Dual Debiasing: Remove Stereotypes and Keep Factual Gender for Fair Language Modeling and Translation

Tomasz Limisiewicz[1*] and **David Mareček**[2] and **Tomáš Musil**[2]
[1]Paul G. Allen School of Computer Science and Engineering, University of Washington
[2] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague
tomlim@cs.washington.edu {marecek, musil}@ufal.mff.cuni.cz

Abstract

Mitigation of biases, such as language models' reliance on gender stereotypes, is a crucial endeavor required for the creation of reliable and useful language technology. The crucial aspect of debiasing is to ensure that the models preserve their versatile capabilities, including their ability to solve language tasks and equitably represent various genders. To address these issues, we introduce *Dual Debiasing Algorithm through Model Adaptation* (*2DAMA*). Novel *Dual Debiasing* enables robust reduction of stereotypical bias while preserving desired factual gender information encoded by language models. We show that *2DAMA* effectively reduces gender bias in language models for English and is one of the first approaches facilitating the mitigation of their stereotypical tendencies in translation. The proposed method's key advantage is the preservation of factual gender cues, which are useful in a wide range of natural language processing tasks.

1 Introduction

Gender representation in large language models (LLMs) has been the topic of significant research effort (Stanczak and Augenstein, 2021; Kotek et al., 2023). Past studies have predominantly focused on such representation to identify and mitigate social biases. Admittedly, biases are a challenging issue limiting the reliability of LLMs in real-world applications. However, we argue that preserving particular types of gender representation is crucial for fairness and knowledge acquisition in language models.

To provide a more detailed perspective, we draw examples of both unwanted and beneficial types of gender signals in LLMs. Undesirable biases are typically inherited from stereotypes and imbalances in the training corpora and tend to be amplified further during model training (Van Der Wal et al., 2022; Gallegos et al., 2024). Biases are manifested in multiple ways, including unequal representation (models are more likely to generate mentions of a specific overrepresented gender), stereotypical associations (particular contexts are associated with one gender based on stereotypical cues, e.g., "*politics and business are male domains*", while "*family is a female domain*"). It has been shown that, due to bias, LLMs struggle with high-stakes decision-making and are prone to produce discriminatory predictions. Examples of such a sensitive application are the automatic evaluation of CVs and biographical notes (De-Arteaga et al., 2019), where some professions are stereotypically associated with a specific gender. Therefore, individuals of another gender could face an unfair disadvantage when assessed by an LLM-based evaluator.

Nevertheless, LLMs should understand and represent gender signals. For instance, chatbots should be persistent in addressing the user with their preferred gender pronouns after they are revealed (Limisiewicz and Mareček, 2022). Adequate representation of gender is also required for knowledge acquisition, for example, in question answering (QA), to correctly answer "*Maria Skłodowska-Curie*" to the question "*Who was the first woman*

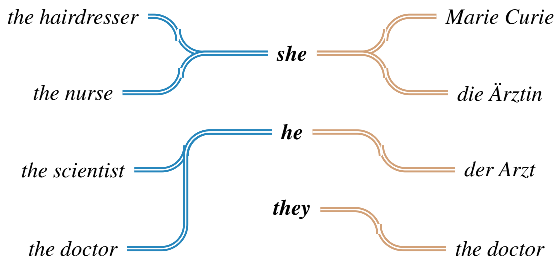


Figure 1: Dual character of gender signals encoded in language models: stereotypical cues are shown on the left, and factual cues are shown on the right-hand side. "*Die Ärztin*" and "*der Arzt*" are respectively female and male German translation for "*the doctor*".

[*]Work done at Charles University

Proceedings of the 18th International Natural Language Generation Conference, pages 148–164
October 29–November 2, 2025. ©2025 Association for Computational Linguistics

to win a Nobel Prize?". Gender sensitivity is even more critical in morphologically rich languages, where gender mentions are much more ubiquitous, e.g., through morphological markings (as in German, Czech, or Russian) (Hellinger and Bußmann, 2002). Examples of dual characters (stereotypical vs. factual) of gender encoding are shown in Figure 1.

To address these intricate ways gender signals are present in natural language, we introduce a new method, *2DAMA*, that post-hoc modifies pre-trained language models to represent gender in an equitable way, i.e., without stereotypical bias but with factual gender information. **As the core contribution, we introduce the novel method of *Dual Debiasing* that aims at our core problem of decreasing bias while maintaining an equitable representation of the factual gender.** Specifically, we aim to reduce the models' reliance on stereotypes in predictions, e.g., given a stereotypical prompt as the one in Figure 2: "*The salesperson laughed because*", we intend to coerce equitable probabilities of possible gender predictions manifested by pronouns "*he*", "*she*", or "*them*". On the contrary, when considering a prompt that contains factual gender information: "*The king laughed because*" the desired output distribution would assign a high probability to the male pronoun.

2 Methodology

In this section, we formally introduce *Dual Debaising Algorithm through Model Adaptation* (*2DAMA*), a new dual debiasing method, and provide theoretical backing for the presented approach. Appendix A contains the proofs and further terminological explanations.

2.1 Background and Novel Methods

In *2DAMA*, we introduce the novel *Dual Debiasing* and incorporate it into the debiasing framework that comprises of two previously established algorithms: *DAMA*, *LEACE*. We provide a clear distinction between previous and novel approaches described in this paper.

Background Methods: DAMA *Debiasing Algorithm through Model Adaptation* (Limisiewicz et al., 2024) is a method for adapting parameters of language models to mitigate the encoding of harmful biases without affecting their general performance. The method employs model editing techniques (Meng et al., 2022) to disassociate specific signals provided in a prompt with the model outputs, i.e., stereotypes in prompts and gendered output. **LEACE** *LEAst-squares Concept Erasure* (Belrose et al., 2023) is a method of concept erasure (such as bias signal) in latent representation.

Novel Methods: DAMA-LEACE (Section 2.2) The first innovation is streamlining the base debiasing algorithm *DAMA*. We achieve it by replacing the Partial Least Squares concept erasure used in *DAMA* with *LEACE*, which does not require pre-defining the dimensionality of erased signals. The core novelty of this work is **2D** *Dual Debiasing*, a new algorithm that we formally introduce in Section 2.3. The method uses covariance matrix decomposition to identify correlates related to bias and protected feature signals. A concept erasure algorithm is modified to erase bias while preserving protected features, such as factual gender.

2.2 Contribution: DAMA-LEACE

LEACE guarantees erasing a specific concept's influence on a latent vector. In a neural network, we can consider a latent vector U to be an output of one of the intermediate layers. *LEACE* aims to de-correlate latent vectors with an unwanted signal (e.g., gender bias), whose distribution is represented as another vector Z.

In model editing, we are interested in how a model's layer maps its input vector U to output vector V (unlike *LEACE*, which focuses on stand-alone latent vector U). We are specifically interested in a transformation that minimizes the distance between the input (keys: U) and the predicted variables (values: V). Such U can be a latent vector obtained by feeding into a model a gendered prompt, while Z is a vector corresponding to stereotypical output.

We reasonably assume that dense layers of trained neural networks (e.g., feed-forward layers in Transformer) fulfill this purpose, i.e.:

$$V = \boldsymbol{S}U - \epsilon, \tag{1}$$

where $\boldsymbol{S}$ is a linear transformation and ϵ a vector of errors. Due to gradient optimization in the model's pre-training, we assume that the feed-forward layer approximates the least solution, i.e., $FF \approx \boldsymbol{S}$.

Taking this assumption, we can present a theorem guaranteeing concept erasure (based on *LEACE*) in the model adaptation algorithm (*DAMA*):

Theorem 1 (**DAMA-LEACE**). *We consider random vectors: U taking values in $\mathbb{R}^m$, V and Z taking values in $\mathbb{R}^n$, where $m \geq n$. Under assumptions that: A) random vectors U, V, Z are centered, and each of them has finite moment; B) the regression relation between U and V fulfill the assumption of ordinary least squares, and there exist least squares estimator $V = \boldsymbol{S}U - \epsilon$.*

Then the objective:

$$\underset{\boldsymbol{P} \in \mathbb{R}\, n \times m}{\arg\min}\ \mathbb{E}\left[||\boldsymbol{P}U - V||^2\right],$$

subject to:

$$\mathbb{C}\mathrm{ov}(\boldsymbol{P}U, Z) = 0$$

is solved by:

$$\boldsymbol{P}^* = \left(\mathbb{I} - \boldsymbol{W}^+ \boldsymbol{P}_{\boldsymbol{W\Sigma}} \boldsymbol{W}\right) \boldsymbol{S},$$

where $\boldsymbol{W}$ is the whitening transformation $(\Sigma^{1/2}_{\boldsymbol{SU},\boldsymbol{SU}})^+$; $\boldsymbol{P}_{\boldsymbol{W\Sigma}}$ is an orthogonal projection matrix onto colspace of $\boldsymbol{W\Sigma}_{\boldsymbol{S}U,Z}$; $\boldsymbol{S}$ is a least squares estimator of V given U: $\boldsymbol{S} = \boldsymbol{\Sigma}_{U,V}\boldsymbol{\Sigma}^{-1}_{U,U}$.[1]

Based on the theorem and the assumption that $FF \approx \boldsymbol{S}$ applying projections would break the correlation between stereotypical keys and gendered values with minimal impact on other correlations stored by the feed-forward layer. We call the algorithm realizing such adaptation in a neural network: *DAMA-LEACE*.

2.3 Contribution: Dual Debiasing

In *Dual Debiasing*, we extend the concept erasure problem by considering two type signals and corresponding random variables: Z_b bias to be erased and Z_f feature to be preserved. We posit that:

Theorem 2 (**DUAL-DEBIASING**). *We consider random vectors X, Z_b, and Z_f in $\mathbb{R}^n$. Under the assumptions that: A) $Z_b \perp Z_f | X$, i.e., Z_b and Z_f are conditionally independent, given X; B) $\boldsymbol{\Sigma}_{X,Z_b}\boldsymbol{\Sigma}^T_{X,Z_f} = 0$, i.e., the variable X is correlated with Z_f and Z_b through mutually orthogonal subspaces of $\mathbb{R}^n$. The solution of the objective:*

$$\underset{\boldsymbol{P} \in \mathbb{R}^{n \times n}}{\arg\min}\, \mathbb{E}\left[||\boldsymbol{P}X - X||^2\right],$$

subject to:

$$\mathbb{C}\mathrm{ov}(\boldsymbol{P}X, Z_b) = 0,$$

satisfies:

$$\mathbb{C}\mathrm{ov}(\boldsymbol{P}X, Z_f) = \mathbb{C}\mathrm{ov}(X, Z_f).$$

[1]Notation: $^+$ denotes Moonrose-Penrose psuedoinverse. For brevity, we use $\boldsymbol{\Sigma}_{V,Z}$ for covariance matrix $\mathbb{C}\mathrm{ov}(V, Z)$. The complete terminological note can be found in Appendix A

The theorem shows that the correlation with the conditionally independent features is left intact by applying *LEACE* erasure to a bias signal. However, the assumption of conditional independence is strong and unlikely to hold when considering the actual signals encoded in the model. Thus, for practical applications, we need to relax the requirements.

In *Dual Debiasing*, we relax the assumption of the theorem in order to consider bias and feature signals that can be conditionally correlated. In constructing the debiasing projection ($\boldsymbol{P}*$), we must decide whether specific dimensions should be nullified or preserved. We propose to nullify dimensions of X with t times higher correlation with Z_f than Z_b, where the threshold t (later referred to as *bias-to-feature threshold*) is empirically chosen. To analyze the correlations we consider correlation matrix $\boldsymbol{W\Sigma}_{X,[Z_f,Z_b]}$. By using singular value decomposition, we can identify the share of variance in each column's first n rows (associated with Z_f) and the latter n rows (associated with Z_b). In modified colspace projection $\tilde{\boldsymbol{P}}_{\boldsymbol{W\Sigma}}$, we only consider the column with t times higher variance with Z_f than with Z_b. Thus the final *Dual Debiasing LEACE* projection $\tilde{\boldsymbol{P}}^* = \left(\mathbb{I} - \boldsymbol{W}^+ \tilde{\boldsymbol{P}}_{\boldsymbol{W\Sigma}} \boldsymbol{W}\right)$ will to large extent preserve the protected feature while reliably erasing bias. In Section 4.2, we experimentally study the impact of feature-to-bias threshold t.

3 Experimental Setting

This section presents an empirical setting to examine the practical application of model editing methods. We describe models, data, and evaluation metrics for gender bias and general performance.

3.1 Models

In experiments, we focus on Llama family models (Touvron et al., 2023; Dubey et al., 2024), which are robust and publicly available language models developed by Meta AI. We analyze Llama 2 models of sizes 7 and 13 billion parameters and Llama 3 with 8 billion parameters. In multilingual experiments, we use ALMA-R 13 billion parameter model (Xu et al., 2024). ALMA-R is based on an instance of Llama 2 model that was fine-tuned to translate using Contrastive Preference Optimization. ALMA-R covers translation between English and five languages (German, Czech, Russian, Icelandic, and Chinese).

EN: “The **salesperson** laughed because” { he | she }
EN: “The **saleseperson** is not working today.” $\rightarrow$ DE: { Der | Die }
EN: “That **saleseperson** is not working today.” $\rightarrow$ CS: { Ten | Ta }

Figure 2: Sterotypical prompts with possible gendered outputs (in brackets) in three languages. We use prompts to obtain stereotypical key vector U, the possible outputs are used to approximate gendered values vector V.

In model editing experiments, we adapt the layers starting from the one found in the two-thirds of the layer stack counted from the input to the output. It is the 26th layer for 13 billion parameter models and the 21st layer for smaller models. For example, the adaptation of 11 mid-upper layers in the 13B model modifies the layers from 26th through 37th.

3.2 Data for Dual Debiasing

Following Limisiewicz et al. (2024), we feed prompts to the model in order to obtain the latent embeddings in the input of intermediate layers. We treat these embeddings as key vectors (U) containing stereotypical or factual gender signals. To obtain gendered value vectors (V), we find the output vector of the layer that would maximize the probability of predicting tokens corresponding to gender.

Language Modeling Prompts For debiasing language models, we use solely English prompts. We design 11 prompt templates, such as “*The **X** laughed because* ___”, where “**X**” should be replaced by profession name. This prompt construction provokes the model to predict one of the gendered pronouns (“*he*”, “*she*”, or “*they*”). To distinguish stereotypical signals for debiasing, we use 219 professions without factual gender that were annotated as stereotypically associated with one of the genders by Bolukbasi et al. (2016).

Multilingual Prompts For debiasing machine translation, we use prompts instructing the model to translate sentences containing the same set of 219 professions to a target language that has the grammatical marking of gender, e.g., “*English: The **X** is there. German:* ___”. The translation model would naturally predict one of the German determiners, which denotes gender (“*Der*” for male or “*Die*” for female). For each model, we adjust the template to include instructions suggested by the ALMA authors. We construct translation prompts for two target languages, Czech and German, proposing 11 templates, and thus 2409 prompts for each language.

Factual Prompts *Dual debiasing* requires using factual prompts to identify the signal to be preserved. For that purpose, we use the same prompt templates as defined above (both English and multilingual) with the distinction of entities used to populate them. For that purpose, we propose 13 pairs of factually male and female entities, e.g., “*king*” – “*queen*”, “*chairman*” – “*chairwoman*”.

The examples of language modeling and multilingual prompts are given in Figure 2. We list all the prompt templates in Appendix B.

3.3 Bias Evaluation

Language Modeling We assess the bias in language generation following the methodology of Limisiewicz et al. (2024). From the dataset of Bolukbasi et al. (2016), we select the held-out set of professions that were not included in the 219 used for debiasing. For each of these professions, annotators had assigned two scores: *factual* score x_f and *stereotypical* score x_s. The scores define how strongly a word (or a prompt) is connected with the male or female gender, respectively, through factual or stereotypical cues. By convention, scores range from -1 for female-associated words to 1 for male ones. We measure the probabilities for gendered prediction for a given prompt $P_M(o|X)$. For that purpose, we use the pronouns o_+ = “he” and o_- = “she”, since they are probable continuations of given prompts. Subsequently, for each prompt, we compute *empirical* score $y = P_M(o_+|X) - P_M(o_-|X)$. We estimate the linear relationship between scores:

$$y = a_s \cdot x_s + a_f \cdot x_f + b_0 \qquad (2)$$

The linear fit coefficients have the following interpretations: a_s is an impact of stereotypical signal on the model's predictions; a_f is an impact of the factual gender of the word. Noticeably, y, x_s, and x_f take values in the same range. The slope coefficient tells how shifts in annotated scores across professions impact the difference in prediction probabilities of male and female pronouns. The intercept b_0 measures how much more probable the male

	Bias in LM			WinoBias	
	$\downarrow a_s$	$\uparrow a_f$	$\downarrow b$	$\downarrow \Delta S$	$\downarrow \Delta G$
Llama 2 7B	0.234	0.311	0.090	33.6	7.3
DAMA	0.144	0.205	0.032	27.3	6.8
DAMA+LEACE	0.118	0.171	0.028	22.9	5.4
2DAMA	0.128	0.187	0.042	22.9	5.7
Llama 2 13B	0.244	0.322	0.097	35.0	0.3
DAMA	0.099	0.160	0.030	26.4	2.4
DAMA+LEACE	0.098	0.159	0.026	26.5	2.4
2DAMA	0.119	0.206	0.023	27.0	1.9
Llama 3 8B	0.262	0.333	0.082	36.8	2.7
DAMA	0.069	0.090	0.144	20.3	4.2
DAMA+LEACE	0.084	0.157	0.082	18.8	2.7
2DAMA	0.140	0.209	0.051	18.7	2.4

Table 1: Bias evaluation for the Llama family models, and their adaptation with different debiasing algorithms (*DAMA*, *DAMA* with *LEACE*, and *2DAMA*). The debiasing adaptation was applied to 12 mid-upper layers for the 13B model and 9 mid-upper layers for the smaller ones. In *2DAMA*, we set bias-to-feature threshold to $t = 0.05$.

	LM	ARC		MMLU
	$\downarrow$ ppl	$\uparrow$ acc C	$\uparrow$ acc E	$\uparrow$ acc
Llama 2 7B	21.28	70.2	42.5	35.5
DAMA	21.51	69.8	42.8	35.2
DAMA+LEACE	23.81	68.3	41.2	34.3
2DAMA	23.66	67.5	42.0	34.0
Llama 2 13B	19.68	72.6	46.8	-
DAMA	18.94	71.6	45.0	-
DAMA+LEACE	19.67	71.3	46.4	-
2DAMA	19.90	71.2	46.1	-
Llama 3 8B	-	67.1	39.9	-
DAMA	-	64.6	38.1	-
DAMA+LEACE	-	63.0	39.8	-
2DAMA	-	63.5	37.9	-

Table 2: General performance in language modeling and reasoning on ARC and MMLU datasets. We present results for Llama family models, and their adaptation with different debiasing algorithms (*DAMA*, *DAMA* with *LEACE*, and *2DAMA*). We do not present perplexity for Llama 3 because the model has different vocabulary and the results are not comparable. For ARC we present results for both **C**hallenge and **E**asy subsets. The hyperparameters are the same as in Table 1

pronouns are than the female pronouns when we marginalize the subject.

Other Bias Manifestations in English We evaluate the bias in coreference resolution based on **WinoBias** dataset (Zhao et al., 2018). We use metrics ΔG and ΔS to evaluate representational and stereotypical bias, respectively. ΔG measures the difference in coreference identification correctness (accuracy) between masculine and feminine entities; similarly, ΔS measures the difference in accuracy between pro-stereotypical and anti-stereotypical instances of gender role assignments.

Translation Stanovsky et al. (2019) proposed using Winograd Challenge sentences for evaluating bias in translation from English into eight languages with morphological marking of gender (e.g., German, Spanish, Russian, Hebrew). In **WinoMT**, the correctness of the translation is computed by the F1 score of correctly generating gender inflection of profession words in the target language. The evaluation of gender bias is analogical, as in WinoBias. ΔG and ΔS measure the difference in F1 scores: male vs. female and pro- vs. anti-stereotypical sets of professions, respectively. The more recent **BUG** (Levy et al., 2021) dataset is based on the same principle of bias evaluation, with the distinction that it contains naturally occurring sentences instead of generic templates used in WinoMT.

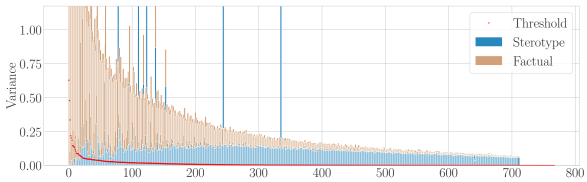


Figure 3: Visualization of dimensions and their variances related to stereotypical and factual gender signals identified by *Dual Debiasing* algorithm in 26th layer of Llama 2 13B. The red dots denote the bias-to-feature threshold $t = 0.05$. In *2DAMA*, the dimension is preserved if stereotypical covariance is below the threshold.

3.4 General Performance Evaluation

Language Modeling We evaluate perplexity on general domain texts from **Wikipedia-103** corpus (Merity et al., 2016).

Reasoning Endtask To assess the models' reasoning capabilities, we compute accuracy on **AI2 Reasoning Challenge (ARC)** (Clark et al., 2018) in both easy and challenging subsets.

Translation To monitor the effect of debiasing on translation quality, we evaluate models on **WMT-22** (Kocmi et al., 2020) parallel corpora with German, Czech, and Russian sentences and their translations in English. We estimate the quality by two automatic metrics: COMET-22 (Rei et al., 2022) and chrf (Popović, 2015).

Layer	Bias Dimesnions		Variance Erased	
	Erased	Preserved	Bias	Factual
26	712	12	99.6%	69.4%
27	774	18	99.4%	64.0%
28	782	22	99.0%	62.4%
29	750	17	99.5%	65.4%
30	713	19	99.5%	64.0%
31	304	12	99.3%	57.6%
32	387	16	99.2%	57.0%
33	469	17	99.2%	60.1%
34	716	21	99.2%	61.3%
35	621	18	99.2%	62.2%
36	406	20	98.9%	54.0%
37	409	18	99.1%	57.2%

Table 3: Number of erased and preserved orthogonal dimensions with *2DAMA* in each feed-forward layer. We call a dimension "*biased*" when it belongs to col-space spanned by covariance matrix between latent representation and bias signal ($\boldsymbol{W}\boldsymbol{\Sigma}_{SU,Z}$). We present the percentage of erased covariance with stereotypical bias and factual gender as the result of the intervention in the layers. The bias-to-feature threshold was set at $t = 0.05$.

4 Debiasing Language Models

In the first batch of the experiments, we evaluate the effectiveness of debiasing language models. In these experiments, we solely focus on tasks in English. We specifically analyze three model editing approaches: *DAMA* as a baseline; *DAMA* in combination with *LEACE*; and *2DAMA*, which employs *Dual Debiasing* to preserve factual gender information.

4.1 Main Results

Model editing reduces bias and preserves the model's performance. All of the considered methods reduce gender bias both in language modeling and coreference resolution (Table 1). Remarkably, we observe that the model's overall performance, i.e., unrelated to gender, is not significantly affected, as demonstrated by perplexity and question-answering results (Table 2). Relatively worse performance preservation was observed for Llama 3, which could be caused by intervening in too many layers.

Streamlining the approach with *LEACE*. We observe that *DAMA-LEACE* reduces bias to a larger extent than baseline *DAMA*. The more substantial debiasing effect comes in pair with a slightly higher drop in general performance, as shown in Table 2. Yet, the deterioration is still small compared to the original models' scores. The crucial benefit of *DAMA-LEACE* is that projection dimensionality does not need to be pre-defined because it is learned implicitly (details in Section 2.2).[2] That motivates us to use *DAMA-LEACE* in further experiments.

Preserving factual gender with *Dual Debiasing*. The coefficients a_s and a_f from Table 1 indicate how much the models' prediction is affected by gender present through stereotypical and factual cues, respectively. We see that *2DAMA* enables, to a significant extent, preserving factual gender information (as indicated by higher a_f coefficient) with a slight increase in susceptibility to gender bias.

4.2 Relationship between Stereotypical and Factual Signals

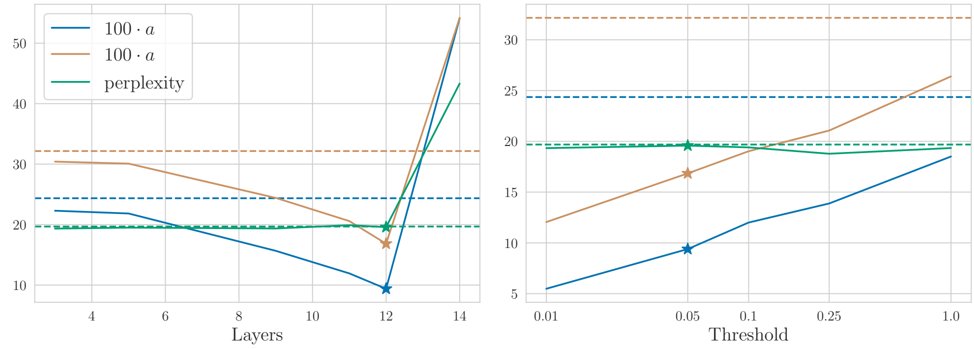
With *Dual Debiasing*, we can analyze the covariance of embedding space orthogonal dimensions in the model's feed-forward layers with the stereotypical and factual signals (as detailed in Section 2.3). In Figure 3, we plot these covariances for each dimension. The visualization reveals that the factual gender is represented by relatively few dimensions with high covariance. In contrast, stereotypical bias is encoded in more-dimensional subspaces, yet each dimension has low covariance.

This observation suggests that in debiasing, we need to exempt just a small subset of dimensions encoding factual gender. Accordingly, further analysis (shown in Table 3) shows that *2DAMA* obtains a reasonable threshold with low bias-to-feature threshold $t = 0.05$. Such a setting preserves only a few dimensions responsible for stereotypical bias in each layer. Such intervention in the model erases $\approx 99\%$ of covariance with a stereotypical signal while keeping over 30% of covariance with a factual gender signal.

4.3 Choice of Hyperparameters

We present the impact of two parameters on the effectiveness of *2DAMA* in Figure 4. The first is the bias-to-feature threshold t. We observe that its choice controls the trade-off between mitigating bias and preserving factual information. We set it a low value of 0.05 because our primary objective is the reduction of bias. The second hyperparameter is the number of layers that should be edited. We confirm the findings of Limisiewicz et al. (2024) that adaptation should applied to approximately

[2]In baseline *DAMA*, the projection dimensionality is preset to $d = 256$ for the 7B model and $d = 512$ for the 13B models.



(a) Bias-to-feature threshold fixed at 0.05 (b) Number of layers fixed at 12

Figure 4: The hyperparameter analysis for *2DAMA* applied to Llama 2 13B model on performance and bias in language modeling. We measured bias on gendered prompts by linear coefficients: a_s and a_f, the language modeling capabilities are measured by perplexity. Stars mark the performance of the best setting. The dashed line corresponds to the scores of the original model.

one-third of the midd-upper layers. Notably, the top two layers (38th and 39th) should be left out.

5 Beyond English: Multilingual Debiasing

In a multilingual setting, we debias a model fine-tuned for translation: ALMA-R 13B (Xu et al., 2024) by employing the collection of the new multilingual debiasing prompts. We specifically evaluate gender bias and quality of translation between English and Czech, German, and Russian.[3]

5.1 Main Results

Model editing generalizes to the multilingual settings. Analogically to experiments for English, we show that model editing reduces bias in translation and has a small impact on the translation quality (as shown in Table 4). We observe some differences in results between the two analyzed languages. Overall, the scores after debiasing are better for German than Czech, indicating that German prompts are of better quality.

***Dual Debiasing* is required to mitigate representational bias.** Our methods are more effective for the stereotypical manifestation of bias ΔS than the representational one ΔG. In the representational bias, we sometimes observe bias increase after model editing. To remedy that, we use *2DAMA* with higher values of feature-to-bias threshold ($t = 1.00$ instead of $t = 0.05$), which tends to preserve more factual signal. Factual gender understanding is especially essential for equitable representation of factual gender in morphologically rich languages, as evidenced by ΔG scores for $t = 1.00$ setting. This finding emphasizes the utility of *2DAMA* in a multilingual setting. [4]

5.2 Cross-lingual Debiasing

An intriguing question of multilingual bias is whether its encoding is shared across languages (Gonen et al., 2022). We test this hypothesis by editing models with prompts in one or multiple languages and testing on another language. The results show evidence of effectiveness in cross-lingual mitigation of stereotypical gender bias. In Table 5b, we observe that some languages are more effective in debiasing than others, e.g., German prompts offer the strongest ΔS reduction for both Czech and German. Whereas to control representational bias mitigation (ΔG), it is recommended to use in-language prompts, as indicated by Czech, German, and Russian results in Table 5a.

6 Related Work

6.1 Model Editing and Concept Erasure

Model editing is a method of applying targeted changes to the parameters of the models to modify information encoded in them. Notable examples of model editing include targeted changes in the model's weight (Mitchell et al., 2022; Meng et al., 2023, 2022) or adaptation with added modules (adapters) (Houlsby et al., 2019; Hu et al., 2022). The technique showed promising results as the tool to erase specific information (Patil et al.,

[3]We do not perform debiasing on Russian prompt (as indicated in Section 3.2. We include Russian in evaluation to observe if debiasing generalizes across languages.

[4]The extended study of hyperparameters in translation debiasing is presented in Appendix C.

Language		Translation to English		Translation from English		WinoMT		BUG	
		↑ comet	↑ chrf	↑ comet	↑ chrf	↓ ΔS	↓ ΔG	↓ ΔS	↓ ΔG
German	ALMA-R 13B	85.0	57.0	86.7	58.1	30.5	3.7	7.8	32.5
	DAMA+LEACE	85.0	56.7	85.3	55.4	20.5	10.0	5.4	33.6
	2DAMA ($t = 0.05$)	84.9	56.7	85.1	54.8	22.6	3.3	4.4	27.8
	2DAMA ($t = 1.00$)	84.9	56.6	85.4	55.4	22.1	-10.1	7.7	28.4
Czech	ALMA-R 13B	87.0	68.6	89.7	53.8	26.3	2.1	11.7	9.2
	DAMA+LEACE	86.9	68.2	88.6	50.1	21.6	17.7	10.4	18.0
	2DAMA ($t = 0.05$)	86.9	68.1	88.5	49.9	18.0	14.6	4.5	11.0
	2DAMA ($t = 1.00$)	86.9	68.1	88.8	50.4	22.4	7.2	8.6	9.8

Table 4: Evaluation of gender bias and quality of translation. In all the methods, ALMA-R was used as the base model. Adaptations were applied to 11 mid-upper feed-forward layers. Translation quality was evaluated on the WMT-22 dataset.

Prompt Lang. ↓	German	Czech	Russian
∅	3.7	2.1	25.7
English	11.1	7.9	31.4
German	3.3	21.6	31.3
Czech	6.2	14.6	32.0
All Above	8.1	23.2	33.4

(a) Representational bias (ΔG)

Prompt Lang. ↓	German	Czech	Russian
∅	30.5	26.3	10.2
English	28.5	21.2	7.0
German	14.4	15.1	4.0
Czech	24.3	17.2	3.9
All Above	24.0	18.7	1.3

(b) Stereotypical bias (ΔS)

Table 5: Bias evaluation based on WinoMT challenge-set. The evaluation language is shown at the top of each column. Each row corresponds to a set of languages for which prompts were used in model adaptation (∅ denotes the model without any adaptation). The debiasing adaptation was performed with *2DAMA* on 11 mid-upper layers with the bias-to-feature threshold set to $t = 0.05$.

2024).

In the literature, bias mitigation was perceived as a theoretically interesting and practical application for concept erasure. Ravfogel et al. (2020, 2022); Belrose et al. (2023) proposed effective linear methods of erasing gender bias from the latent representation of language models. Other approaches aimed to edit pre-trained language models to reduce their reliance on stereotypes. They include: causal intervention (Vig et al., 2020), model adapters (Fu et al., 2022), rate-distortion (Chowdhury and Chaturvedi, 2022), or targeted weight editing (Limisiewicz et al., 2024).

6.2 Debiasing Machine Translation

Machine translation systems have been shown to exhibit gender bias in their predictions (Savoldi et al., 2021). The problem is especially severe in translation from languages that do not grammatically mark gender (e.g., English, Finnish) to ones that do (e.g., German, Czech, Spanish) because translation requires predicting gender, which is not indicated in the reference (Stanovsky et al., 2019). There have been a few past attempts to mitigate biases in translation systems (Saunders and Byrne, 2020; Iluz et al., 2023; Zmigrod et al., 2019). Nevertheless, these approaches are based on fine-tuning for non-stereotypical sentences, which increases the model's specialization but significantly reduces usability, e.g., in tasks unrelated to gender (Luo et al., 2023).

One key constraint of multilingual debiasing is the scarcity of bias annotations in various languages. Notable datasets were introduced by Levy et al. (2021); Névéol et al. (2022). The difficulty of obtaining reliable cross-lingual bias resources stems from the need for deep knowledge of culture in addition to understanding a language. To the best of our knowledge, we are the first to propose a method for debiasing LLM in machine translation tasks.

7 Conclusion

We highlight the importance of considering the dual character of gender encoding in model editing. The theoretical and empirical results show that our novel model editing methods: *2DAMA* effectively reduces the impact of stereotypical bias on the predictions while preserving equitable representation of (factual) gender based on grammar and semantics. Maintaining the factual component of gender representation is crucial for debiasing in languages other than English, for which gender markings are ubiquitous. Furthermore, our method does not significantly deteriorate the high performance of LLMs in various evaluation settings unrelated to gender.

Limitations

The main drawback of the *Dual Debiasing* approach is the high likelihood of stereotypical and factual signals being correlated, as mentioned in Section 2.3. We hypothesize that the model attained this correlation from training data because the distinction between factual and stereotypical gender cues is often vague and depends on context. Nevertheless, we show that with *Dual Debiasing* we can control the tradeoff, and with proper choice of hyperparameters, we can keep strong factual signals while discarding the majority of bias.

Another drawback of our method is that we observe a small deterioration in non-gender-related tasks, such as language modeling and translation. Some of the drop may be attributed to the fact that test sets may exhibit representational bias. For instance, there could be a higher frequency of male than female mentions, which would unfairly advantage a biased model in evaluation.

Acknowledgements

This work was supported by the project "Human-centred AI for a Sustainable and Adaptive Society" (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union. It was also supported by the grant 23-06912S of the Czech Science Foundation. We have been using language resources and tools developed, stored, and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

References

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. 2022. Learning fair representations via rate-distortion maximization. *Transactions of the Association for Computational Linguistics*, 10:1159–1174.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem 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* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128. ACM.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona

Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Morris L. Eaton. 1983. The Gauss-Markov Theorem in

Multivariate Analysis. Technical report, University of Minnesota.

Chin-Lun Fu, Zih-Ching Chen, Yun-Ru Lee, and Hung-yi Lee. 2022. AdapterBias: Parameter-efficient Token-dependent Representation Shift for Adapters in NLP Tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2608–2621. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Hila Gonen, Shauli Ravfogel, and Yoav Goldberg. 2022. Analyzing gender representation in multilingual models. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 67–77, Dublin, Ireland. Association for Computational Linguistics.

Marlis Hellinger and Hadumod Bußmann, editors. 2002. *Gender Across Languages: The Linguistic Representation of Women and Men, Volume 2*, volume 2 of *Impact: Studies in Language and Society*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Bar Iluz, Tomasz Limisiewicz, Gabriel Stanovsky, and David Mareček. 2023. Exploring the impact of training data distribution and subword tokenization on gender bias in machine translation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 885–896, Nusa Dua, Bali. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in large language models. *Proceedings of The ACM Collective Intelligence Conference*.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomasz Limisiewicz and David Mareček. 2022. Don't forget about pronouns: Removing gender bias in language models without losing factual gender information. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 17–29, Seattle, Washington. Association for Computational Linguistics.

Tomasz Limisiewicz, David Mareček, and Tomáš Musil. 2024. Debiasing algorithm through model adaptation. In *The Twelfth International Conference on Learning Representations*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *ArXiv*, abs/2308.08747.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *NeurIPS*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-Editing Memory in a Transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Vaidehi Patil, Peter Hase, and Mohit Bansal. 2024. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In

The Twelfth International Conference on Learning Representations.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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 Cotterell. 2022. Linear Adversarial Concept Erasure. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *CoRR*, abs/2112.14168.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, D. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, A. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. 2022. The birth of bias: A case study on the evolution of gender bias in an English language model. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 75–75, Seattle, Washington. 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. *CoRR*, abs/2004.12265.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *Preprint*, arXiv:2401.08417.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Proofs

A.1 Terminological Note

For brevity of theorems and proofs, we adopt the following notation convention:

Definition 1 (**Moore-Penrose Pseudoinvers**). *We denote Moore-Penrose pseudoinverse of matrix $\boldsymbol{M}$ as $\boldsymbol{M}^+$:*

$$\boldsymbol{M}^+ = (\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T$$

Definition 2 (**Matrix Square Root**). *We denote a positive semi-definite square root of positive semi-definite matrix $\boldsymbol{M}$ as $\boldsymbol{M}^{1/2}$.*

Definition 3 (**Covariance Matrix**). *For two random vectors: $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$. We denote the covariance matrix as:*

$$\boldsymbol{\Sigma}_{X,Y} = \mathbb{C}\mathrm{ov}(X, Y)$$

A.2 LEACE Theorem

For reference, we present the original *LEACE* theorem from Belrose et al. (2023). The proof can be found in ibid..

Theorem 3 (**LEACE**). *We consider random vectors X and Z taking values in $\mathbb{R}^n$. Both random vectors are centered, each with a finite moment.*

Then the objective:

$$\underset{\boldsymbol{P}\in\mathbb{R}^{n\times n}}{\arg\min}\, \mathbb{E}\left[||\boldsymbol{P}X - X||^2\right]$$

subject to:

$$\mathbb{C}\mathrm{ov}(\boldsymbol{P}X, Z) = 0$$

is solved by:

$$\boldsymbol{P}^* = \mathbb{I} - \boldsymbol{W}^+\boldsymbol{P}_{\boldsymbol{W\Sigma}}\boldsymbol{W},$$

where $\boldsymbol{W}$ is the whitening transformation $(\Sigma_{V,V}^{1/2})^+$; $\boldsymbol{P}_{\boldsymbol{W\Sigma}}$ is an orthogonal projection matrix onto colspace of $\boldsymbol{W}\boldsymbol{\Sigma}_{V,Z}$.

A.3 Proof for DAMA-LEACE Theorem

We formalize the requirements and implications of that assumption in the following theorem:

Theorem 4 (**Gauss-Markov: Probabilistic Least Squares**). *We consider random vectors: U taking values in $\mathbb{R}^m$, V, and Z taking values in $\mathbb{R}^n$; both are centered and have finite second moments. We seek the linear regression model given by:*

$$V = \boldsymbol{S}U - \epsilon,$$

given the following assumptions:

A ***No Multicollinearity****: there is no linear relationship among the independent variables, i.e., matrix $\Sigma_{U,U}$ is of full rank m.*

B ***Exogeneity****: the expected value of error terms given independent variables $\mathbb{E}[\epsilon|U] = 0$, this also implies that $\mathbb{C}\mathrm{ov}(\epsilon, U) = 0$.*

C ***Homoscedasticity****: the covariance of the error terms is constant and does not depend on the independent variables $\mathbb{C}\mathrm{ov}(\epsilon, \epsilon|U) = \sigma\,\mathbb{I}$.*

Then, the ordinary least squares estimator is given by the formula:

$$\boldsymbol{S}^* = \boldsymbol{\Sigma}_{U,V}\boldsymbol{\Sigma}_{U,U}^{-1}$$

Such estimator is ***best linear unbiased estimator*** *and minimizes the variance of error terms: $Tr(\mathbb{C}\mathrm{ov}(\epsilon, \epsilon))$.*

The proof of the Theorem 4 can be found in the classical statistics literature. For instance, Eaton (1983) presents proof for the multivariate case presented above.

Equipped with the theorems above, we are ready to present the theorem that is the main theoretical contribution of this work:

Theorem 1. *We consider random vectors: U taking values in $\mathbb{R}^m$, V and Z taking values in $\mathbb{R}^n$, where $m \geq n$. Under assumptions that: A) random vectors U, V, Z are centered, and each of them has finite moment; B) the regression relation between U and V fulfill the assumption of ordinary least squares, and there exist least squares estimator $V = \boldsymbol{S}U - \epsilon$.*

Then the objective:

$$\underset{\boldsymbol{P}\in\mathbb{R}\,n\times m}{\arg\min}\, \mathbb{E}\left[||\boldsymbol{P}U - V||^2\right],$$

subject to:

$$\mathbb{C}\mathrm{ov}(\boldsymbol{P}U, Z) = 0$$

is solved by:

$$\boldsymbol{P}^* = \left(\mathbb{I} - \boldsymbol{W}^+\boldsymbol{P}_{\boldsymbol{W\Sigma}}\boldsymbol{W}\right)\boldsymbol{S},$$

where $\boldsymbol{W}$ is the whitening transformation $(\Sigma_{\boldsymbol{S}U,\boldsymbol{S}U}^{1/2})^+$; $\boldsymbol{P}_{\boldsymbol{W\Sigma}}$ is an orthogonal projection matrix onto colspace of $\boldsymbol{W}\boldsymbol{\Sigma}_{\boldsymbol{S}U,Z}$; $\boldsymbol{S}$ is a least squares estimator of V given U: $\boldsymbol{S} = \boldsymbol{\Sigma}_{U,V}\boldsymbol{\Sigma}_{U,U}^{-1}$.

Proof. For simplicity, we will decompose the problem into independent optimization objectives corresponding to each dimension in $\mathbb{R}^n$. For the ith dimension, we write the objective as:

$$\arg\min_{\boldsymbol{P}_i \in \mathbb{R}^n} \mathbb{E}\left[\boldsymbol{P}_i^T V - V_i\right]^2 \quad \text{s.t.} \quad \mathbb{C}\text{ov}(\boldsymbol{P}_i U, Z) = 0, \tag{3}$$

where $\boldsymbol{P}_i$ is ith column of matrix $\boldsymbol{P}$. From the assumption (B) of the theorem, we can represent the linear relation between U and V, as $\boldsymbol{S}U = V+\epsilon$, where ϵ is an error term of regression. We use this property to rewrite the minimization objective from expression 3, as:

$$\arg\min_{\widetilde{\boldsymbol{P}_i} \in \mathbb{R}^n, \boldsymbol{S} \in \mathbb{R}^{m \times n}} \mathbb{E}\left[\widetilde{\boldsymbol{P}_i}^T \boldsymbol{S}U - V_i\right]^2 \tag{4}$$

We manipulate the term under $\arg\min$ to rewrite it as a sum of three terms:

$$\begin{aligned}\mathbb{E}\left[\widetilde{\boldsymbol{P}_i}^T \boldsymbol{S}U - V_i\right]^2 &= \mathbb{E}\left[\widetilde{\boldsymbol{P}_i}^T (V+\epsilon) - V_i\right]^2 = \\ &= \mathbb{E}\left[\widetilde{\boldsymbol{P}_i}^T (V+\epsilon) - (V_i + \epsilon_i) + \epsilon_i\right]^2 = \\ &= \underbrace{2E\left[\left(\widetilde{\boldsymbol{P}_i}^T (V+\epsilon) - (V_i+\epsilon_i)\right)\epsilon_i\right]}_{\text{I}} + \\ &+ \underbrace{\mathbb{E}[\epsilon_i]^2}_{\text{II}} + \underbrace{\mathbb{E}\left[\widetilde{\boldsymbol{P}_i}^T (V+\epsilon) - (V_i+\epsilon_i)\right]^2}_{\text{III}}\end{aligned} \tag{5}$$

We will now consider each of the three summands one by one to find the solution to the optimization objective $P^* = \widetilde{P}^* S^*$.

Summand I zeros out. We show that by observing that the summand is doubled covariance[5]:

$$\begin{aligned}&E\left[\left(\widetilde{\boldsymbol{P}_i}^T (V+\epsilon) - (V_i+\epsilon_i)\right)\epsilon_i\right] = \\ &= \mathbb{C}\text{ov}\left(\widetilde{\boldsymbol{P}_i}^T (V+\epsilon) - (V_i+\epsilon_i), \epsilon_i\right) = \\ &= \left(\widetilde{\boldsymbol{P}_i}^T - \mathbb{1}_i^T\right)\mathbb{C}\text{ov}(V+\epsilon, \epsilon) = \\ &= \left(\widetilde{\boldsymbol{P}_i}^T - \mathbb{1}_i^T\right)\boldsymbol{S}\,\mathbb{C}\text{ov}(U, \epsilon)\end{aligned} \tag{6}$$

From assumption B of Theorem 4 (exogeneity) and, by extension, assumption of this theorem, we have that $\mathbb{C}\text{ov}(U, \epsilon) = 0$ and thus Summand I zeros out.

[5]From the fact that both factors under $\mathbb{E}$ are centered.

Summand II by the conclusion of Theorem 4 is minimized by setting:

$$\boldsymbol{S}^* = \boldsymbol{\Sigma}_{\boldsymbol{U},\boldsymbol{V}} \boldsymbol{\Sigma}_{\boldsymbol{U},\boldsymbol{U}}^{-1} \tag{7}$$

We can also set $\boldsymbol{S}$ to $\boldsymbol{S}^*$ in summand III, as the variable under $\mathbb{E}$ is independent of ϵ, as shown in the previous paragraph. By finding $\boldsymbol{S}^*$, we have solved part of the objective in expression 4.

Summand III we find the matrix $\widetilde{P}$ minimizing the value of the summand under constraines. By rewriting $\mathbb{C}\text{ov}(P_i U, Z)$ as $\mathbb{C}\text{ov}(\widetilde{\boldsymbol{P}_i}(V+\epsilon), Z)$, we observe that minimizing the value of summand III under constraint is analogical to solving the problem stated in LEACE (Theorem 3):

$$\begin{gathered}\arg\min_{\widetilde{\boldsymbol{P}_i} \in \mathbb{R}^n} \mathbb{E}\left[\widetilde{\boldsymbol{P}_i}^T (V+\epsilon_i) - (V_i+\epsilon)\right]^2 \\ \text{such that} \quad \mathbb{C}\text{ov}(\widetilde{\boldsymbol{P}}_i(V+\epsilon), Z) = 0\end{gathered} \tag{8}$$

We find the solution based on Theorem 3, where we substitute X with $V+\epsilon$ and find $\widetilde{\boldsymbol{P}}^* = \mathbb{I} - \boldsymbol{W}^+ \boldsymbol{P}_{\boldsymbol{W}\boldsymbol{\Sigma}} \boldsymbol{W}$, where $\boldsymbol{W}$ is the whitening transformation $(\Sigma_{V+\epsilon,V+\epsilon}^{1/2})^+$; $\boldsymbol{P}_{\boldsymbol{W}\boldsymbol{\Sigma}}$ is an orthogonal projection matrix onto colspace of $\boldsymbol{W}\boldsymbol{\Sigma}_{V+\epsilon,Z}$

Conclusion for summands II and III, we independently found the matrices minimizing their values. We obtain the matrix P^* solving our original objective in expression 3 by multiplying them:

$$P^* = \widetilde{P}^* S^* = \left(\mathbb{I} - \boldsymbol{W}^+ \boldsymbol{P}_{\boldsymbol{W}\boldsymbol{\Sigma}} \boldsymbol{W}\right) \boldsymbol{\Sigma}_{U,V} \boldsymbol{\Sigma}_{U,U}^{-1} \tag{9}$$

□

A.4 Proof for Dual-Debiasing Theorem

Theorem 2. *We consider random vectors X, Z_b, and Z_f in $\mathbb{R}^n$. Under the assumptions that: A) Z_b and Z_f $Z_b \perp Z_f | X$, i.e., Z_b and Z_f are conditionally independent, given X; B) $\boldsymbol{\Sigma}_{X,Z_b} \boldsymbol{\Sigma}_{X,Z_f}^T = 0$, i.e., the variable X is correlated with Z_f and Z_b through mutually orthogonal subspaces of $\mathbb{R}^n$. The solution of the objective:*

$$\arg\min_{\boldsymbol{P} \in \mathbb{R}^{n \times n}} \mathbb{E}\left[||\boldsymbol{P}X - X||^2\right],$$

subject to:

$$\mathbb{C}\text{ov}(\boldsymbol{P}X, Z_b) = 0,$$

satisfies:

$$\mathbb{C}\text{ov}(\boldsymbol{P}X, Z_f) = \mathbb{C}\text{ov}(X, Z_f).$$

Proof. First, we observe that the assumption A) can be generalized to any coordinate system. For an orthogonal matrix $\boldsymbol{W}$, we have:

$$\boldsymbol{\Sigma}_{\boldsymbol{W}X,Z_b}\boldsymbol{\Sigma}_{\boldsymbol{W}X,Z_f}^T = \boldsymbol{W}\boldsymbol{\Sigma}_{X,Z_b}\boldsymbol{\Sigma}_{X,Z_f}^T\boldsymbol{W}^T = 0 \tag{10}$$

This guarantees the orthogonality of spaces spanned by columns of two orthogonality matrices. The property will be useful for the second part of the proof:

$$Col(\boldsymbol{\Sigma}_{\boldsymbol{W}X,Z_b}) \perp Col(\boldsymbol{\Sigma}_{\boldsymbol{W}X,Z_f}) \tag{11}$$

Secondly, we remind the reader that the solution to the objective provided in the theorem (based on Theorem 3) is as follows:

$$\boldsymbol{P}^* = \mathbb{I} - \boldsymbol{W}^+ \boldsymbol{P}_{\boldsymbol{W}\boldsymbol{\Sigma}} \boldsymbol{W} \tag{12}$$

Now, we evaluate the covariance matrix between $\boldsymbol{P}^*X$ and Z_f to check that it is the same as the covariance matrix between X and Z_f.

$$\begin{aligned} \mathbb{C}\mathrm{ov}(\boldsymbol{P}^*X, Z_f) &= \\ = \boldsymbol{\Sigma}_{X,Z_f} - \boldsymbol{W}^+\boldsymbol{P}_{\boldsymbol{W}\boldsymbol{\Sigma}}\boldsymbol{W}\boldsymbol{\Sigma}_{X,Z_f} &= \\ = \boldsymbol{\Sigma}_{X,Z_f} - \boldsymbol{W}^+\boldsymbol{P}_{\boldsymbol{W}\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_{\boldsymbol{W}X,Z_f} & \end{aligned} \tag{13}$$

we note that $\boldsymbol{P}_{\boldsymbol{W}\boldsymbol{\Sigma}}$ is the projection matrix onto the column space of $\boldsymbol{\Sigma}_{\boldsymbol{W}X,Z_f}$. From that fact and Equation 11, we have:

$$\boldsymbol{P}_{\boldsymbol{W}\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_{\boldsymbol{W}X,Z_f} = 0 \tag{14}$$

Thus the last component in Equation 13 nullifies and we conclude that:

$$\mathbb{C}\mathrm{ov}(\boldsymbol{P}^*X, Z_f) = \boldsymbol{\Sigma}_{X,Z_f} = \mathbb{C}\mathrm{ov}(X, Z_f) \tag{15}$$

□

B Prompts

B.1 Monolingual Prompts

The list of 11 prompt templates is given in Table 6. The term <profession> is substituted by 219 professions without factual gender (from Bolukbasi et al., 2016) and 26 gendered entities ("*man*", "*boy*", "*gentleman*", "*father*", "*son*", "*brother*", "*husband*", "*king*", "*prince*", "*uncle*", "*nephew*", "*groom*", "*duke*", "*grandfather*", "*widower*", "*hero*", "*wizard*", "*baron*", "*emperor*", "*anchorman*", "*clergyman*", "*chairman*", "*alumnus*", "*woman*", "*girl*", "*lady*", "*mother*", "*daughter*", "*sister*", "*wife*", "*queen*", "*princess*", "*aunt*", "*niece*", "*bride*", "*duchess*", "*grandmother*", "*widow*", "*heroine*", "*witch*", "*baroness*", "*empress*", "*anchorwoman*", "*clergywomen*", "*chairwomen*", "*alumna*") which makes 2409 stereotypical and 286 factual prompts in total used for *Dual Debiasing*.

B.2 Multilingual Prompts

For debiasing the translation models, we use 11 English-German and 11 English-Czech prompt templates shown in Table 7. These prompts were designed to be as general as possible, and so that the target language prefix does not include any gender information, while the proposed completions, which are always one-token words, do specify the gender. In German, these completions may be determiners. Czech does not use determiners, but gender is included, for example, in the past form of the verb "to be". We use the same sets of professions, and therefore, we generate additional data of equal size for each language. To match the intended use, we wrap them into ALMA-R translation prompt template:

Translate this from {src_lang} to {tgt_lang}:
{src_lang}: <English source>
{tgt_lang}: <Target prefix>

English prompt	Completions
The <profession> wanted that	[he, she, they]
The <profession> laughed because	[he, she, they]
The <profession> went home because	[he, she, they]
The <profession> desired that	[he, she, they]
The <profession> wished that	[he, she, they]
The <profession> cried because	[he, she, they]
The <profession> ate because	[he, she, they]
The <profession> said that	[he, she, they]
The <profession> ran because	[he, she, they]
The <profession> stayed up because	[he, she, they]
The <profession> whispered that	[he, she, they]

Table 6: Monolingual English prompt templates.

C Additional Results

C.1 Stereotypical and Factual Signals across Layers

In Figure 6, we observe the variances with stereotypical and factual gender signals in subsequent layers. We see that the number of biased dimensions differs across layers. Nevertheless, we observe the same pattern in each layer: stereotypical signal is encoded in a relatively small number of

English source	German prefix	Completions
This is the <profession>.	Das ist ___	[der, die]
There is the <profession>.	Da ist ___	[der, die]
The <profession> is not working today.	___	[Der, Die]
The <profession> was fired.	___	[Der, Die]
The <profession> is busy.	___	[Der, Die]
Do you know the <profession>	Kennen Sie ___	[den, die]
I was there with the <profession>	Ich war dort mit ___	[dem, der]
I asked the <profession>.	Ich fragte ___	[den, die]
We met the <profession>.	Wir trafen ___	[den, die]
I answered the <profession>.	Ich antwortete ___	[dem, der]
The salary of the <profession> has increased.	Das Gehalt ___	[des, der]

English source	Czech prefix	Completions
This is that <profession>.	To je ___	[ten, ta]
There is that <profession>.	Tam je ___	[ten, ta]
That <profession> is not working today.	___	[Ten, Ta]
That <profession> was fired.	___	[Ten, Ta]
That <profession> is busy.	___	[Ten, Ta]
I was a <profession> two years ago.	Před dvěma lety jsem ___	[byl, byla]
You were a <profession> two years ago.	Před dvěma lety jste ___	[byl, byla]
If only I were a <profession>.	Kdybych tak ___	[byl, byla]
I was a <profession> at that time.	V té době jsem ___	[byl, byla]
You were a <profession> at that time.	V té době jsi ___	[byl, byla]
You were a <profession> at that time.	V té době jste ___	[byl, byla]

Table 7: Multilingual prompt templates for English-to-German and English-to-Czech translation

dimensions with high variance, while the stereotypical variance is spread across more dimensions with lower values in each.

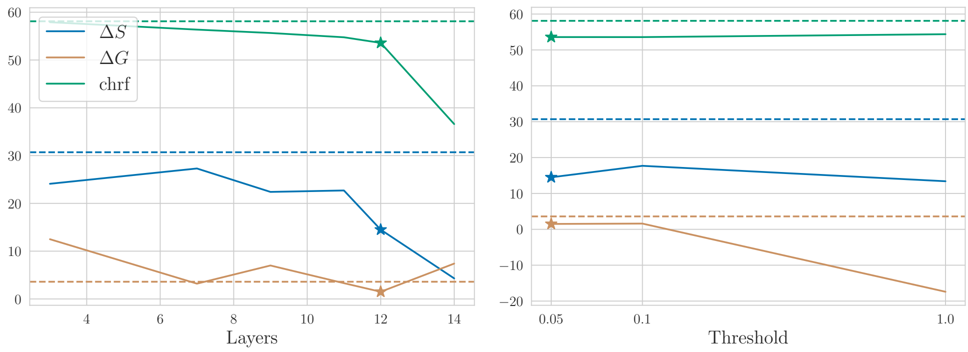
C.2 Choice of Hyperparameters in Translation

Analogically to Section 4.3, we present the impact of bias-to-feature threshold t and the number of edited layers on translation to German in Figure 5. We observe that stronger factual regularization (high t) helps in reducing representational bias (ΔG) yet offers weaker stereotypical bias mitigation (ΔS). Similar to the results in language modeling, the best performance is obtained when editing 12 mid-upper layers with $t = 0.05$.

D Technical Details

To find the value representation V, we run gradient optimization for 20 steps with Adam scheduler (Kingma and Ba, 2015) and learning rate: $lr = 0.5$. We picked the following regularization constants: $\lambda_1 = 0.0625$ and $\lambda_2 = 0.2$.

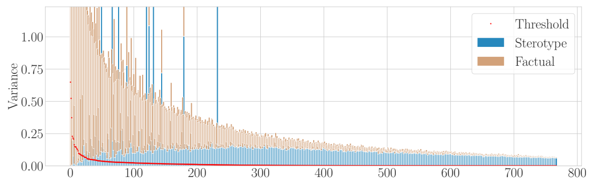
The optimization was run on a Nvidia A40 GPU. For Llama 2 7B, processing one prompt took around 10 seconds.



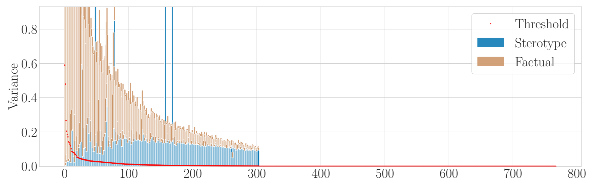
(a) Bias-to-feature threshold fixed at 0.05

(b) Number of layers fixed at 12

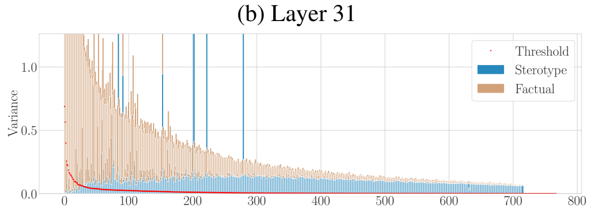
Figure 5: The hyperparameter analysis for *2DAMA* applied to ALMA-R 13B model on performance and bias in translation to German. We measured bias via WinoMT metrics ΔS and ΔG. The translation quality to Germna is measured by chrf on WMT-22. Stars mark the performance of the best setting. The dashed line corresponds to the scores of the original model.



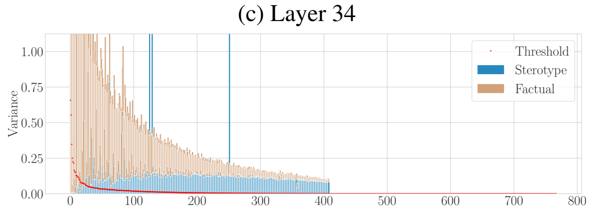
(a) Layer 28



(b) Layer 31



(c) Layer 34



(d) Layer 37

Figure 6: Visualization of dimensions and their variances related to stereotypical and factual gender signals identified by *Dual Debiasing* algorithm across different layers of Llama 2 13B.