Intervention Lens: From Representation Surgery to String Counterfactuals

Anonymous ACL submission

Abstract

Interventions targeting the representation space of language models (LMs) have emerged as an effective means to influence model behavior. Such methods are employed, for example, to eliminate or alter the encoding of demographic information such as gender within the model's representations and, in so doing, create a counterfactual representation. However, because the intervention operates within the representation space, understanding precisely what aspects of the text it modifies poses a challenge. In this paper, we give a method to convert representation counterfactuals into string counterfactuals. We demonstrate that this approach enables us to analyze the linguistic alterations corresponding to a given representation space intervention and to interpret the features utilized to encode a specific concept. Moreover, the resulting counterfactuals can be used to mitigate bias in classification through data augmentation.

1 Introduction

Interventions performed in the representation space of LMs (generally $\mathbb{R}^D$) have proven to be effective in understanding and exerting control over neural language models (Ravfogel et al., 2020, 2021; Geva et al., 2021; Elazar et al., 2021; Ravfogel et al., 2022, 2023; Belrose et al., 2023b; Li et al., 2023). One popular set of techniques **erases** the linear subspace associated with human-interpretable concepts, e.g., GENDER or SENTIMENT. Another widely used approach is to **steer** representations from one class to another, e.g. shifting it toward a region in the representation space associated with a different class (Subramani et al., 2022; Li et al., 2023; Ravfogel et al., 2021; Singh et al., 2024). For instance, they could steer a representation into a region in the representation space associated with negative sentiment, thereby creating *counterfactual representations*. In this paper, we propose a technique to derive the string instantiations that correspond to representation-level counterfactuals, which we denote as *string counterfactuals*.

Collectively, we refer to representation space intervention techniques as **representation surgery** because they (surgically) intervene on the encoding of a concept in the representation while keeping the rest of the representation as similar as possible. In this sense, representation surgery resembles a causal intervention (Vig et al., 2020; Geiger et al., 2021; Feder et al., 2021; Geiger et al., 2022; Guerner et al., 2023; Lemberger and Saillenfest, 2024) and we will (informally) use causal language throughout the paper, referring to such modifications in the representation space as interventions. In notation, we write $f_{c \to c'} : \mathbb{R}^D \to \mathbb{R}^D$ for a function that performs such an intervention.

While representation surgery techniques can create counterfactual variants of the original representations, they do not create them at the level of natural language text. In this work, we tackle the problem of generating the counterfactual *string* that corresponds to a specific representation intervention. Despite the plethora of research on represen-

Original Biography

Providing legal representation in Florida for a variety of different issues , Barry Wax was selected to Super Lawyers for 2017 - 2018 . He is admitted to practice before the courts in Florida .

MiMiC M➞F

In 2016, Ms. Wax was selected by her peers to be selected by Florida Super Lawyers as a Florida Super Lawyer. She represents clients in a variety of practice areas, including labor and employment, real estate, bankruptcy, and family law. Barris is

LEACE M➞∅

In 2018, Barry Wax was selected as a Super Lawyer in Florida. His practice focuses on providing legal representation to clients in a variety of practice areas including :

MiMiC+ M➞F

In 2016, Ms. Barry was selected by her peers to be selected by the Florida Super Lawyers. She represents clients in all areas of family law, including but not limited to : legal malpractice, spousal care, child custody, and legal malpractice...

Figure 1: String counterfactuals induced by different representation-space interventions.

tation surgery, translating such interventions into string counterfactuals is a less-studied problem. We refer to this process as **intervention lens**, as it allows us to interpret representation space interventions in natural language, akin to representation-level interpretability techniques (Meng et al., 2022; nostalgebraist, 2020; Belrose et al., 2023a; Ghandeharioun et al., 2024). Constructing string counterfactuals serves various practical purposes. First, it serves as an approach for **meta-interpretability**, aiding in the interpretation of commonly employed representational intervention techniques, which are often used for interpretability. By pulling representational interventions back to the string, we can observe the lexical and higher-level semantic shifts triggered by the intervention. Second, we believe that string counterfactuals are a natural choice for data augmentation. Indeed, we showcase their potential to address fairness concerns in a real-world classification problem.

We follow Morris et al. (2023) in developing an approach for generating string counterfactuals from representation interventions. Let Σ be an alphabet. Consider a neural network that performs a mapping from a string $\boldsymbol{s} \in \Sigma^*$ to a representation $\mathbf{h} = \mathsf{enc}(\boldsymbol{s}) \in \mathbb{R}^D$. Morris et al. (2023) propose an iterative algorithm to approximate the inverse function $\mathsf{enc}^{-1} \colon \mathbb{R}^D \to \Sigma^*$. We exploit the Morris et al.'s (2023) algorithm to construct a *string* counterfactual that constructs a string that corresponds to a surgical intervention in the representation space. Using the notation introduced so far, we are interested in computing $\boldsymbol{s}' = \mathsf{enc}^{-1}(f(\mathsf{enc}(\boldsymbol{s})))$. To the extent that enc^{-1} constitutes a suitable inverse, we expect $\boldsymbol{s}'$ to be a minimally different version of $\boldsymbol{s}$ that reflects the difference between $\mathbf{h}$ and $\mathbf{h}'$ reflected in the representation space.

We perform experiments on a dataset of short biographies, annotated with gender and profession (De-Arteaga et al., 2019). We find that swapping the gender in the representation space and then generating the inverse is an effective method of producing string counterfactuals. The counterfactuals exhibit some degree of gender bias, for example, a preference for including words with a professional association in male biographies, providing evidence that LMs encode subtle cues that correlate with gender, beyond pronouns (§4.2). We further show that counterfactuals can be used for data augmentation to increase fairness in a multiclass classification (§4.2.2), i.e., we show that profession

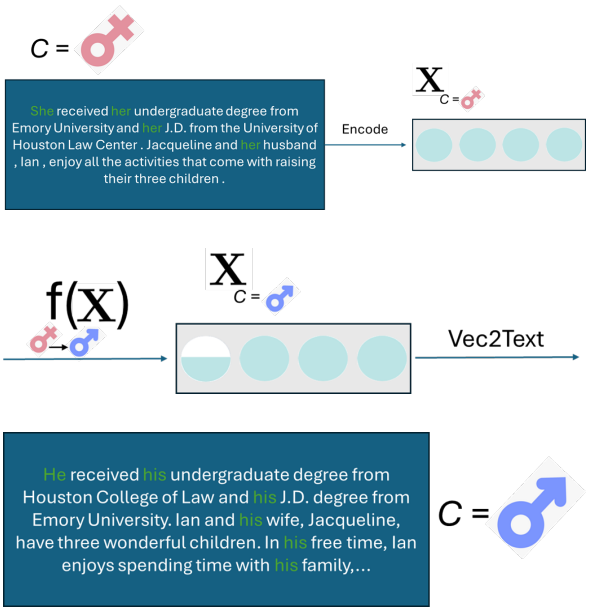

Figure 2: **Intervention lens** overview, generating natural language counterfactuals through representation surgery.

classifiers trained in a dataset comprising both original and counterfactual biographies (with respect to gender) exhibit reduced gender bias compared to classifiers trained solely on original biographies.

2 Representation Surgery

We give a more in-depth overview of representation surgery. Many neural networks for natural language processing construct a function $\mathsf{enc} \colon \Sigma^* \to \mathbb{R}^D$ that maps a string of symbols over Σ, e.g., a natural language text, to a real-valued representation in $\mathbb{R}^D$. We call such functions **language encoders**. In §1, we introduce a function $f \colon \mathbb{R}^D \to \mathbb{R}^D$ that performs the intervention in the representation space. We consider three types of representation interventions, each discussed in a labeled paragraph below. First, however, we will introduce some general notation.

Notation. Let p be a language model,[1] i.e., a distribution over Σ^*, let $\mathsf{enc} \colon \Sigma^* \to \mathbb{R}^D$ be a language encoder. Let $\mathcal{C} = \{0, 1\}$ be a binary set that stands for the different values for a concept. Binary concepts denote whether a given

[1] In this text, p is fully decoupled from the language encoder enc. For instance, our notation allows for p to some approximation to or the actual human language model (to the extent one believes in the human language model as a construct). However, we also allow p to be deeply related to enc. For instance, in an autoregressive language model, enc could be produced by the representation of EOS.

property is present or not, e.g., whether or not a string $\boldsymbol{s} \in \Sigma^*$ is a biography of a man or of a woman. Furthermore, let $\phi\colon \Sigma^* \to \mathcal{C}$ be a concept encoding function.[2] We define the distribution

$$p_c(\boldsymbol{s}) \stackrel{\text{def}}{=} p(\boldsymbol{s} \mid C = c) \propto p(\boldsymbol{s})\mathbb{1}\{\phi(\boldsymbol{s}) = c\}. \quad (1)$$

Then, for each $c \in \mathcal{C}$, define the following $\mathbb{R}^D$-valued random variable

$$\mathbf{X}_c(\boldsymbol{s}) = \mathsf{enc}(\boldsymbol{s})\colon \Sigma^* \to \mathbb{R}^D, \quad (2)$$

which is distributed according to

$$\begin{aligned}\mathbb{P}(\mathbf{X}_c = \mathbf{h}) &= \mathbb{P}(\mathbf{X}_c^{-1}(\mathbf{h})) \\ &= \sum_{\boldsymbol{s}\in\Sigma^*} p_c(\boldsymbol{s})\mathbb{1}\{\mathbf{h} = \mathsf{enc}(\boldsymbol{s})\}. \end{aligned} \quad (3)$$

LEACE (Belrose et al., 2023b). LEACE is a spectral algorithm that induces log-linear guardedness (Ravfogel et al., 2023), i.e., it minimally (in the L_2 sense) changes the $\mathbb{R}^D$-valued random variables $\mathbf{X}_c$ for all $c \in \mathcal{C}$ such that there does not exist a log-linear classifier that operates at better than the accuracy of the majority class. To achieve guardedness, LEACE finds an oblique projection matrix $\mathbf{P}$ of rank $|\mathcal{C}| - 1$ and a translation vector $\mathbf{b}$, which are then used to define the following intervention function

$$f^{\text{L}}_{\mathcal{C}\to\emptyset}(\mathbf{X}_c) \stackrel{\text{def}}{=} \mathbf{P}\mathbf{X}_c + \mathbf{b}. \quad (4)$$

MiMiC (Singh et al., 2024). MiMiC, in contrast to LEACE, does not merely erase the target concept from the representations, but rather takes the representations of one class (e.g., MSC), and minimally changes it such that it resembles the representations of the other class (e.g., FEM). More precisely, it equates the first two moments of the *source* class-conditional distribution to the *destination* class-conditional distribution, i.e., MiMiC finds a function $f^{\text{M}}_{c\to c'}$ such that

$$\mathbb{E}\left[f^{\text{M}}_{c\to c'}(\mathbf{X}_c)\right] = \mathbb{E}\left[\mathbf{X}_{c'}\right] \quad (5a)$$

$$\mathbb{V}\left[f^{\text{M}}_{c\to c'}(\mathbf{X}_c)\right] = \mathbb{V}\left[\mathbf{X}_{c'}\right]. \quad (5b)$$

In the case where the random variables $\mathbf{X}_c$ and $\mathbf{X}_{c'}$ are Gaussian distributed, MiMiC guarantees that the Wasserstein distance (Kantorovich, 1960) between $\mathbf{X}_c$ and $\mathbf{X}_{c'}$ is minimized. In this case, the distance will be zero.

[2] We (simplistically) assume each string $\boldsymbol{s}$ contains exactly one concept. Future work will relax this assumption.

MiMiC+. With MiMiC+, we further push the representations in the direction that connects the class-conditional means of the representations belonging to the two classes. Let $\mathbf{v} \stackrel{\text{def}}{=} \mathbb{E}\left[\mathbf{X}_c\right] - \mathbb{E}\left[\mathbf{X}_{c'}\right]$. Given a representation $\mathbf{X}_c(\boldsymbol{s})$, we linearly transform the output of MiMiC+ as follows

$$f^{\text{M+}}_{c\to c'}(\mathbf{X}_c) \stackrel{\text{def}}{=} f^{\text{M}}_{c\to c'}(\mathbf{X}_c) + \alpha\mathbf{v}, \quad (6)$$

where α is a non-negative scalar. Intuitively, we move the representations towards the mean of $\mathbf{X}_c$.

3 Representation Inversion

The generative process through which natural language text is created is complex and difficult to model. However, in some respects, it is well-approximated by language models. Concepts like gender are often conveyed subtly, and merely modifying overt indicators such as pronouns and names may not suffice (Maudslay et al., 2019). Instead, we leverage the fact that neural encoders capture nuanced manner in which these concepts manifest in texts. Intervening in such representations is feasible, even *without* the ability to enumerate or fully understand all linguistic features relevant to a concept. Using representational surgery, we intervene on a concept encoded in the representation generated by an encoder after the intervention, we apply an inverter model $\mathsf{enc}^{-1}(\cdot)$ that maps the representation back to a string, yielding an approximate string counterfactual $\boldsymbol{s}' = \mathsf{enc}^{-1}(f(\mathsf{enc}(\boldsymbol{s})))$.

Morris et al. (2023). Let $\boldsymbol{s} \in \Sigma^*$ be a sentence and let $\mathsf{enc}(\boldsymbol{s})$ be its representation. Our goal is to convert $\mathsf{enc}(\boldsymbol{s})$ back into a string. Morris et al.'s (2023) method starts by fine-tuning a language model that can be used to reconstruct an initial hypothesis $\hat{\boldsymbol{s}}_0$ of the inverse $\mathsf{enc}^{-1}(\boldsymbol{s})$ given the representation $\mathsf{enc}(\boldsymbol{s})$. Then, a second language model is fine-tuned to reconstruct another hypothesis $\hat{\boldsymbol{s}}_1$ conditioned on the initial $\hat{\boldsymbol{s}}_0$, $\mathsf{enc}(\hat{\boldsymbol{s}}_0)$, $\mathsf{enc}(\boldsymbol{s})$ and the difference vector $\mathsf{enc}(\boldsymbol{s}) - \mathsf{enc}(\hat{\boldsymbol{s}}_0)$. This process is repeated K times—each time $k \in [K]$, the step consists of fine-tuning the second language model conditioned on $\hat{\boldsymbol{s}}_{k-1}$, $\mathsf{enc}(\hat{\boldsymbol{s}}_{k-1})$, $\mathsf{enc}(\boldsymbol{s})$ and the difference vector $\mathsf{enc}(\boldsymbol{s}) - \mathsf{enc}(\hat{\boldsymbol{s}}_{k-1})$. The procedure ends when $\mathsf{enc}(\hat{\boldsymbol{s}}_k)$ is sufficiently close to $\mathsf{enc}(\boldsymbol{s})$ or the computational budget is exceeded. Then, $\hat{\boldsymbol{s}}_k$ is returned by the method as the inverse $\mathsf{enc}^{-1}(\boldsymbol{s})$. Empirically, Morris et al. (2023) find $K > 1$ iterations produces a more faithful inverse.

Putting it all together. Now, for a concept $c \in \mathcal{C}$ and an intervention function f_c that intervenes on that concept, we generate a counterfactual string by taking the inverse of the encoding of the string, post-intervention. Formally, the counterfactual string correspond to the following Σ^*-valued random variable:

$$\boldsymbol{R}_{c\to c'}(\boldsymbol{s}) = \mathsf{enc}^{-1}(f_{c\to c'}(\mathsf{enc}(\boldsymbol{s})), \quad (7)$$

which is distributed according to

$$\begin{aligned} p_{c\to c'}(\boldsymbol{s}') &= \mathbb{P}(\boldsymbol{R}^{-1}_{c\to c'}(\boldsymbol{s}')) \quad (8) \\ &= \sum_{\boldsymbol{s}\in\Sigma^*} p_c(\boldsymbol{s})\mathbb{1}\{\boldsymbol{s}' = \mathsf{enc}^{-1}(f_{c\to c'}(\mathsf{enc}(\boldsymbol{s}))\}. \end{aligned}$$

4 Experimental Evaluation

We now present our experimental results, focusing on gender-based interventions that modify the gender attribute in short biographical texts. We evaluate the semantic changes induced by these counterfactual interventions (§4.2), assess their quality (§4.1) and demonstrate that the resulting counterfactual texts are effective in mitigating gender bias (§4.2.2).

Inversion model. We train a variant of the inversion model of Morris et al. (2023) over sequences of 64 tokens and further fine-tune it on the BiasBios dataset. See Appendix A for more details.

Dataset We conduct experiments on the BiasInBios dataset (De-Arteaga et al., 2019), a large collection of short biographies sourced from the Internet. Each biography in this dataset is annotated with the subject's gender and their profession.[3] We create natural language counterfactuals after intervention in the encoding of the gender property. Our string counterfactuals are then being used to study the way gender is encoded in the LM in which we intervene (§4.2), and as a data augmentation tool to mitigate bias (§4.2.2).

Pipeline implementation We trained a dedicated inversion model (Morris et al., 2023) on the representations of biographies extracted from the last layer of GTR-base model (Ni et al., 2022), obtained by averaging words embedding into a single paragraph representation. After creating a high-quality inversion model, we applied one of the intervention methods to the dataset of extracted biographies representations to obtain representation-level counterfactuals. For MiMiC and MiMiC+ interventions we set the regularization term to be 10^{-5}, furthermore we set $\alpha = 2$ for MiMiC+. Lastly, we applied the trained inversion model to the intervened representations to yield the desired string counterfactuals. Note that although the inversion model is also a GTR-base model, it is not a prerequisite of the method, and any model could be employed to create the biographies representation, as demonstrated by (Chen et al., 2024). For further explanations on the inversion model training setup check Appendix A.

4.1 Evaluating Counterfactuals Quality

We assess the semantic and grammatical coherent of the generated counterfactuals, and their correctness.

4.1.1 Automatic evaluation

To assess the quality of the generated counterfactuals, we computed the average perplexity of the resulting texts for each intervention technique. Perplexity is a standard measure of LM performance, with lower values indicating that the model finds the text more predictable, and thus of higher quality in terms of fluency. We also calculated perplexity for the original biographies and for the reconstructed biographies, without any intervention (i.e., applying the inversion process of Morris et al. (2023) without any intervention). The latter establishes a baseline for the degradation induced by the inversion process itself.

The results, shown in Tab. 1, reveal that the reconstructed biographies (without intervention) consistently achieve lower perplexity than the original biographies, suggesting that the reconstruction process itself simplifies the text and makes it more predictable. Moreover, the counterfactuals generated by the three intervention methods (LEACE, MiMiC, MiMiC+) show only a small increase in perplexity compared to the reconstructed biographies. This indicates that the interventions introduce minimal degradation in fluency, effectively preserving the overall quality of the generated text.

4.1.2 Human evaluation

We conducted human annotation experiments using Amazon Mechanical Turk (MTurk) to evaluate the quality of the counterfactuals and the effectiveness of our method, as described in detail in Appendix C. Five annotators, all native English speakers from the US, UK, and Australia, were

[3] The dataset contains 28 distinct professions.

Scenario	Mistral7b	GPT-2
Original biographies	22.62	104.67
Reconstructed biographies (no intervention)	18.17	52.58
LEACE counterfactuals	18.82	53.42
MiMiC counterfactuals	18.29	51.55
MiMiC+ counterfactuals	19.14	48.84

Table 1: Average Perplexity for Original, Reconstructed, and Counterfactual Biographies using the different intervention techniques generated by `Mistral7b` and `GPT-2` (Jiang et al., 2023; Radford et al., 2019).

recruited and compensated for their participation. The annotators were tasked with assessing three aspects of the generated texts: (1) readability, (2) grammatical correctness, and (3) gender specification of the subject entity. The first two aim to measure the *quality* of the counterfactual strings, while the latter aims to measure its *correctness*, i.e., whether we successfully intervened in the concept of interest.

For tasks (1) and (2), annotators were presented with pairs of texts (Original and counterfactual) and asked to compare them in terms of readability and grammatical correctness, indicating which text was better or if they were comparable. For task (3), annotators were asked to determine the gender of the subject entity in each text (male, female, or unclear).

Quality. We performed statistical tests to evaluate whether the interventions had a significant effect on the annotators' responses regarding readability and grammatical correctness. The results are summarized in Appendix C (Tab. 5) based on Tab. 3 and Tab. 4. For most interventions (LEACE, MiMiC, and MiMiC+), the p-values from the one-tailed binomial tests for readability and grammar were greater than 0.05, indicating no significant preference for the original text over the counterfactuals. This suggests that our method did not degrade the quality of the text in terms of readability and grammar. However, for the MiMiC - FEM $\rightarrow$ MSC and MiMiC+ - FEM $\rightarrow$ MSC interventions, the p-values for readability were significantly less than 0.05 ($p = 1.60 \times 10^{-5}$ and $p = 9.05 \times 10^{-8}$, respectively), indicating that the original text was preferred over the counterfactual in terms of readability. This suggests that these interventions may have caused a degradation in readability when altering gender from female to male.

Correctness. To assess whether the interventions effectively altered the perceived gender in the texts, we analyzed the annotators' gender specification responses using chi-square tests of independence. A significant chi-square statistic indicates that the distribution of gender identification is dependent on the intervention, meaning that the intervention successfully changed how annotators perceived the gender of the subject entity. The results, shown in Table 5, indicate that for all interventions, the p-values were extremely low (well below 0.05), and the chi-square values were significantly high, demonstrating a strong effect of the interventions.

For example, in the MiMiC - FEM $\rightarrow$ MSC intervention, originally female biographies were annotated as male after the intervention in 82% of the cases, compared to only 3% male in the original texts. This significant shift is reflected in the chi-square statistic of 130.56 (p-value 4.45×10^{-29}). Similarly, in the MiMiC - MSC $\rightarrow$ FEM intervention, originally male biographies were annotated as female after the intervention in 86% of the cases, compared to only 10% female in the original texts, with a chi-square statistic of 131.44 (p-value 2.87×10^{-29}). These results demonstrate that annotators found the gender interventions were correct and provide clear evidence that the interventions were effective in altering gender representation within the texts.

In contrast, LEACE being an erasure method, resulted in mixed outcomes. For instance, when LEACE was applied to originally female biographies, the proportion of texts perceived as male increased from 11% to 66%, and those perceived as female decreased from 85% to 26%, with 8% being labeled as unclear. This reflects its nature as an erasure technique rather than a steering method (see Fig. 1 and Appendix D).

4.2 Semantic Changes in the Counterfactuals

In the previous section, we validated the semantic coherence and correctness of the counterfactuals. In this section, we analyze the specific changes incurred in the inversion process. This analysis is performed over sentences from the dev set of BiasBios dataset whose lengths are 64 tokens or less: 7,578 biographies in the MSC $\rightarrow$ FEM direction and 6,982 biographies in the FEM $\rightarrow$ MSC direction.

Pointwise Mutual Information. To quantitatively evaluate the word-level changes induced by the counterfactual generation process, we analyze

Word Likelihood Changes in Different Methods

	Increased	Decreased
MiMiC MSC → FEM	*"ms", "she's", "bri", "marie", "mrs", "girl", "herself", "jennifer", "nicole", "domestic", "anne", "nancy", "maternal"*	*"et", "himself", "kau", "enterprise", "prof", "anthony", "edward", "iot", "acoustic", "days", "hardware", "late"*
MiMiC FEM → MSC	*"mr", "him", "he's", "himself", "developer", "chris", "robert", "veterinary", "stephen"*	*"she's", "mrs", "girl", "herself", "nicole", "female", "desire", "abuse", "lingerie"*
LEACE (originally FEMALE**)**	*"him", "he's", "mr", "plays", "showcase", "authority", "pleasure", "watch", "adventure"*	*"clutter", "uncomfortable", "classrooms", "compassion", "experiencing", "participant", "babies", "engaging"*
LEACE (originally MALE**)**	*"ms", "colleagues", "grace", "leaders", "happy", "presenter", "she's", "advocates", "teach"*	*"et", "elite", "kau", "direction", "theater", "mentor", "hollywood", "photojournalism"*
MiMiC+ MSC → FEM	*"ms", "women's", "she's", "marie", "maternal", "girl", "female", "nicole", "elizabeth", "maternity", "joy"*	*"he", "his", "mr", "him", "michael", "robert", "daniel", "charles", "peter"*
MiMiC+ FEM → MSC	*"he's", "mr", "him", "developer", "daniel", "robert", "jeremy", "adam", "plays"*	*"she", "her", "ms", "women", "mary", "jennifer", "marie", "herself", "joy"*

Figure 3: Word log likelihood ratio changes in Different methods. We present the words whose log-ratio increased or decreased the most following the intevention.

the words whose probabilities change the most between the original and counterfactual sentences. Let $c, c' \in \mathcal{C}$ be two concepts. In the case of concept *erasure*, we may have $c' = \emptyset \notin \mathcal{C}$. We now considered two random multisets of M strings

$$S_{c\to c} = \wr \boldsymbol{s}^{(m)} \mid \boldsymbol{s} \sim p_{c\to c} \wr_{m=1}^{M} \tag{9a}$$

$$S_{c\to c'} = \wr \boldsymbol{s}^{(m)} \mid \boldsymbol{s} \sim p_{c\to c'} \wr_{m=1}^{M}. \tag{9b}$$

Then, we define a unigram distribution over Σ induced from $S_{c\to c}$ and $S_{c\to c'}$ as follows

$$p(w \mid c \to c) \propto \#(w \in S_{c\to c}) \tag{10a}$$

$$p(w \mid c \to c') \propto \#(w \in S_{c\to c'}). \tag{10b}$$

Then, taking $p(c \to c) = p(c \to c') = 1/2$, we define the pointwise mutual information as follows

$$\mathrm{PMI}(w, c \to c') \stackrel{\text{def}}{=} \log \frac{2p(w, c \to c')}{p(w)}. \tag{11}$$

Manipulation then reveals that the difference of two PMIs is the log odds ratio:

$$\mathrm{PMI}(w, c \to c') - \mathrm{PMI}(w, c \to c') \tag{12a}$$

$$= \log \frac{p(w, c \to c')}{p(w, c \to c)}. \tag{12b}$$

Intuitively, the difference of two PMIs identifies the words whose frequency increased or decreased the most after the intervention, normalized by the amount of change incurred by the inversion process alone, i.e., inversion without an intervention. We additionally add a smoothing term of $1e^{-6}$ when calculating the logarithmic odds ratio. Finally, we sort the vocabulary according to the log-odds ratio, omitting words whose frequency is less than 5.

4.2.1 Results

In this section, we analyze the changes in log ratios across different methods when manipulating gender concepts. See Fig. 1 and Appendix D for a sample of the original and counterfactual sentences. See also Fig. 3 for a subset of the most-changed words (according to log-likelihood ratio), with the entire lists available in Appendix B. We explore how words increase or decrease in likelihood when gendered concepts such as FEM and MSC are removed or altered, and we highlight the thematic shifts associated with these changes. For each method—LEACE, MiMiC, and MiMiC+—we find notable trends that reveal underlying gender associations in language models.

Overall trends. As anticipated, in the FEM → MSC direction, masculine pronouns and titles such as "*he's*", "*him*", "*mr*", and "*himself*" experienced the most significant increase in likelihood. Conversely, in the MSC → FEM direction, the largest changes were observed with feminine pronouns and titles like "*she's*", "*ms*", "*mrs*", and "*herself*". Beyond pronouns, we find that some more subtle changes sometimes occur, reflecting biases in the dataset. Furthermore, in the direction MSC → FEM the counterfactuals of the biographies of doctors often omit the "*Dr.*" prefix and replace it with "*Ms*". Specific terms associated with professional and technical domains, such as "*developer*", "*managers*", "*esl*", and "*llp*", exhibited an increased frequency in the FEM → MSC direction, as we discuss below.

The counterfactuals generated by MiMiC+ exhibit an overuse of stereotypical markers of the target gender, adding pronouns when they are not necessary or introducing new stereotypical information as depicted in Fig. 1. This intervention tends to significantly modify the overall structure of the sentence. The inversion process is not perfect, and at times inflicts some changes to the original text, such as paraphrasing (see §4.1).

LEACE Recall that LEACE aims to remove the ability do distinguish between male and female representations. We find that post-intervention, originally-female-focused texts now exhibit in a decrease in words related to social engagement, care, and education. This reduction is evident from terms like "*uncomfortable*", "*strangers*", "*volunteering*", and "*babies*", which are often associated with social roles and nurturing activities. Educational and experiential terms like "*seminars*", "*classrooms*", and "*participant*" also show a decreased likelihood, reflecting a diminished focus on traditionally feminine-associated educational themes.

Conversely, we witness an increase in the likelihood of words associated with masculine pronouns and themes related to action, authority, and success. Words like "*him*", "*he's*", and "*mr*" show a rise in likelihood, as well as action-oriented words such as "*adventure*", "*watch*", and "*serve*", demonstrating a shift towards traditionally masculine concepts.

Opposite trends are shown when examining originally-male-focused biographies. Those show a decrease in references to cultural, artistic, and professional domains. Words like "*elite*", "*theater*", and "*mentor*" diminish in likelihood, suggesting a reduction in masculine-associated professional and artistic spheres. Meanwhile, an increased likelihood of words related to collaboration, leadership, and personal growth is observed. Words like "*colleagues*", "*leaders*", and "*advocates*" show a rise, reflecting themes of teamwork and leadership more commonly associated with femininity. Positive emotions and personal growth terms such as "*grace*" and "*happy*" also increase, signaling a shift toward nurturing and empathetic language.

MiMiC For the MiMiC method, altering the male gender concept to female reveals a shift towards more stereotypically feminine references. Words like "*ms*", "*she's*", and "*mrs*" increase, as do female names like "*marie*", "*jennifer*", and "*nicole*". Social and emotional words, such as "*happy*" and "*colleagues*", also rise in likelihood, indicating a stronger presence of themes related to interpersonal relationships and caregiving.

When altering the female gender concept to male using MiMiC, we observe an increase in male-specific references, with words like "*mr*", "*him*", "*he's*", and "*himself*" rising in likelihood. Male names, such as "*dahl*", "*chris*", and "*stephen*", also become more prominent, along with professional and technical terms like "*developer*" and "*managers*". Conversely, terms related such as "*she's*", "*mrs*", and "*girl*", decrease in likelihood, as well as names like "*marie*", "*nicole*", "*anne*", "*stephanie*", "*susan*" and terms from the social sphere such as "*inspire*", "*uncomfortable*", "*desire*", "*strangers*", "*classrooms*". This reflects a reduced focus on female-associated themes, particularly around care and emotional expression.

In summary, across methods and gender concept manipulations, we observe clear patterns of thematic shifts. Removing or altering gender concepts in language models leads to changes in words associated with social roles, authority, and professional domains, reflecting underlying gender biases. These findings highlight the importance of understanding and addressing gender biases in language model development.

4.2.2 Counterfactual Data Augmentation

We have established that the proposed pipeline creates high quality and specific counterfactuals. In this section, we make use of the counterfactuals to increase fairness in multiclass classification.

The BiasBios dataset exhibit an imbalance in

Setting	Accuracy ↑	F1 ↑	True Positive Rate Gender Gap ↓
Original biographies	86.42±0.0	79.63±0.01	14.27±01
Reconstructed biographies (no intervention)	81.52±0.0	72.08±0.0	13.69±0.0
Biographies without gender indication	85.33±0.0	77.75±0.0	11.22±0.0
Original biographies + LEACE counterfactuals	**86.59**±0.0	**81.8**±0.0	12.95±0.0
Original biographies + MiMiC counterfactuals	86.12±0.0	80.92±0.0	12.13±0.02
Original biographies + MiMiC+ counterfactuals	85.76±0.0	80.31±0.01	**10.59**±0.01

Table 2: Multi-class classification results from a log-linear model trained on top of `roberta-base` (Liu et al., 2019).

the representation of men and women in various professions, leading to observed biases in the profession classifiers trained on this data (De-Arteaga et al., 2019). In our next experiment, we show how our generated string counterfactuals can be used for data augmentation. By adding counterfactual examples with the opposite gender label, we aim to mitigate the model's dependence on gender.

Setup. We encode each biography as the last layer representation of a GTR-base model (Ni et al., 2022). Subsequently, an intervention is applied, followed by decoding the intervened representation using the trained inversion model. A beam size of 4 was used, along with 20 correction steps using the pre-trained Natural Questions corrector from Morris et al. (2023). This iterative process was carried out for each of the three intervention techniques: LEACE, MiMiC, and MiMiC+. The results are averaged over 3 models (see Appendix A). Following previous work (De-Arteaga et al., 2019) we quantify bias by the RMS true positive rate (TPR) gap of a profession classifier between genders:

$$\text{TPR}_{\text{RMS}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\text{TPR-Gap}(\text{y}_k)^2}. \quad (13)$$

Where $\text{TPR-Gap}(\text{y}_k)$ is the difference between genders in the expected probability of predicting a given profession y, conditioned on the true profession being y. For instance, for $y =$ "*professor*", a profession classifier with high TPR gap would assign different probabilities to predicting the true class "*professor*" for males and females.

Models. We train a log-linear profession classifier on top of language model `roberta-base` (Liu et al., 2019) to predict the profession of the subject of the biography. The classifiers are trained on the original biographies D^{orig}, the inversion of the biographies without intervention D^{inv}, the biographies without gender indications, such as pronouns ("Biographies without gender indication" in Tab. 2) $\text{D}^{\text{no gender}}$, and the dataset $\text{D}^{\text{inv+orig}}_{f_{c\to c'}}$ that consists of the original biographies in addition to the corresponding counterfactuals created by LEACE MiMiC and MiMiC+ ($\alpha = 2$ for all experiments)

Results. All results are presented in Tab. 2. Classifiers trained on the augmented dataset achieve lower TPR values (better fairness), even more so than classifiers trained on the biographies after the omission of overt gender markers. At the same time, the augmentation does not damage, and even improves, main-task performance (profession classification), indicating that augmentation the dataset with intervention-induced string counterfactuals is a viable way to encourage the classifier to show invariance to the sensitive information (in out case, gender).

5 Conclusion

We introduced **intervention lens**, a method for converting representation space interventions in language models into string-level counterfactuals. This approach bridges the gap between abstract representation manipulations and concrete textual changes, and allows us to derive the latter from the former. We demonstrated that the resulting counterfactuals are semantically coherent, and that they surface some biases in the encoding of complex concepts such as gender. We additionally showed that the counterfactuals can assist in mitigating bias in classification through data augmentation.

Our experiments highlight the potential of string counterfactuals for interpreting features used to encode concepts like demographic information, with important implications for fairness in NLP. However, the quality of counterfactuals depends on the inversion model, and our focus on binary attributes could be expanded in future work. In future work, we aim to refine the inversion process and extend the method to other interventions.

Limitations

Quality of the inversion model. Our counterfactual generation method consists of two component: the intervention function f and the inversion models enc^{-1}. We aimed to disentangle these two factors in our evaluation by comparing the counterfactuals generated with the interventions with the counterfactuals without intervention. However, a complete disentanglement of these two factors is difficult, and it is possible that some of the changes we witness should be attributed to a non-perfect inversion process, rather than to the intervention. We observe that the inversion model is indeed not perfect, and often introduce slight variations in the text (e.g., by modifying numbers or geographical locations, or introducing lexical paraphrases). These changes might be undesired in certain use cases. However, an improvement in the inversion model is orthogonal to our method.

Causal interventions. As the generative process of natural language texts is opaque, we inevitably rely on markers that people often identify with the property of interest (gender) in our evaluation. Future work should use controlled, synthetic setting to test the degree to which the counterfactuals reflect the true causal factors related to the concept of interest.

Representation of gender. We inevitably use an existing dataset with binary gender labels. We acknowledge that this is a simplification, as gender is a complex nonbinary construct.

Ethical Considerations

In all scenarios involving the potential application of automated methods in real-world contexts, we strongly recommend exercising caution and conducting thorough evaluations of the representativeness of the data, its alignment with real-world phenomena, and possible adverse societal implications. Gender bias is a complex and multifaceted issue, and we view the experiments conducted in this paper as an initial exploration of strategies for mitigating the negative impacts of language models, rather than a definitive solution to real-world bias challenges. As highlighted in the limitations section, the utilization of binary gender labels arises from limitations in available data, and we anticipate that future research will facilitate more nuanced examinations of how gender, as a construct, manifests in text.

References

Marc Baboulin, Alfredo Buttari, Jack Dongarra, Jakub Kurzak, Julie Langou, Julien Langou, Piotr Luszczek, and Stanimire Tomov. 2009. Accelerating scientific computations with mixed precision algorithms. *Computer Physics Communications*, 180(12):2526–2533.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023a. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

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

Yiyi Chen, Heather Lent, and Johannes Bjerva. 2024. Text embedding inversion security for multilingual language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7808–7827, Bangkok, Thailand. 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*, page 120–128, New York, NY, USA. Association for Computing Machinery.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscope: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*.

Clément Guerner, Anej Svete, Tianyu Liu, Alexander Warstadt, and Ryan Cotterell. 2023. A geometric notion of causal probing. *arXiv preprint arXiv:2307.15054*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Leonid V. Kantorovich. 1960. Mathematical methods of organizing and planning production. *Management Science*, 6(4):366–422.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Pirmin Lemberger and Antoine Saillenfest. 2024. Explaining text classifiers with counterfactual representations. *arXiv preprint arXiv:2402.00711*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

nostalgebraist. 2020. Interpreting gpt: The logit lens.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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, Yoav Goldberg, and Ryan Cotterell. 2023. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431, Toronto, Canada. Association for Computational Linguistics.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022. Kernelized concept erasure. 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.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. 2024. MiMiC: Minimally modified

counterfactuals in the representation space. *arXiv preprint arXiv:2402.09631*.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Appendix

A Experimental setup

Training an inversion model. Morris et al. (2023) introduced an approach for converting representations into strings. To effectively invert the representations derived from the BiasBios dataset, we trained a dedicated inversion model on sequences with a length of 64 tokens from the Natural Questions dataset (Kwiatkowski et al., 2019). This decision was informed by the observation that the median length of the biography within the BiasBios dataset is 72 tokens. The model architecture is GTR-base (Ni et al., 2022) as originally used in vec2text Morris et al. (2023). The inversion process consists of two components: the inversion model and a corrector model (both are GTR-base LMs). Empirical results demonstrate that training both components improves the quality of the reconstructed text. The training procedure involved training the inversion model for 30 epochs on the Natural Questions dataset (Kwiatkowski et al., 2019) with a batch size of 4096, followed by fine-tuning for an additional 20 epochs on the BiasBios dataset (De-Arteaga et al., 2019) with a reduced batch size of 512. Afterward, the corrector model was trained on the BiasBios dataset for 10 epochs, using a batch size of 128 samples.

Training profession classifiers. To quantify the counterfactuals' causal effect on predicting an individual's profession, we utilized `roberta-base` (Liu et al., 2019) classifiers trained with the different counterfactuals and the correspondent original biographies as outlined in Tab. 2. We trained each classifier with three different seeds, reporting the mean+-std metrics of the models with the lowest validation loss. Each classifier was trained for 10 epochs on the entire BiasBios biographies with a length of up to 64 tokens. This comprised 7,578 male biographies and 6,982 female biographies. The training set for the counterfactual classifiers encompassed, for each sample, its corresponding counterfactual. We used a batch size of 1024 samples for training and 4096 for evaluation. Furthermore, 6% of the samples were used for the warm-up of the learning rate, with the initial learning rate set to $2 * 10^{-5}$. We additionally employed half-precision quantization (fp16) for the network's weights (Baboulin et al., 2009; Micikevicius et al., 2017). The results reported at Tab. 2 were calculated on the entire BiasBios dev set, i.e., 39,369 samples, truncated to a 64 tokens sequence length.

B Word PMI Analysis

We provide the words most changed due to MiMiC , LEACE and MiMiC+ interventions below.

B.1 MiMiC

- words whose likelihood most decreased in direction MSC → FEM : ["*et*", "*himself*", "*kau*", "*enterprise*", "*really*"","*prof*", "*anthony*", "*ch*", "*edward*", "*iot*", "*0560*", "*1978*", "*acoustic*", "*biggest*", "*steven*", "*founding*", "*days*", "*hardware*", "*patience*", "*late*", "*reputed*", "*3d*", "*run*", "*stephen*", "*trustee*", "*boy*", "*theater*", "*join*", "*detection*", "*rather*"]

- words whose likelihood most increased in direction MSC → FEM :

  ["*ms*", "*she's*", "*", "*bri*", "*marie*", "*mrs*", "*girl*", "*herself*"", "*jennifer*"", "*002412*", "*nicole*", "*women's*", "*happy*", "*newborn*", "*andrea*", "*domestic*", "*exploring*", "*mn*", "*colleagues*", "*setting*", "*anne*", "*elizabeth*", "*1215242727*", "*donna*", "*geriatric*", "*nancy*", "*upon*", "*maternal*", "*picture*", "*1215191916*"]

- words whose likelihood most decreased in direction FEM → MSC :

  ["*she's*", "*mrs*", "*girl*", "*marie*", "|", "*herself*", "*clutter*"", "*inspire*", "*uncomfortable*", "*nicole*", "*female*", "*promotes*", "*anne*"", "*desire*", "*13*", "*abuse*", "*lingerie*", "*caring*", "*elder*", "*strangers*", "*classrooms*", "*stephanie*", "*mn*", "*susan*", "*refugee*", "*runway*", "*21*", "*within*", "*59*", "*plants*"]

- words whose likelihood most increased in direction FEM → MSC:

[“*mr*“, “*him*“, “*he’s*“, “*dahl*“", “*himself*“, “*1st*“, “*peers*“, “*plays*“, “*.0*“, “*2019*“, “*developer*“, “*chris*“, “*x*“, “*robert*“, “*veterinary*“, “*esl*“, “*lifetime*“, “*llp*“, “*wallpapers*“, “*adventure*“, “*chance*“, “*managers*“, “*watch*“, “*humour*“, “*murya*“, “*1003021313*“, “*stephen*“, “*list*“, “*say*“, “*concerned*“]

B.2 LEACE

- words whose likelihood most decreased when removing the FEM concept: [“*clutter*“, “*uncomfortable*“, “*strangers*“, “*front*“, “*classrooms*“,“*volunteering*“, “*0000*“, “*never*“, “*travelling*“, “*seminars*“, “*compassion*“, “*cute*“, “*humanitarian*“, “*pre-*“, “*experimental*“, “*accredited*“, “*experiencing*“, “*partnerships*“, “*distribution*“, “*off*“, “*participant*“, “*implementing*“, “*babies*“, “*funny*“, “*die*“, “*photographing*“, “*1903021717*“, “*words*“, “*engaging*“, “*engages*“]

- words whose likelihood most increased when removing the FEM concept:

  [“*him*“, “*he’s*“, “*mr*“, “*hunger*“, “*eat*“, “*himself*“, “*plays*“, “*hot*“, “*showcase*“, “*inspiring*“, “*fair*“, “*authority*“, “*1979*“, “*llp*“, “*watch*“, “*pleasure*“, “*cns*“, “*beyond*“, “*failure*“, “*per*“, “*meets*“, “*suny*“, “*adventure*“, “*agricultural*“, “*serve*“, “*greater*“, “*luxury*“, “*idea*“, “*night*“, “*reuters*“]

- words whose likelihood most decreased when removing the MSC concept:

  [“*et*“, “*elite*“, “*kau*“, “*ch*“, “*pastoral*“, “*direction*“, “*0560*“, “*choice*“, “*august*“, “*patience*“, “*cinema*“, “*restaurant*“, “*58*“, “*theater*“, “*join*“, “*rather*“, “*composing*“, “*tn*“, “*reviewer*“, “*kent*“, “*core*“, “*effect*“, “*mentor*“, “*significant*“, “*entertainment*“, “*hollywood*“, “*something*“, “*photojournalism*“, “*friend*“, “*demand*“]

- words whose likelihood most increased when removing the MSC concept:

  [“*ms*“, “*colleagues*“, “*grace*“, “*prepare*“, “*leaders*“, “*mediations*“, “*greater*“, “*setting*“, “*grown*“, “*happy*“, “*publication*“, “*writers*“, “*similar*“, “*presenter*“, “*counsels*“, “*1903021515*“, “*employee*“, “*19th*“, “*bi*“, “*she’s*“, “*wilderness*“, “*bad*“, “*embedded*“, “*believer*“, “*detail*“, “*promotion*“, “*advocates*“, “*teach*“, “*mri*“, “*dedication*“]

B.3 MiMiC+

- words whose likelihood most decreased in direction MSC $\rightarrow$ FEM : [“*he*“, “*his*“, “*mr*“, “*him*“, “*he’s*“, “*michael*“, “*william*“, “*et*“, “*elite*“, “*mark*“, “*andrew*“, “*robert*“, “*man*“, “*paul*“, “*brian*“, “*richard*“, “*himself*“, “*daniel*“, “*engineer*“, “*funded*“, “*alan*“, “*joseph*“, “*charles*“, “*distributed*“, “–“, “*peter*“, “*developer*“, “*kau*“, “*subject*“, “*adam*“]

- words whose likelihood most increased in direction MSC $\rightarrow$ FEM :

  [“*ms*“, “*women’s*“, “*she’s*“, “*marie*“, “*maternal*“, “*girls*“, “*girl*“, “*1417191916*“, “*1417191997*“, “*empowerment*“, “*michelle*“, “*nicole*“, “*female*“, “*jennifer*“, “*elizabeth*“, “*mrs*“, “*nurses*“, “*parenting*“, “*mary*“, “*promotion*“, “*practitioners*“, “*birth*“, “*empowering*“, “*holistic*“, “*mom*“, “*mothers*“, “*maternity*“, “*woman’s*“, “*crisis*“, “*joy*“]

- words whose likelihood most decreased in direction FEM $\rightarrow$ MSC :

  [“*she*“, “*her*“, “*ms*“, “*women*“, “*she’s*“, “*mother*“, “*ki*“, “*women’s*“, “*mrs*“, “*woman*“, “*elementary*“, “*january*“, “*mary*“, “*daughter*“, “*girl*“, “*jennifer*“, “*marie*“, “|“, “*assisting*“, “*lisa*“, “*jessica*“, “*herself*“, “*elizabeth*“, “*joy*“, “*sexual*“, “*pregnancy*“, “*amy*“, “*sexuality*“, “*opportunities*“, “*rachel*“]

- words whose likelihood most increased in direction FEM $\rightarrow$ MSC:

  [“*he’s*“, “*mr*“, “*him*“, “*guy*“, “*x*“, “***“, “*ka*“, “*himself*“, “*developer*“, “*daniel*“, “*1st*“, “*robert*“, “*1003021313*“, “*juicy*“, “*jeremy*“, “*nephrology*“, “*peers*“, “*chairman*“, “*adam*“, “*hardware*“, “*bi*“, “*matthew*“, “*mark*“, “*acoustic*“, “*//*“, “*christopher*“, “*plays*“, “*.0*“, “*player*“, “*forum*“]

Intervention	Same	Original	Counterfactual
LEACE- FEM	29	55	16
LEACE- MSC	20	52	28
MiMiC- FEM → MSC	8	71	21
MiMiC- MSC → FEM	15	41	44
MiMiC+- FEM → MSC	10	76	14
MiMiC+- MSC → FEM	13	41	46

Table 3: Readability Annotation Results

C Human annotation

We conducted human annotation experiments to evaluate the quality of the interventions using Amazon Mechanical Turk (MTurk). Five annotators, all native English speakers from the US, UK, and Australia, were recruited for this task. The annotators were compensated for their work in line with standard MTurk rates. This selection process ensured that the annotators had a high degree of fluency in English.

Annotators were required to complete three tasks: (1) assess the readability of pairs of sentences (2) assess their grammatical correctness, and (3) determine the subject entity gender for each sentence. These tasks were designed to evaluate the quality and correctness of the generated counterfactuals compared to the original biographies, following the annotation guidelines (see Appendix C).

In tasks (1) and (2), annotators were presented with two texts, labeled Text A and Text B. They were asked to compare the readability and grammatical correctness of the texts, selecting which was more readable and grammatically correct, or indicating that both were comparable. In task (3), annotators were asked to identify the gender of the subject entity in the sentence (male, female, or unclear).

To analyze the results, we performed a Chi-Square Test of Independence to statistically evaluate whether there was a significant difference in the annotation responses before and after applying the interventions.

Intervention	Same	Original	Counterfactual
LEACE- FEM	48	36	16
LEACE- MSC	58	29	13
MiMiC- FEM → MSC	57	37	6
MiMiC- MSC → FEM	52	30	18
MiMiC+- FEM → MSC	49	38	13
MiMiC+- MSC → FEM	45	34	21

Table 4: Grammar Annotation Results

Hypotheses We formulated our hypotheses separately for each task and applied the appropriate statistical tests:

- **Readability and Grammar (One-Tailed Binomial Test)**
  - **Null Hypothesis** (H_0): The original text is *not* preferred over the counterfactual text in terms of readability/grammar; that is, the probability of preferring the original biography is less than or equal to 0.5.
  - **Alternative Hypothesis** (H_1): The original text is preferred over the counterfactual text in terms of readability/grammar; that is, the probability of preferring Text A is greater than 0.5.
- **Gender Specification (Chi-Square Test of Independence)**
  - **Null Hypothesis** (H_0): The distribution of gender identification is independent of the intervention; the intervention does not affect how annotators perceive the gender of the subject entity.

– **Alternative Hypothesis** (H_1): The distribution of gender identification is dependent on the intervention; the intervention affects how annotators perceive the gender of the subject entity.

Intervention	Test	Test Statistic	P-value
LEACE - Originally Female	Readability (Binomial Test)	$k = 55, n = 100$	$p = 0.18$
	Grammar (Binomial Test)	$k = 36, n = 100$	$p = 1.00$
	Gender Specification (Chi-Square Test)	$\chi^2 = 71.98$	$p = 2.34 \times 10^{-16}$
LEACE - Originally Male	Readability (Binomial Test)	$k = 52, n = 100$	$p = 0.38$
	Grammar (Binomial Test)	$k = 29, n = 100$	$p = 1$
	Gender Specification (Chi-Square Test)	$\chi^2 = 24.87$	$p = 3.97 \times 10^{-6}$
MiMiC - Female to Male	Readability (Binomial Test)	$k = 71, n = 100$	$p = 1.60 \times 10^{-5}$
	Grammar (Binomial Test)	$k = 37, n = 100$	$p = 1.00$
	Gender Specification (Chi-Square Test)	$\chi^2 = 130.56$	$p = 4.45 \times 10^{-29}$
MiMiC - Male to Female	Readability (Binomial Test)	$k = 41, n = 100$	$p = 0.97$
	Grammar (Binomial Test)	$k = 30, n = 100$	$p = 1.00$
	Gender Specification (Chi-Square Test)	$\chi^2 = 131.44$	$p = 2.87 \times 10^{-29}$
MiMiC Plus - Female to Male	Readability (Binomial Test)	$k = 76, n = 100$	$p = 9.05 \times 10^{-8}$
	Grammar (Binomial Test)	$k = 38, n = 100$	$p = 0.99$
	Gender Specification (Chi-Square Test)	$\chi^2 = 103.94$	$p = 2.69 \times 10^{-23}$
MiMiC Plus - Male to Female	Readability (Binomial Test)	$k = 41, n = 100$	$p = 0.97$
	Grammar (Binomial Test)	$k = 34, n = 100$	$p = 1.00$
	Gender Specification (Chi-Square Test)	$\chi^2 = 134.58$	$p = 5.97 \times 10^{-30}$

Table 5: Test Results for Readability, Grammar, and Gender Specification tasks across LEACE MiMiC and MiMiC+ interventions. P-values above 0.05 in the binomial tests indicate no significant preference for the original text over the counterfactual, suggesting that the interventions did not degrade text quality. P-values below 0.05 in the chi-square tests indicate statistically significant differences in gender specification after the interventions.

Results The results of the statistical tests are summarized in Tab. 5, Tab. 6 and Tab. 4.

For the **Readability** and **Grammar** tasks, we performed one-tailed binomial tests. The number of times the original text was preferred (k) and the total number of observations (n) are reported, along with the p-values.

For the **Gender Specification** task, we performed chi-square tests of independence, reporting the chi-square statistic and the p-value.

Conclusions

- **Readability and Grammar:** For most interventions, the p-values from the one-tailed binomial tests are greater than 0.05, indicating that we fail to reject the null hypothesis. This suggests that the original text was not significantly preferred over the counterfactual in terms of readability and grammatical correctness, implying that the interventions did not degrade text quality.

  However, for the *MiMiC - Female to Male* and *MiMiC+ - Female to Male* interventions, the p-values for readability are less than 0.05 ($p = 1.60 \times 10^{-5}$ and $p = 9.05 \times 10^{-8}$, respectively). This means we reject the null hypothesis in these cases, indicating that the original text was significantly preferred over the counterfactual in terms of readability. This suggests that these interventions may have led to a degradation in readability when altering gender from female to male.

- **Gender Specification:** For all interventions, the p-values from the chi-square tests are significantly less than 0.05, leading us to reject the null hypothesis. This indicates that the interventions had a statistically significant effect on how annotators perceived the gender of the subject entity. Therefore, the interventions were effective in altering the perceived gender in the texts.

An agreement between the annotators was measured by Fleiss' κ score (Fleiss, 1971). For task (1), comparing the readability of the sentence pairs, Fleiss' κ was 0.23, indicating fair agreement among the

Technique	Data Type	Male	Female	Unclear
LEACE - Fem	Original	11	85	4
	Intervention	66	26	8
LEACE - MSC	Original	92	5	3
	Intervention	64	32	4
MiMiC - FEM → MSC	Original	3	97	0
	Intervention	82	17	1
MiMiC - MSC → FEM	Original	90	10	0
	Intervention	9	86	5
MiMiC PLUS - FEM → MSC	Original	12	87	1
	Intervention	84	16	0
MiMiC PLUS - MSC → FEM	Original	91	9	0
	Intervention	9	88	3

Table 6: Gender Annotation Results for Different Intervention Techniques

annotators. For task (2), comparing the grammar level of the sentence pairs, Fleiss' κ was 0.21, indicating fair agreement among the annotators. For task (3), determining the subject entity's gender, Fleiss' κ was 0.6, indicating moderate to substantial agreement.

The counterfactual was randomly presented as Text A or Text B with a uniform distribution. Moreover the counterfactual sentence was generated by applying one of the three intervention techniques followed by the inversion model. The samples were drawn uniformly with replacement using a random sampling generator.

The exact annotation guidelines provided to the annotators are given below in Appendix C.

Annotation Guidelines

Overview

You will be provided with two texts, labeled Text A and Text B.
Your task is to evaluate these texts based on their:

- **Readability**
- **Grammatical correctness**
- **Entity gender specification**

Examples

Read the following two texts (Text A and B) and answer the following questions:

Text A:
In this capacity he will assist clients in matters involving estates, trusts, wills, guardianships, asset disputes, powers of attorney, and advanced medical directives.

Text B:
In this capacity, she will assist clients in a variety of medical matters, including elder care, medical malpractice, wills, trusts, powers of attorney, guardianships, and advanced medical directives.

Question 1: Which of the texts A or B is more readable and understandable? If both texts are comparable in terms of readability, select *Same*.
Answer: Same

Question 2: Which of the texts A or B is more grammatically correct? If both texts are comparable in terms of grammar, select *Same*.
Answer: Same
Question 3: Is the subject entity male, female, or unclear?
Answer:

- Text A: Male
- Text B: Female

Text A:
She studied at the Wimbledon School of Art 1980-84 and later on with Cecil Collins and Sybil Andrews. She has traveled extensively, setting up homes and painting in Kenya, Dubai, Canada, and Jerusalem.
Text B:
He studied at the London College of Art with Andrew Davies and Sybil Kennedy. Since 1987, he has traveled to New Zealand, Canada, Israel, Kenya, Australia, and New Zealand, where he studied a range of painting-in-residences including...

Question 1: Which of the texts A or B is more readable and understandable? If both texts are comparable in terms of readability, select *Same*.
Answer: Text A
Question 2: Which of the texts A or B is more grammatically correct? If both texts are comparable in terms of grammar, select *Same*.
Answer: Same

Readability and Grammatical Correctness

Read the following two texts (Text A and B) and answer the following questions:
Text A:
${text_a}
Text B:
${text_b}

Question 1: Which of the texts A or B is more readable and understandable? If both texts are comparable in terms of readability, select *Same*.
Possible Answers:

- Same
- Text A
- Text B

Question 2: Which of the texts A or B is more grammatically correct? If both texts are comparable in terms of grammar, select *Same*.
Possible Answers:

- Same
- Text A
- Text B

Gender Tagging

For each text, determine the gender of the subject entity.

Text A:

${text_a}

Is the subject entity male, female, or unclear?

Possible Answers:

- Male
- Female
- Unclear

Text B:

${text_b}

Is the subject entity male, female, or unclear?

Possible Answers:

- Male
- Female
- Unclear

D Intervention inversion sample

In Tab. 7 we provide a random sample of the counterfactuals generated by the different methods.

Method	Inversion without intervention	Intervention + Inversion
MiMiC FEM → MSC	A 2017 Sports Illustrated Swimsuit Trend Hero, she talks about how she learned to embrace her body's curves, embrace their natural curves, and never let it get in the way of her career.	A 2017 Sports Illustrated Swimsuit Hero, he embraces his flaws to become a hero, but never lets risk get in the way of his self - confidence.
MiMiC FEM → MSC	Prior to moving to New York, Naomi worked as a registered nurse for six years, specializing in cardiology. During this time, Naomi also worked in the Emergency Department at Duke University Medical Center. Naomi has a Master's Degree in Nursing. She is	Prior to joining Manhattan, she worked as an RN in the Emergency Department, specializing in cardiology. During her time at UMDNJ - NewYork, Naomi completed a Master's Degree in Cardiology.
MiMiC MSC → FEM	Dr. Ahn attended medical school at Northwestern University, Feinberg School of Medicine and completed a residency at the Children's Memorial Hospital. He is board certified by the American Board of Pediatrics.	Dr. Ahn attended medical school at Northwestern University Feinberg School of Medicine and completed her residency in Pediatrics at the Children's Hospital. She is board - certified by the American Academy of Nurse Practitioners in Pediatrics.
MiMiC MSC → FEM	He earned his Ph.D. in business administration from Penn State University's Smeath School of Business. His research interests include business process innovation, data mining, information security, and healthcare operations research.	She earned her doctorate in business administration from Penn State University's Smeal School of Business and her master's and bachelor's degrees from Penn State University. Her research interests include healthcare operations research, e - health, innovation, patient satisfaction, process
LEACE (Originally female)	She earned Bachelor of Science degrees in Biology and Chemistry from Metropolitan State College in DeKalb, MN. Dean obtained her Master of Science degree in physician assistant studies from Denver Medical College.	He earned a Bachelor of Science degree in Biology and Chemistry from Metropolitan State College in Denver, Colorado. Dean obtained his Master of Medical Science degree in physician assistant studies from Western Michigan University.
LEACE (Originally female)	She has devoted her life to helping people to achieve their health goals through diet, nutrition and lifestyle counseling. Diana can be contacted at : danielle.danielle.co.uk	He has helped hundreds of people to achieve their health and lifestyle through nutrition. Diana helps people to navigate their way of life through nutrition and counseling. For more information, please visit : http://www.daniellebakhron.com
LEACE (Originally male)	He focuses his practice on understanding how to protect data and privacy across the entire enterprise. Mr. Robinson also works with federal and state agencies to develop data security standards, best practices, cybersecurity best practices, and best practices to protect data and privacy across the entire enterprise.	She focuses her practice on partnering with businesses, government agencies, and private companies to develop and implement best practices in data security, data privacy, and cybersecurity. Mr. Robinson also counsels businesses on how to protect their data assets and comply with federal data security standards.
LEACE (Originally male)	He graduated with a degree in Architecture and Design from the Architectural Association School of Design and Design. His architectural practice is based in Notting Hill London.	He graduated from the Architectural Association School of Architecture and Design based in Notting Hill, London. He also has a degree in architecture and design. His practice is based in notting Hill.
MiMiC+ FEM → MSC	She is the co - author of Films, Culture, and Popular Culture in Argentina. Her research and writings have appeared or are forthcoming in Latin American Studies, Latin American Studies, Latin American Culture, and other journals.	He is the author or co - author of many books and many essays on Latin American culture. His main field of study is the history of theater, cinema, and popular culture in Argentina. He was the Editor of Cinemas, Cultures, and Cultures in Argentina,
MiMiC+ FEM → MSC	Her work specializes in counseling clients in information security - related incidents, risk management, data collection and compliance and assisting clients with information security - related risks and compliance.	His work specializes in information security counseling and compliance for clients, dealing with security - related incidents, data - related risk management and software - related compliance incidents.
MiMiC+ MSC → FEM	He practices in Fayetteville, Arkansas and has the professional credentials of MD. The NPI Number for Robert L. Firstseine is 1134262688 and he holds a License No. 67 (Arkansas).	She practices in Fayetteville, Arkansas and has the professional credentials of RN. The NPI Number for Firstbornee L. Roberts is 1396764048 and she holds a License No. 325 (Arkansas).
MiMiC+ MSC → FEM	Dr. Russ Meka has recently obtained his Ph.D. in Global Affairs at the University of Michigan, Ann Arbor, USA. He has also been a Research Fellow at the Center for Globalization and Society, Bergen University, Norway.	Ms. Meka has earned her Ph.D. in Global Affairs from Rutgers University, Newark, NJ. She is also an Assistant Professor at Rutgers University, Newark, NJ. Her research interests include global medicine, social justice,

Table 7: Random sample of inverted representations without intervention, alongside an intervention + inversion.