3] Scaling by two allows the metric to provide a unified scale for comparing both binary and multiclass attributes.

where μ_c is the average TPR across all groups for class c. This formulation captures the overall variation in model performance across all groups simultaneously. Then, to obtain a single fairness scalar for the model, we aggregate the gaps calculated for every class c using RMS again:

$$\text{Total Fairness Score} = \sqrt{\frac{1}{|\{c\}|}\sum_{c}(\text{Gap}_c)^2}$$

This aggregation ensures that the model is penalized for disparities in both the positive and negative classes of the main task.

D.2 Predictability of Z from H: Validation of Implicit Bias

As discussed in §4, our method assumes that the profile representation $\mathbf{H}$ contains information about the target attribute $\mathbf{Z}$; otherwise, there would be little attribute-related signal to erase. We validate this assumption by probing $\mathbf{Z}$ from $\mathbf{H}$ on the same user-disjoint datasets used in our main experiments. Reputation, engagement, and gender are approximately balanced to make the probing task less sensitive to majority-class prediction, while **DisplayName** clusters and country retain their natural distributions because they are highly imbalanced and long-tailed. Because the probing split is also user-disjoint, no `UserId` and corresponding `Self-description` appears in more than one split, preventing leakage of user-specific profile information between train, validation, and test sets. Table 5 shows that profile-text embeddings encode substantial attribute-related information, providing a concrete motivation for textual erasure in representation space.

E Additional Results

This appendix provides additional results omitted from the main text for space. §E.1 reports per-model results under the main erasure setting, corresponding to the model-averaged results in the main paper. §E.2 reports model-averaged sensitivity to the number of removed directions k, illustrating the fairness–utility trade-off as erasure strength varies.

E.1 Per-Model Results

The main text reports results averaged across three backbone models. Here, we provide the corresponding per-model results for BERT (Table 6), Llama (Table 7), and Mistral (Table 8) under the same user-disjoint split and main erasure setting. For implicit H-SAL, we use $k = 50$, which gives the strongest debiasing effect in our experiments. SAL uses bias-specific ranks, corresponding to the number of explicit attribute groups. SAL (random) uses the same k as H-SAL, but removes randomly selected directions as a control.

E.2 Sensitivity to the Number of Removed Directions

The main results use $k = 50$ for H-SAL, which gives the strongest debiasing effect overall. To show the fairness–utility trade-off, we additionally report model-averaged results for $k \in \{5, 10, 24, 45\}$ in Table 9–12. While larger k generally reduces TPR-Gap more strongly, $k = 24$ often provides a good balance between debiasing and preserving helpfulness accuracy.

Dataset	Language Model	Probing Accuracy				
		Reputation	Gender	Location	Engagement	DisplayName
StackOverflow	BERT	67.4	74.2	51.0	67.7	38.1
	Llama3.1-8B	68.0	74.1	60.6	67.0	29.9
	Mistral-7B	67.4	73.2	58.0	66.9	26.5
Mathematics	BERT	73.1	74.9	58.2	68.0	40.1
	Llama3.1-8B	71.8	72.7	64.6	66.0	31.1
	Mistral-7B	70.6	70.5	63.4	64.5	25.8
SuperUser	BERT	64.0	66.0	48.0	63.6	32.9
	Llama3.1-8B	63.0	63.2	56.2	62.1	25.7
	Mistral-7B	61.7	66.0	55.3	62.0	22.7
English	BERT	69.0	61.9	57.9	66.6	25.9
	Llama3.1-8B	69.5	68.7	67.1	66.1	23.8
	Mistral-7B	69.3	61.9	64.4	67.1	22.1

Table 5: **Predictability of protected attributes from profile-text representations.** We report test accuracy for predicting the protected attribute $\mathbf{Z}$ from the representation $\mathbf{H}$ of users' self-descriptions. Higher accuracy indicates that the embedding retains more information about the attribute, implying stronger attribute relation to textual representations. For Location and DisplayName, there are five classes.

Bias Type	Method	**StackOverflow**		**Mathematics**		**SuperUser**		**English**	
		Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
DisplayName	Baseline	68.4	5.9	63.5	4.9	74.9	2.9	68.2	8.3
	SAL (random)	↓0.5 68.0	↓1.8 4.1	↓0.4 63.2	↓1.5 3.4	↓0.5 74.5	↑0.3 3.2	↓0.7 67.5	↓0.3 8.0
	SAL	↓0.6 67.8	↓4.3 1.6	↓1.0 62.5	↓1.7 3.2	↓0.3 74.6	↑0.6 3.5	↑0.2 68.5	↓0.5 7.8
	H-SAL	↓1.5 66.9	↓5.3 0.6	↓1.3 62.3	↓4.3 0.6	↓1.1 73.8	↓1.4 1.5	↓2.0 66.3	↓4.1 4.1
Gender	Baseline	55.7	0.5	57.3	2.9	58.2	6.1	53.9	6.7
	SAL (random)	↓0.8 55.0	↑1.2 1.7	↓0.5 56.8	↓0.6 2.2	↓3.5 54.8	↓1.7 4.4	↓1.5 52.4	↑2.3 8.9
	SAL	↑0.4 56.1	↑0.7 1.2	↑0.4 57.6	↓1.1 1.8	↑0.5 58.8	↓1.0 5.0	↓0.2 53.8	↑1.2 7.8
	H-SAL	↓2.5 53.3	↑0.3 0.8	↓3.2 54.1	↓2.0 0.9	↓2.4 55.9	↑0.7 6.7	↓2.6 51.3	↑2.8 9.5
Country	Baseline	66.4	10.7	60.0	12.8	63.5	3.6	61.3	13.0
	SAL (random)	↓0.4 66.0	↓1.3 9.3	↓0.4 59.5	↓1.8 11.1	63.6	↓0.9 2.7	61.3	↓3.0 9.9
	SAL	↓0.3 66.1	↓7.2 3.5	↓0.1 59.9	↓1.9 10.9	↓0.3 63.2	↓1.2 2.3	↓0.5 60.8	↓1.1 11.9
	H-SAL	↓2.2 64.2	↓9.2 1.5	↓4.0 56.0	↓7.9 5.0	↓0.6 62.9	↓0.6 3.0	↓1.6 59.7	↓6.7 6.3
Reputation	Baseline	68.4	6.3	63.5	7.3	74.9	6.6	68.2	12.4
	SAL (random)	↓0.5 68.0	↓1.3 4.9	↓0.4 63.2	↓0.9 6.3	↓0.5 74.5	↑0.5 7.1	↓0.7 67.5	↓3.2 9.2
	SAL	↓0.6 67.8	↓5.8 0.4	↓0.8 62.8	↓6.0 1.3	↓0.8 74.1	↓5.3 1.3	↓1.7 66.5	↓6.2 6.2
	H-SAL	↓1.5 66.9	↓6.2 0.1	↓1.3 62.3	↓6.8 0.5	↓1.1 73.8	↓4.0 2.6	↓2.0 66.3	↓8.0 4.4
Engagement	Baseline	67.5	6.2	63.0	5.8	69.7	5.9	67.5	13.4
	SAL (random)	↓0.6 66.9	↓1.1 5.1	↓0.3 62.7	↓1.1 4.6	↓0.1 69.6	↓0.7 5.1	↓0.4 67.1	↓2.1 11.4
	SAL	↓0.8 66.7	↓5.8 0.4	↓0.5 62.5	↓4.5 1.2	↑0.1 69.8	↓3.7 2.2	↓0.4 67.1	↓7.0 6.5
	H-SAL	↓1.7 65.8	↓6.2 0.0	↓1.8 61.2	↓5.2 0.6	69.7	↓3.4 2.5	↓1.1 66.4	↓6.1 7.4

Table 6: **BERT results under the main erasure setting. Main** denotes helpfulness prediction accuracy, and **TPR-Gap** denotes generalized disparity, computed as $2\times$ the standard deviation of true positive rates across groups. SAL uses bias-specific ranks, while SAL (random) and H-SAL use $k = 50$ removed directions for all attributes. The helpfulness label is balanced in each split; reputation, engagement, and gender are approximately balanced, while display-name clusters and country retain their natural distributions. Baseline denotes no erasure.

Bias Type	Method	StackOverflow		Mathematics		SuperUser		English	
		Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
DisplayName	Baseline	69.7	4.5	64.2	5.0	69.2	6.0	61.0	10.2
	SAL (random)	↓0.4 69.3	↓0.2 4.3	↑0.1 64.3	↓0.4 4.6	↑0.2 69.5	↑0.7 6.7	↓2.0 59.1	↓1.9 8.3
	SAL	↓0.8 68.8	↓1.9 2.6	↓1.3 62.8	↓2.4 2.6	↓0.1 69.1	↑1.3 7.3	↓0.5 60.5	↓2.6 7.6
	H-SAL	↓1.7 68.0	↓2.9 1.5	↓2.2 62.0	↓3.1 1.9	↓1.3 68.0	↓0.6 5.4	↓1.8 59.2	↓2.5 7.7
Gender	Baseline	56.6	0.0	54.2	2.2	58.5	13.5	54.9	13.8
	SAL (random)	↓0.3 56.3	↑0.4 0.4	↓0.1 54.1	↑1.4 3.6	58.5	13.5	↓0.9 53.9	↑0.7 14.5
	SAL	↓0.4 56.1	↑0.5 0.5	↑0.4 54.5	↑1.5 3.6	58.5	↑1.0 14.5	↑0.4 55.3	↓1.7 12.1
	H-SAL	↓1.8 54.7	↑3.4 3.4	↓1.6 52.5	↓0.6 1.6	↓0.5 58.0	↓0.1 13.3	↓2.8 52.1	↓4.2 9.6
Country	Baseline	67.9	12.8	61.7	8.8	63.4	8.3	57.8	6.5
	SAL (random)	↓0.2 67.6	↓0.1 12.7	↓0.1 61.5	8.8	63.3	↓0.2 8.1	↑0.4 58.2	↑1.2 7.8
	SAL	↓0.6 67.3	↓10.0 2.8	↓0.1 61.6	↓3.4 5.4	↓0.1 63.3	↓5.7 2.6	↑1.0 58.8	↓0.1 6.5
	H-SAL	↓2.1 65.8	↓10.8 2.0	↓3.9 57.8	↓4.5 4.3	↓0.8 62.6	↓5.2 3.0	↓0.7 57.1	↑4.4 10.9
Reputation	Baseline	69.7	9.9	64.2	13.9	69.2	10.8	61.0	7.8
	SAL (random)	↓0.4 69.3	↓0.8 9.1	↑0.1 64.3	↓0.4 13.5	↑0.2 69.4	10.7	↓2.0 59.1	↑0.5 8.3
	SAL	↓1.0 68.6	↓9.7 0.2	↓1.6 62.6	↓12.4 1.5	↓1.1 68.2	↓4.8 6.0	↓1.0 60.0	↓5.1 2.7
	H-SAL	↓1.7 68.0	↓9.6 0.3	↓2.2 62.0	↓11.5 2.4	↓1.2 68.0	↓5.6 5.2	↓1.8 59.2	↓4.6 3.2
Engagement	Baseline	69.2	8.4	63.9	10.1	65.4	8.9	60.8	6.7
	SAL (random)	↓0.3 68.9	↓0.7 7.7	↓0.2 63.7	↓0.9 9.2	↓0.2 65.2	9.0	↓0.1 60.7	↓1.4 5.3
	SAL	↓1.3 67.9	↓7.9 0.5	↓1.3 62.5	↓7.5 2.6	↓0.7 64.8	↓5.3 3.6	↓0.9 59.9	↓5.6 1.2
	H-SAL	↓2.2 67.0	↓7.7 0.7	↓2.8 61.0	↓8.0 2.1	↓0.9 64.6	↓5.2 3.7	↓1.7 59.1	↓1.7 5.0

Table 7: **Llama-3.1-8B results under the main erasure setting.** Results follow the same user-disjoint evaluation protocol and method definitions as Table 6.

Bias Type	Method	StackOverflow		Mathematics		SuperUser		English	
		Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
DisplayName	Baseline	69.5	4.3	64.2	4.4	69.2	3.5	61.3	11.2
	SAL (random)	↓0.2 69.3	↓0.1 4.2	↓0.3 63.9	↑0.8 5.2	↑0.2 69.4	↓0.7 2.8	↓0.3 61.0	↓2.4 8.8
	SAL	↓0.8 68.7	↓2.4 1.9	↓1.1 63.1	↓1.8 2.7	↓0.3 68.8	↓1.1 2.3	↓2.1 59.2	↑0.6 11.8
	H-SAL	↓1.8 67.8	↓2.7 1.6	↓2.3 62.0	↓2.3 2.1	↓0.6 68.5	↓1.6 1.9	↓2.3 59.0	↓4.3 6.9
Gender	Baseline	57.8	0.4	52.7	3.5	59.6	15.2	53.9	9.3
	SAL (random)	↓1.2 56.6	↓0.1 0.2	↑1.5 54.2	↓1.0 2.5	↓1.3 58.2	↓0.9 14.4	↓1.5 52.4	↑2.3 11.6
	SAL	↓1.6 56.1	↑0.5 0.8	↑0.4 53.2	↓0.6 2.9	↑0.5 60.1	15.2	↑0.2 54.1	↑0.1 9.4
	H-SAL	↓3.0 54.8	0.3	↑0.6 53.3	↓1.9 1.6	↑0.8 60.4	↓4.0 11.2	53.9	↑0.3 9.6
Country	Baseline	67.8	10.9	61.4	7.5	63.3	9.1	58.9	10.6
	SAL (random)	↓0.3 67.5	↓0.6 10.3	↑0.1 61.6	↑0.4 7.9	↓0.6 62.7	↑0.8 9.9	↓0.4 58.6	↓0.7 9.9
	SAL	↓0.9 66.9	↓7.2 3.8	↓0.2 61.3	↓2.7 4.9	↓0.4 62.9	↓4.9 4.2	↓0.2 58.7	↑0.3 10.9
	H-SAL	↓2.5 65.3	↓8.3 2.6	↓3.8 57.6	↓4.2 3.3	↓1.3 62.0	↓4.6 4.5	↓2.2 56.7	↑0.4 11.0
Reputation	Baseline	69.5	9.7	64.2	14.2	69.2	8.9	61.3	9.2
	SAL (random)	↓0.2 69.3	↓0.6 9.1	↓0.3 63.9	↓0.9 13.3	↑0.2 69.4	8.8	↓0.3 61.0	↑0.4 9.5
	SAL	↓0.8 68.7	↓9.6 0.1	↓1.4 62.9	↓11.5 2.7	↓0.4 68.7	↓4.5 4.4	↓1.2 60.1	↓3.4 5.8
	H-SAL	↓1.8 67.8	↓9.2 0.5	↓2.3 62.0	↓12.0 2.2	↓0.6 68.5	↓7.0 1.8	↓2.3 59.0	↓2.0 7.2
Engagement	Baseline	69.1	8.5	64.0	9.5	66.6	10.5	62.4	8.6
	SAL (random)	↓0.2 68.8	↓1.2 7.3	↓0.1 63.9	↓0.6 8.9	↓0.1 66.4	↓1.3 9.2	↓1.2 61.2	↓2.6 6.0
	SAL	↓1.3 67.8	↓8.0 0.5	↓1.4 62.7	↓7.0 2.5	↓0.6 65.9	↓7.1 3.4	↓2.7 59.7	↓5.4 3.3
	H-SAL	↓2.2 66.9	↓7.8 0.8	↓2.8 61.2	↓6.9 2.6	↓1.3 65.2	↓6.0 4.5	↓2.8 59.6	↓1.4 7.2

Table 8: **Mistral-7B results under the main erasure setting.** Results follow the same user-disjoint evaluation protocol and method definitions as Table 6.

Bias Type	Method	StackOverflow		Mathematics		SuperUser		English	
		Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
DisplayName	Baseline	69.2	4.9	64.0	4.8	71.1	4.1	63.5	9.9
	SAL (random)	69.2	↓0.1 4.8	64.0	4.8	↓0.1 71.0	↓0.1 4.1	↓0.2 63.3	↑0.2 10.1
	SAL	↓0.8 68.4	↓2.9 2.0	↓1.1 62.8	↓2.0 2.8	↓0.3 70.9	↑0.3 4.4	↓0.8 62.7	↓0.8 9.1
	H-SAL	↓1.3 67.9	↓3.3 1.6	↓1.7 62.3	↓3.0 1.8	↓0.5 70.6	↓0.4 3.7	↓0.6 62.9	↓1.5 8.4
Gender	Baseline	56.7	0.3	54.7	2.8	58.8	11.6	54.3	9.9
	SAL (random)	↓0.6 56.1	↑1.0 1.3	↑0.5 55.2	↓0.1 2.8	↓1.0 57.8	↑0.1 11.7	↓0.3 54.0	↑1.4 11.3
	SAL	↓0.5 56.1	↑0.5 0.8	↑0.4 55.1	↓0.1 2.8	↑0.4 59.1	11.6	↑0.1 54.4	↓0.1 9.8
	H-SAL	↓1.4 55.3	↑0.8 1.1	↓0.7 54.0	↑0.6 3.5	↑0.8 59.6	↑0.8 12.4	↓0.4 53.9	↓1.4 8.6
Country	Baseline	67.4	11.5	61.0	9.7	63.4	7.0	59.3	10.1
	SAL (random)	67.4	↑0.1 11.5	61.0	↓0.1 9.7	63.4	↓0.1 6.9	↑0.1 59.4	↑0.4 10.5
	SAL	↓0.6 66.8	↓8.1 3.4	↓0.1 60.9	↓2.7 7.1	↓0.3 63.1	↓3.9 3.0	↑0.1 59.5	↓0.3 9.8
	H-SAL	↓1.7 65.7	↓8.3 3.2	↓2.3 58.7	↓1.3 8.4	↓0.7 62.7	↓3.3 3.7	59.4	↓2.6 7.4
Reputation	Baseline	69.2	8.6	64.0	11.8	71.1	8.8	63.5	9.8
	SAL (random)	69.2	↓0.1 8.5	64.0	↓0.2 11.6	↓0.1 71.0	↓0.2 8.5	↓0.2 63.3	↑0.2 9.9
	SAL	↓0.8 68.4	↓8.4 0.2	↓1.2 62.7	↓10.0 1.8	↓0.8 70.3	↓4.9 3.9	↓1.3 62.2	↓4.9 4.9
	H-SAL	↓1.3 67.9	↓6.5 2.1	↓1.7 62.3	↓8.0 3.8	↓0.5 70.6	↓2.6 6.2	↓0.6 62.9	↓3.1 6.7
Engagement	Baseline	68.6	7.7	63.6	8.5	67.2	8.4	63.6	9.6
	SAL (random)	↓0.1 68.5	↓0.1 7.6	63.6	8.5	↓0.1 67.1	↓0.1 8.3	63.5	↑0.5 10.1
	SAL	↓1.1 67.5	↓7.2 0.5	↓1.1 62.5	↓6.4 2.1	↓0.4 66.8	↓5.4 3.1	↓1.3 62.2	↓6.0 3.6
	H-SAL	↓1.5 67.1	↓6.1 1.5	↓1.7 61.9	↓4.9 3.5	↓0.4 66.8	↓4.0 4.4	↓1.1 62.5	↓2.6 7.0

Table 9: **Average results under the user-disjoint split with H-SAL at $k = 5$.** Results are averaged across BERT, Llama-3.1-8B, and Mistral-7B. **Main** denotes helpfulness prediction accuracy, and **TPR-Gap** denotes generalized disparity. SAL uses bias-specific ranks, while SAL (random) and H-SAL use $k = 5$ removed directions for all attributes. Other evaluation details follow Table 1.

Bias Type	Method	StackOverflow		Mathematics		SuperUser		English	
		Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
DisplayName	Baseline	69.2	4.9	64.0	4.8	71.1	4.1	63.5	9.9
	SAL (random)	↓0.1 69.1	↓0.2 4.7	64.0	4.8	↓0.1 71.0	↓0.3 3.9	↓0.5 63.0	↓0.4 9.5
	SAL	↓0.8 68.4	↓2.9 2.0	↓1.1 62.8	↓2.0 2.8	↓0.3 70.9	↑0.3 4.4	↓0.8 62.7	↓0.8 9.1
	H-SAL	↓1.4 67.8	↓3.7 1.2	↓1.8 62.1	↓2.9 1.8	↓0.6 70.5	↓0.5 3.6	↓0.8 62.7	↓3.4 6.5
Gender	Baseline	56.7	0.3	54.7	2.8	58.8	11.6	54.3	9.9
	SAL (random)	↓0.3 56.4	↑0.8 1.1	↑0.3 55.0	2.9	↓1.5 57.3	↑0.1 11.6	↓0.6 53.7	↑1.7 11.7
	SAL	↓0.5 56.1	↑0.5 0.8	↑0.4 55.1	↓0.1 2.8	↑0.4 59.1	11.6	↑0.1 54.4	↓0.1 9.8
	H-SAL	↓1.9 54.8	↑1.6 1.9	↓1.0 53.8	↓0.2 2.7	↑0.7 59.5	↓0.9 10.7	↓0.3 54.0	↓0.9 9.0
Country	Baseline	67.4	11.5	61.0	9.7	63.4	7.0	59.3	10.1
	SAL (random)	67.3	↑0.1 11.6	61.1	↓0.1 9.7	↓0.1 63.3	↑0.2 7.1	↓0.2 59.1	↓1.0 9.1
	SAL	↓0.6 66.8	↓8.1 3.4	↓0.1 60.9	↓2.7 7.1	↓0.3 63.1	↓3.9 3.0	↑0.1 59.5	↓0.3 9.8
	H-SAL	↓1.9 65.5	↓8.2 3.3	↓2.7 58.3	↓2.7 7.0	↓0.9 62.5	↓2.8 4.2	59.3	↓1.9 8.1
Reputation	Baseline	69.2	8.6	64.0	11.8	71.1	8.8	63.5	9.8
	SAL (random)	↓0.1 69.1	↓0.1 8.6	64.0	↓0.3 11.5	↓0.1 71.0	↓0.2 8.6	↓0.5 63.0	↓0.8 9.0
	SAL	↓0.8 68.4	↓8.4 0.2	↓1.2 62.7	↓10.0 1.8	↓0.8 70.3	↓4.9 3.9	↓1.3 62.2	↓4.9 4.9
	H-SAL	↓1.4 67.8	↓7.2 1.5	↓1.8 62.1	↓8.9 2.9	↓0.6 70.5	↓3.2 5.6	↓0.8 62.7	↓3.7 6.1
Engagement	Baseline	68.6	7.7	63.6	8.5	67.2	8.4	63.6	9.6
	SAL (random)	↓0.1 68.5	↓0.2 7.5	63.6	8.4	↓0.1 67.1	↓0.4 8.1	↓0.2 63.3	↓0.5 9.1
	SAL	↓1.1 67.5	↓7.2 0.5	↓1.1 62.5	↓6.4 2.1	↓0.4 66.8	↓5.4 3.1	↓1.3 62.2	↓6.0 3.6
	H-SAL	↓1.6 67.0	↓6.3 1.4	↓1.9 61.7	↓5.4 3.1	↓0.7 66.5	↓3.9 4.5	↓1.2 62.4	↓3.0 6.6

Table 10: **Average results under the user-disjoint split with H-SAL at $k = 10$.** Results follow the same model-averaged evaluation protocol and method definitions as Table 9.

Bias Type	Method	StackOverflow		Mathematics		SuperUser		English	
		Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
DisplayName	Baseline	69.2	4.9	64.0	4.8	71.1	4.1	63.5	9.9
	SAL (random)	↓0.3 68.9	↓0.6 4.3	64.0	↓0.2 4.6	↓0.1 71.0	↓0.2 3.9	↓0.6 63.0	↓1.2 8.7
	SAL	↓0.8 68.4	↓2.9 2.0	↓1.1 62.8	↓2.0 2.8	↓0.3 70.9	↑0.3 4.4	↓0.8 62.7	↓0.8 9.1
	H-SAL	↓1.5 67.7	↓3.8 1.1	↓2.0 62.0	↓3.2 1.6	↓1.0 70.1	↓0.5 3.6	↓1.1 62.4	↓3.2 6.7
Gender	Baseline	56.7	0.3	54.7	2.8	58.8	11.6	54.3	9.9
	SAL (random)	↓0.4 56.3	↑0.8 1.1	↑0.4 55.1	2.9	↓0.6 58.2	↑0.9 12.4	↓0.6 53.6	↑2.0 11.9
	SAL	↓0.5 56.1	↑0.5 0.8	↑0.4 55.1	↓0.1 2.8	↑0.4 59.1	11.6	↑0.1 54.4	↓0.1 9.8
	H-SAL	↓1.8 54.9	↑1.1 1.4	↓0.8 53.9	↓1.7 1.2	↓1.0 57.8	↑0.2 11.8	↓0.7 53.6	↓1.7 8.2
Country	Baseline	67.4	11.5	61.0	9.7	63.4	7.0	59.3	10.1
	SAL (random)	↓0.2 67.1	↓0.3 11.1	↓0.1 60.9	↓0.4 9.3	↓0.1 63.3	↓0.2 6.7	59.4	↓1.8 8.3
	SAL	↓0.6 66.8	↓8.1 3.4	↓0.1 60.9	↓2.7 7.1	↓0.3 63.1	↓3.9 3.0	↑0.1 59.5	↓0.3 9.8
	H-SAL	↓2.0 65.4	↓9.1 2.4	↓3.1 57.9	↓5.6 4.1	↓0.9 62.5	↓3.0 4.0	↓0.9 58.4	↓2.6 7.4
Reputation	Baseline	69.2	8.6	64.0	11.8	71.1	8.8	63.5	9.8
	SAL (random)	↓0.3 68.9	↓0.6 8.0	64.0	↓0.5 11.4	↓0.1 71.0	8.7	↓0.6 63.0	↓0.6 9.2
	SAL	↓0.8 68.4	↓8.4 0.2	↓1.2 62.7	↓10.0 1.8	↓0.8 70.3	↓4.9 3.9	↓1.3 62.2	↓4.9 4.9
	H-SAL	↓1.5 67.7	↓7.9 0.7	↓2.0 62.0	↓9.7 2.1	↓0.9 70.2	↓4.4 4.3	↓1.1 62.4	↓4.4 5.3
Engagement	Baseline	68.6	7.7	63.6	8.5	67.2	8.4	63.6	9.6
	SAL (random)	↓0.2 68.4	↓0.7 7.0	↓0.1 63.5	↓0.5 8.0	67.2	↓0.3 8.1	↓0.3 63.3	↓0.3 9.3
	SAL	↓1.1 67.5	↓7.2 0.5	↓1.1 62.5	↓6.4 2.1	↓0.4 66.8	↓5.4 3.1	↓1.3 62.2	↓6.0 3.6
	H-SAL	↓1.8 66.8	↓7.1 0.6	↓2.3 61.3	↓6.3 2.2	↓0.7 66.5	↓4.5 3.9	↓1.4 62.2	↓2.6 7.0

Table 11: **Average results under the user-disjoint split with H-SAL at $k = 24$.** Results follow the same model-averaged evaluation protocol and method definitions as Table 9.

Bias Type	Method	StackOverflow		Mathematics		SuperUser		English	
		Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
DisplayName	Baseline	69.2	4.9	64.0	4.8	71.1	4.1	63.5	9.9
	SAL (random)	↓0.4 68.8	↓0.7 4.1	64.0	↓0.7 4.1	71.1	↑0.3 4.5	↓0.7 62.8	↓1.7 8.2
	SAL	↓0.8 68.4	↓2.9 2.0	↓1.1 62.8	↓2.0 2.8	↓0.3 70.9	↑0.3 4.4	↓0.8 62.7	↓0.8 9.1
	H-SAL	↓1.6 67.6	↓3.5 1.4	↓1.8 62.1	↓3.2 1.6	↓1.1 70.1	↓0.7 3.4	↓1.8 61.7	↓4.3 5.6
Gender	Baseline	56.7	0.3	54.7	2.8	58.8	11.6	54.3	9.9
	SAL (random)	↓0.9 55.8	↑0.9 1.2	↑0.3 55.0	↑0.1 2.9	↓0.9 57.9	↓2.0 9.6	↓0.9 53.4	↑2.5 12.4
	SAL	↓0.5 56.1	↑0.5 0.8	↑0.4 55.1	↓0.1 2.8	↑0.4 59.1	11.6	↑0.1 54.4	↓0.1 9.8
	H-SAL	↓2.3 54.4	↑1.1 1.4	↓1.3 53.4	↓1.2 1.6	↓2.0 56.8	↓0.5 11.1	↓1.1 53.1	9.9
Country	Baseline	67.4	11.5	61.0	9.7	63.4	7.0	59.3	10.1
	SAL (random)	↓0.4 67.0	↓0.8 10.7	↓0.1 60.9	↓0.4 9.3	↓0.2 63.2	↓0.2 6.8	59.4	↓1.2 8.8
	SAL	↓0.6 66.8	↓8.1 3.4	↓0.1 60.9	↓2.7 7.1	↓0.3 63.1	↓3.9 3.0	↑0.1 59.5	↓0.3 9.8
	H-SAL	↓2.3 65.1	↓9.3 2.2	↓3.7 57.3	↓5.9 3.8	↓0.8 62.6	↓3.1 3.9	↓1.5 57.8	↓1.2 8.9
Reputation	Baseline	69.2	8.6	64.0	11.8	71.1	8.8	63.5	9.8
	SAL (random)	↓0.4 68.8	↓0.8 7.8	64.0	↓0.7 11.1	71.1	↑0.2 8.9	↓0.7 62.8	↓1.4 8.3
	SAL	↓0.8 68.4	↓8.4 0.2	↓1.2 62.7	↓10.0 1.8	↓0.8 70.3	↓4.9 3.9	↓1.3 62.2	↓4.9 4.9
	H-SAL	↓1.6 67.6	↓8.3 0.3	↓1.8 62.1	↓9.9 2.0	↓1.1 70.1	↓5.1 3.7	↓1.8 61.7	↓4.8 5.0
Engagement	Baseline	68.6	7.7	63.6	8.5	67.2	8.4	63.6	9.6
	SAL (random)	↓0.4 68.2	↓1.1 6.6	↓0.2 63.5	↓0.8 7.7	↓0.2 67.0	↓0.3 8.1	↓0.8 62.8	↓0.8 8.8
	SAL	↓1.1 67.5	↓7.2 0.5	↓1.1 62.5	↓6.4 2.1	↓0.4 66.8	↓5.4 3.1	↓1.3 62.2	↓6.0 3.6
	H-SAL	↓2.0 66.6	↓7.2 0.5	↓2.5 61.1	↓6.6 1.8	↓0.8 66.4	↓4.8 3.6	↓1.8 61.8	↓2.4 7.2

Table 12: **Average results under the user-disjoint split with H-SAL at $k = 45$.** Results follow the same model-averaged evaluation protocol and method definitions as Table 9.