MULTIGROUP SOCIAL DEBIASING OF LLMS VIA NEURON DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Mitigating social biases in language models is a prerequisite for their responsible deployment. Many benchmarks and debiasing methods operationalize bias as a difference between two groups, a target group and a single alternative, so a model can reduce this pairwise gap while its preferences over the remaining groups stay biased. We define social bias as the divergence from uniform of the distribution a model induces over a set of demographic groups, with the completion held fixed so that any preference is attributable to the group. To this end, we introduce `Multigroup-StereoSet`, an extension of `StereoSet` in which each target group is complemented by distractor groups selected for maximal dispersion over real-world taxonomies, while stereotypical and anti-stereotypical completions are held fixed. We then propose DAID, a two-stage method that first identifies the MLP neurons whose suppression moves the group distribution toward uniform and then fine-tunes only these neurons' parameters. Across four models on English text, DAID reduces the median max/min group-probability ratio from 23–35 to 1.3–2.3 on *nationality* and from 97–138 to below 1.5 on *occupation*, and increases `StereoSet` ICAT by 6–16 points, at a cost of 1–5.4 points on `MMLU`. Model-editing and activation-intervention baselines raise ICAT by up to 15 points but leave the ratio on *nationality* between 14 and 121. We find that suppressing the identified neurons without training also yields near-uniform distributions but degrades language modelling capabilities.

1 INTRODUCTION

Large language models encode social presuppositions present in their training corpora [Dodge et al. 2021, Caliskan et al. 2017], and these associations persist in recent models even when their behaviour appears unbiased [Hofmann et al. 2024, Bai et al. 2025]. Most bias benchmarks, and the debiasing approaches evaluated on them, measure bias as a contrast between two groups, a target group and a single alternative [Nadeem et al. 2021, Nangia et al. 2020]. Critiques of this literature find such benchmarks internally inconsistent and their encoded biases rarely grounded in specific harms [Field et al. 2021, Blodgett et al. 2020, 2021, Hanna et al. 2020]. Furthermore, their problem formulation assumes a single stereotype per target and treats all other groups as one alternative, so a model can reduce the pairwise gap while its preferences over the remaining groups stay biased.

Consider a fixed continuation such as "He loved to drink Vodka." following a context that names a demographic group. A pairwise benchmark compares the probability of this continuation for the target group with its probability for a single alternative group [Nangia et al. 2020], or with the probability of an anti-stereotypical continuation for the same group [Nadeem et al. 2021]. However, a model can assign similar probability to the target group and its alternative while strongly favouring or disfavouring other groups.

We therefore define social bias as the divergence from uniform of the distribution a model induces over a *set* of demographic groups. Because the completion is held fixed, any deviation from uniform is attributable to the group named in the context. Therefore, we construct `Multigroup-StereoSet`, an extension of the intersentence split of `StereoSet` [Nadeem et al. 2021] in which each target group is complemented by distractor groups selected for maximal dispersion. We then propose DAID, a two-stage debiasing method (Fig. 1). First, a greedy search over `Multigroup-StereoSet` identifies the MLP neurons whose suppression moves the group distribution closest to uniform, using a first-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

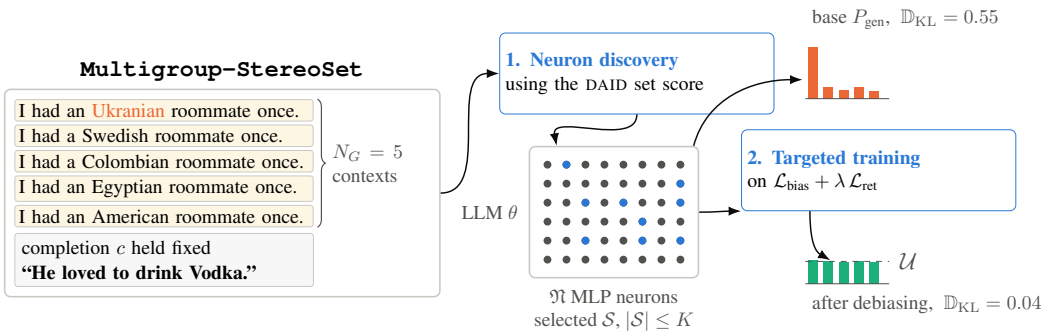

Figure 1: The debiasing procedure consists of two steps. The first is neuron identification within the intermediate representation of the MLP modules of a transformer, followed by a targeted fine-tuning of corresponding MLP down projection parameters. The debiasing procedure utilizes the `Multigroup-StereoSet` extension built from real-world social cues. The reported divergences are the MultiGroup Divergence (MGD), taken with respect to the uniform distribution $\mathcal{U}$.

order approximation that costs one forward and one backward pass per example. Second, we fine-tune only the down-projection columns of these neurons to minimize the divergence of the group distribution from uniform.

We evaluate our method on 4 open-source models—`Qwen2.5-1.5B` [Qwen Team 2025], `Llama-3.2-3B` [Grattafiori et al. 2024], `phi-2` [Javaheripi et al. 2023], `gemma-2-2b-it` [Gemma Team 2024]—across 2 demographic axes, *nationality* and *occupation*, for English text, against prompting, re-weighting, model-editing, and activation-intervention baselines. On nationality, DAID reduces the median group-probability ratio from 23–35 to 1.3–2.3 and increases `StereoSet` ICAT by 6–16 points. On occupation, evaluated on two models, it reduces the ratio from 97–138 to below 1.5. In comparison, model-editing and activation-intervention baselines increase ICAT by up to 15 points but leave the ratio on nationality between 14 and 121. We also find that suppressing the selected neurons without training reduces the median ratio of `Qwen2.5-1.5B` from 23 to 1.7, but decreases `StereoSet` ICAT from 70.4 to 60.3 and, under mean ablation, reduces `MMLU` accuracy to chance.

To summarize, our main contributions are:

1. A multigroup formulation of social bias as the divergence from uniform over a demographic set, and `Multigroup-StereoSet`, a dataset in which the completion is fixed and the societal group varies over maximally dispersed real-world groups.
2. An empirical comparison showing that improvements on pairwise bias benchmarks do not transfer to multigroup parity.
3. DAID, a two-stage method that identifies MLP neurons by their effect on the group distribution and fine-tunes only their parameters, which yields near-uniform group distributions at a small cost in general capability.

2 Related Work

2.1 Benchmarking Societal Bias

Extrinsic measurement of social bias in AI systems evaluates the downstream effects and real-world impact of biased behaviour. This relies on comprehensive risk assessment [Department for Digital, Culture, Media & Sport 2026] and it is hard to measure. In contrast, intrinsic benchmarks simulate this real-world impact, serving as controllable testbeds for debiasing and mitigation techniques. `StereoSet` [Nadeem et al. 2021] and `CrowS-Pairs` [Nangia et al. 2020] pair examples into a stereotypical and anti-stereotypical sentences. Other benchmarks like `BOLD` [Dhamala et al. 2021] measure sentiment and regard scores for open-ended generations, and `BBQ` [Parrish et al. 2022] pose MCQs under ambiguous and disambiguated contexts. `SoFa` [Marchiori Manerba et al. 2024], on

the other hand, probe models by assessing their disparate treatment across identities within several social axes. We adopt this multigroup view, but differ both in construction and use. `SoFa` concatenates a predefined set of identities with short stereotypical assertions, held fixed across examples. We instead paraphrase the naturally occurring contexts of `StereoSet` and select the contrast groups per example by maximizing demographic dispersion from the target. The resulting dataset, `Multigroup-StereoSet`, serves as the training signal for our debiasing objective.

2.2 Debiasing techniques

We group debiasing methods by their intervention region. The first line of work updates model weights. Auto-Debias [Guo et al. 2022] fine-tunes encoder models on masked prompts to equalize output distributions across groups. Model editing, developed to update factual knowledge [Gupta et al. 2024, Tan et al. 2024], has also been adapted to debiasing. BiasEdit [Xu et al. 2025] trains editor hyper-networks to emit parameter updates for a fixed subset of feed-forward blocks, supervised by the stereotype *v/s* anti-stereotype pair together with a retention loss. Our method also updates a small subset of feed-forward parameters, but it selects these parameters by search instead of freezing them in advance, and it optimizes a multigroup objective instead of a pairwise one.

A second line of work targets internal activations. Liu et al. [2024] attribute social bias to neurons in masked language models using integrated gap gradients, adapted from input feature attribution [Sundararajan et al. 2017]. Lin et al. [2026] attribute bias in two directions within the output projections of decoder transformers and set the identified units to fixed values at inference. Both approaches use localization to intervene at inference and leave the parameters unchanged. Zhao et al. [2025] instead fine-tune only the neurons they identify as relevant to safety alignment. We follow this strategy for social bias and use localization to restrict a parameter update trained on a multigroup objective.

A third line of work operates at inference time without training, through prompting or token reweighting during decoding [Schick et al. 2021]. DDP [Li et al. 2025] analyzes how social context influences model decisions and constructs mitigation prompts from this analysis. These methods require no training, but Yang et al. [2025] argue that their gains on bias benchmarks overstate the actual reduction in bias.

3 Methodology

Social bias in a model can be defined as its tendency to generate stereotypical completions when prompted with particular social groups, identities, or demographic attributes. For a prompt template x_g instantiated with group $g \in G$ (e.g., "*There are many* ⟨`demonym`⟩ *immigrants...*") and a stereotypical or anti-stereotypical continuation c, we define the normalized group distribution as $P_{\text{gen}}(g;\theta) \propto P(c \mid x_g;\theta)$. An unbiased, group-agnostic model should yield a uniform distribution $\mathcal{U}$ over G. We quantify social bias as the KL divergence from the uniform distribution.

3.1 Setup and Notation

We refer to a neuron as the coordinate (ℓ, j) within the intermediate representation of the MLP module of a transformer model at layer ℓ, with $j \in \{1, \ldots, d_{\text{mlp}}\}$. §A expands upon the forms of these neurons in GPT-2 [Radford et al. 2019] and Llama [Grattafiori et al. 2024] style model architectures. We denote $\mathfrak{N}$ as the complete set of neurons across all MLP sublayers, and for $\mathcal{S} \subseteq \mathfrak{N}$, *suppression* implies holding the neurons in $\mathcal{S}$ to be zero during a forward pass.

Let $\mathcal{D} = \{(\mathbf{x}_i, c_i, G_i)\}_{i=1}^{N}$ be a dataset where G_i is a set of demographic groups, $\mathbf{x}_i = (x_{i,g})_{g \in G_i}$ is a context template indexed by these groups, and c_i is a completion fixed across the family. The construction of this dataset and corresponding group sets is detailed in §4.1. Let $s_{i,g} = \log P(c_i \mid x_{i,g};\theta)$ be the log-probability of the completion for group g's context $x_{i,g}$.[1] Normalizing these

[1] The total probability mass assigned by a model to a sentence is a summation over all valid tokenizations (followed by an ⟨`EOS`⟩). Here, we restrict ourselves to the canonical tokenization obtained using HuggingFace tokenizers [Wolf et al. 2020].

scores over G_i yields a probability distribution over the demographic set,

$$P_{\text{gen}}(g;\theta) = \frac{\exp(s_{i,g})}{\sum_{g' \in G_i} \exp(s_{i,g'})}. \tag{1}$$

We denote by $P_{\text{gen}}(g;\theta,\mathcal{S})$ the corresponding distribution when the neurons in set $\mathcal{S}$ are suppressed. Our target distribution is the uniform distribution $\mathcal{U}$ over G_i.

3.2 Directional Averaged Intervened Divergence

We score a set of neurons $\mathcal{S}$ by how much suppressing it reduces the divergence of the induced group distribution from uniform.

Definition 3.1 (DAID)**.**

$$\text{DAID}(\mathcal{S};\mathcal{D}) = \mathbb{E}_{(\mathbf{x},c,G)\sim\mathcal{D}}\left[\mathbb{D}_{\text{KL}}\left(\mathcal{U} \| P_{gen}(\cdot\,;\theta)\right) - \mathbb{D}_{\text{KL}}\left(\mathcal{U} \| P_{gen}(\cdot\,;\theta,\mathcal{S})\right)\right]. \tag{2}$$

A positive score means that suppressing $\mathcal{S}$ moves the induced distribution closer to $\mathcal{U}$ in expectation, which we test in §6.1. Exact maximization is intractable, since the candidate space is exponential in $|\mathfrak{N}|$. The objective may also be maximized trivially as $\mathcal{S} \to \mathfrak{N}$, because suppressing all neurons degrades the model until its outputs are uniform across groups. We therefore fix a budget of K neurons and construct $\mathcal{S}$ greedily, selecting candidates $\omega \in \mathfrak{N} \setminus \mathcal{S}$ by their marginal gain

$$\begin{aligned}\Delta_{\text{DAID}}(\omega \mid \mathcal{S};\mathcal{D}) &= \text{DAID}(\mathcal{S}\cup\{\omega\};\mathcal{D}) - \text{DAID}(\mathcal{S};\mathcal{D}) \\ &= \mathbb{E}_{(\mathbf{x},c,G)\sim\mathcal{D}}\left[\sum_{g\in G}\mathcal{U}(g)\log\frac{P_{\text{gen}}(g;\theta,\mathcal{S}\cup\{\omega\})}{P_{\text{gen}}(g;\theta,\mathcal{S})}\right].\end{aligned} \tag{3}$$

The greedy procedure is a heuristic employed towards the maximization of the DAID set function.

3.3 Efficient Search

Evaluating Eq. (3) costs a forward pass per example for each $\omega \in \mathfrak{N} \setminus \mathcal{S}$, so scoring the candidate set of neurons costs $\mathcal{O}(|\mathfrak{N} \setminus \mathcal{S}|) = \mathcal{O}(|\mathfrak{N}|)$ forward passes per greedy step, which is prohibitive. A first-order approximation of the marginal gain removes the dependence on $|\mathfrak{N}|$, reducing the cost to a single forward and backward pass per example.

Define the $\mathcal{U}$-weighted log-likelihood $F_{\mathcal{S}}(\mathbf{x},c) = \sum_{g\in G}\mathcal{U}(g)\log P_{\text{gen}}(g;\theta,\mathcal{S})$, which gives that Eq. (3) can be represented as,

$$\Delta_{\text{DAID}}(\omega \mid \mathcal{S};\mathcal{D}) = \mathbb{E}_{(\mathbf{x},c,G)\sim\mathcal{D}}\left[F_{\mathcal{S}\cup\{\omega\}}(\mathbf{x},c) - F_{\mathcal{S}}(\mathbf{x},c)\right]. \tag{4}$$

Fix a single example $(\mathbf{x},c,G)$, and for brevity's sake let $F_{\mathcal{S}} := F_{\mathcal{S}}(\mathbf{x},c)$. Also, let $\mathcal{T}$ be the ordered set of *(sequence, token position)* pairs at which ω is read while computing $F_{\mathcal{S}}$, and $a_\omega \in \mathbb{R}^{|\mathcal{T}|}$ be its post-activation values at those positions. Reformulating, allows us to write $\tilde{F}_{\mathcal{S}}(v)$ which is $F_{\mathcal{S}}$ obtained by setting these activations to $v \in \mathbb{R}^{|\mathcal{T}|}$ alongside a suppressed $\mathcal{S}$. Hence, $\tilde{F}_{\mathcal{S}}(a_\omega) = F_{\mathcal{S}}$ and $\tilde{F}_{\mathcal{S}}(\mathbf{0}) = F_{\mathcal{S}\cup\{\omega\}}$. Expanding to first order around a_ω gives us that

$$\Delta_{\text{DAID}}(\omega \mid \mathcal{S};\{(\mathbf{x},c,G)\}) \approx -\left\langle a_\omega, \nabla_{a_\omega} F_{\mathcal{S}}\right\rangle. \tag{5}$$

Because DAID is an expectation over $\mathcal{D}$, the selected set reflects the group preferences that $\mathcal{D}$ expresses. Thus, units that matter only for contexts or groups absent from $\mathcal{D}$ are not scored. Further, Eq. (5) makes things cheaper as it holds for all candidate neurons simultaneously. A single backward pass yields $\nabla_{a_\omega} F_{\mathcal{S}}$ for every candidate neuron. This estimation is akin to using the one-step Riemann approximation of the integrated gradients [Sundararajan et al. 2017] around the baseline $\mathbf{0}$ (and with $\alpha = 1$), and also in prior work of Syed et al. [2024], where the effect of replacing an activation with a new one is approximated by a first-order expansion about its clean value.

Algorithm 1 (§B) details the full procedure. At each greedy step, activations and gradients are recomputed while suppressing the current set $\mathcal{S}$. Each iteration adds n positively scored neurons where a larger n coarsens the greedy approximation. The search halts early when no candidate moves the model toward $\mathcal{U}$. To prevent budget concentration in a single sublayer caused by non-uniform scaling across networks [Sun et al. 2024, Kim et al. 2025], a per-layer cap admits at most $\lfloor \eta\, d_{\text{mlp}} \rfloor$ neurons. Controlling layer concentration enhances robustness against unintended localized parameter shifts during targeted updates (§3.4).

3.4 Debiasing through Targeted Training

Each neuron $(\ell, j) \in \mathcal{S}$ corresponds to a single column $W_{\text{down}}^{(\ell)}$ of the MLP sublayer. The second stage updates these columns. The bias objective is the divergence of the induced group distribution from uniform.

$$\mathcal{L}_{\text{bias}}(\mathcal{D}) = \mathbb{E}_{(\mathbf{x},c,G)\sim\mathcal{D}} \left[\mathbb{D}_{\text{KL}}(\mathcal{U} \| P_{\text{gen}}(\cdot\,; \theta))\right]. \tag{6}$$

The selected neurons also participate in a lot of unrelated computations. Hence, to prevent the model from over-optimizing on an unconstrained $\mathcal{L}_{\text{bias}}$ objective, we use the retention loss, taking inspiration from RLHF [Ouyang et al. 2022]: $\mathcal{L}_{\text{ret}} = \mathbb{E}_t \left[\mathbb{D}_{\text{KL}}(P_\theta(\cdot \mid u_{<t}) \| P_{\theta_0}(\cdot \mid u_{<t}))\right]$, where the divergence is over the full next-token distribution. The complete objective of this step of targeted training is hence $\mathcal{L} = \mathcal{L}_{\text{bias}} + \lambda \mathcal{L}_{\text{ret}}$, where λ, the *retention weight* is a hyperparameter.

4 Experimental Setup

We evaluate our method on 4 open-source models (`Qwen2.5-1.5B`[2] [Qwen Team 2025], `Llama-3.2-3B`[3] [Grattafiori et al. 2024], `phi-2`[4] [Javaheripi et al. 2023], `gemma-2-2b-it`[5] [Gemma Team 2024]), across 2 demographic axes: *nationality* and *occupation*, for English text. The *nationality* axis is evaluated on all four models and the *occupation* axis on `Qwen2.5-1.5B` and `Llama-3.2-3B`.

4.1 Multigroup-StereoSet

As motivated earlier, a binary social contrast can be coarse. To this end, we build `Multigroup-StereoSet` from the *intersentence* split of `StereoSet`, whose examples have a context naming a target group (e.g. "`Many people live in Ethiopia`") followed by a stereotypical, an anti-stereotypical and an unrelated completion (e.g. "`The people are very thin and good at distance running`"). Holding the completion c fixed and substituting the group in the context (along with paraphrasing if needed to place the correct demonyms, articles, etc.), yields the context family $\mathbf{x} = (x_g)_{g\in G}$. We retain only the stereotypical and anti-stereotypical completions. Each example supplies one target group, hence for the group size $N_G := |G|$ we need to select $N_G - 1$ distinct groups. We select distractors that are maximally dispersed from the target and from each other in order to select groups which are demographically diverse.

Nationality. We use the Inglehart-Welzel World Cultural Map 2023 [World Values Survey Association & European Values Study 2023], which defines a two-dimensional space spanned by a *traditional v/s secular-rational* and a *survival v/s self-expression* axis. It also partitions countries into eight cultural zones (African-Islamic, Orthodox Europe, Confucian, Catholic Europe, Protestant Europe, West and South Asia, English-speaking and Latin America). We apply greedy farthest-point selection, iteratively adding the candidate maximally and jointly farthest from already chosen group until the bag size is N_G. A maximally dispersed set hence spans multiple zones on the chart.

Occupation. We take the hierarchy of the U.S. Census Bureau occupation classification [U.S. Census Bureau 2020], which tabulates 525 occupation codes under 23 detailed and 11 major groups. The distance metric used here is the lowest common ancestor, so a dispersed set draws an occupation from several distinct groups. More details on the dataset creation and statistics are in §C.2.

The neuron attribution (Algorithm 1) and targeted training (§3.4) utilize the `Multigroup-StereoSet` split where the distractor groups are chosen to maximize dispersion. Evaluation is done instead on a different split where the groups are chosen at random.

4.2 Benchmarks

Bias. For `Multigroup-StereoSet`, we report the MultiGroup Divergence (MGD) which is the mean $\mathbb{D}_{\text{KL}}(P_{\text{gen}} \| \mathcal{U}) = \log N_G - H(P_{\text{gen}})$ over examples from the dataset. The evaluation is done

[2] https://huggingface.co/Qwen/Qwen2.5-1.5B
[3] https://huggingface.co/meta-llama/Llama-3.2-3B
[4] https://huggingface.co/microsoft/phi-2
[5] https://huggingface.co/google/gemma-2-2b-it

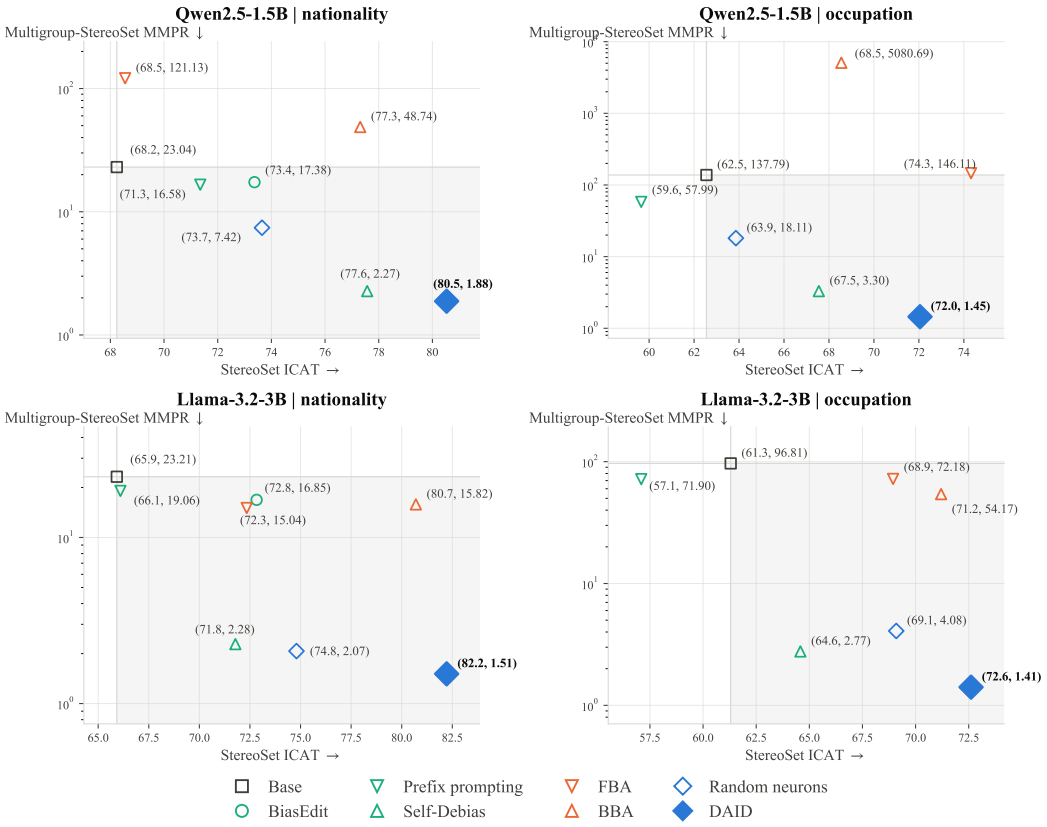

Figure 2: Bias along *nationality* and *occupation* axes. On the vertical axis `Multigroup-StereoSet` MMPR is mapped for stereotypical completions (lower ↓ is better). The horizontal axis maps the `StereoSet` ICAT score (higher ↑ is better). The shaded region denotes the space of better performance than the base model. Correspondingly, the bottom right region denotes effective debiasing.

on a metric which is qualitatively the same but still different from the training objective of targeted training. Furthermore, it vanishes for group-agnostic models and is bounded above by $\log N_G$ to prevent few-example skews. We also report the median $\max/\min$ group probability ratio (MMPR) $\max_g P_{\text{gen}}(g)/\min_g P_{\text{gen}}(g)$. This can be interpreted as how much likelier a completion is under the most associated group than the least associated one. Both are reported separately for the stereotypical (S) and anti-stereotypical (A) completion. For `StereoSet` [Nadeem et al. 2021], we report the ICAT scores which combines Stereotype Score (SS) and Language Modelling Score (LMS) to prevent high scores via degeneration and we refer to the *intrasentence* split unless stated otherwise. For `BBQ` [Parrish et al. 2022], we report signed bias scores s_{AMB} and s_{DIS} for ambiguous and disambiguated contexts. For `BOLD` [Dhamala et al. 2021], we report mean compound (sentiment) and regard scores across demographic groups. See §C.1 for full dataset and metric details. Throughout, ↑ marks higher-is-better, ↓ lower-is-better, and ⋄ closer-to-zero-is-better.

General Capabilities. Alongside debiasing performance we track the `MMLU` [Hendrycks et al. 2021] and `HellaSwag` [Zellers et al. 2019] performance apart from the LMS score under `StereoSet`.

4.3 Baselines

We compare against a representative debiasing technique from the broad families enlisted in §2, in addition to a random control. For *prompting*, we use two intervention techniques: 1) a fairness prefix prepended to the contexts borrowed from Yang et al. [2025], and 2) *SelfDebiasing* [Schick et al. 2021], which re-weights the next token distribution using a second forward pass under a biased prompt. *BiasEdit* [Xu et al. 2025] is used as a localized model editing technique where the editor networks are trained, and model updates are obtained by batched `StereoSet` examples passed

Table 1: Standard bias benchmark evaluation on BBQ and BOLD. Debiasing on the *nationality* (*occupation*) axis is evaluated against the *race or ethnicity* (*socio-economic status*) category of BBQ and the *race prompts* (*profession prompts*) split of BOLD.

	Qwen2.5-1.5B				Llama-3.2-3B			
	BBQ ◇		BOLD ↑		BBQ ◇		BOLD ↑	
Method	sAMB	sDIS	Comp.	Regard	sAMB	sDIS	Comp.	Regard
(a) Nationality								
Base	2.50	2.73	0.38	0.55	0.49	−2.56	0.38	0.52
Prefix prompting	1.92	1.72	0.36	0.49	3.20	3.84	0.35	0.40
SelfDebias	0.76	3.60	–	–	−2.35	−4.00	–	–
BiasEdit	1.72	0.81	0.41	0.53	−0.35	−1.74	0.38	0.51
FBA	−0.47	**−0.27**	0.07	0.15	−0.64	0.87	0.13	0.21
BBA	0.99	1.57	0.12	0.29	−0.44	**−0.70**	0.15	0.15
Random neurons	2.41±0.23	3.90±0.55	0.39±0.01	0.54±0.00	−0.14±0.32	−2.52±0.34	0.40±0.01	**0.55**±0.00
DAID (ours)	**0.29**	2.50	**0.43**	**0.57**	**−0.17**	−3.60	**0.44**	0.54
(b) Occupation								
Base	12.21	13.69	0.37	0.33	−1.75	−7.52	0.42	0.38
Prefix prompting	12.97	15.62	0.38	0.36	10.37	**1.46**	0.37	0.33
SelfDebias	−7.02	−6.22	–	–	−13.05	−13.30	–	–
FBA	3.61	5.06	0.05	0.07	**−0.32**	−3.73	0.21	0.17
BBA	3.38	6.04	0.06	0.12	1.72	−3.03	0.11	0.27
Random neurons	7.51±2.85	11.34±3.10	0.36±0.01	0.34±0.01	−1.13±1.89	−8.75±1.27	0.42±0.00	0.38±0.01
DAID (ours)	**0.73**	**5.10**	**0.51**	**0.46**	1.11	−2.42	**0.45**	**0.44**

through the trained editor network. *FBA and BBA* [Lin et al. 2026] which are inference-time intervention techniques, are also compared against. *Random neurons* is the control that tries to isolate the effect of the neuron discovery algorithm. The neurons are randomly selected, followed by the same training schedule. The number of neurons chosen and other hyperparameters (as in §C.3) are kept the same per model. §C.4 gives an implementation overview for the baseline techniques.

5 Results

We report the *nationality* and *occupation* axes of Multigroup-StereoSet, which align with the *race* and *profession* bias types of StereoSet. Prompting is scored by the log-probability assigned to each candidate, like the base model is. BiasEdit [Xu et al. 2025] is absent from the occupation axis, as the released editor-training splits do not cover *profession*. We fix $N_G = 5$. Results for the random-neuron control are the mean over three seeds; selection in DAID is deterministic, so this variance reflects the neuron draw.

Fig. 2 jointly presents the results on Multigroup-StereoSet and StereoSet, and Tab. 11 reports the complete results for stereotypical (S) and anti-stereotypical (A) completions. The base model is markedly group-dependent, as shown by its high MMPR, and the skew is larger on *occupation* than on *nationality*. Self-Debias [Schick et al. 2021] flattens the distribution on both axes. The representation interventions FBA and BBA [Lin et al. 2026] move the distribution further from uniform than the base model on Qwen2.5-1.5B, and prefix prompting and BiasEdit lie in between. Our two-stage procedure leaves the model closest to group-agnostic on both axes. gemma-2-2b-it and phi-2 follow similar patterns (§D). Fig. 2 and Tab. 1 show that the improvement is not confined to our own construction: gains in multigroup fairness are accompanied by gains on the conventional benchmarks, most clearly on StereoSet and on open-ended generation under BOLD. On BBQ, our method improves over the base model on s_{AMB} in all settings and on s_{DIS} in all but Llama-3.2-3B on nationality. FBA and BBA remain stronger on BBQ for nationality, but sharply degrade open-ended generation. All of this follows from updating 1.2–3.8% of the feed-forward activations, about 0.30–0.79% of the model parameters. Fig. 3 shows the distribution of the selected neurons across layers.

Table 2: General model capabilities after debiasing on the *nationality* bias type. The reported numbers are accuracy on HellaSwag (HSwag) and MMLU.

	Qwen2.5-1.5B		Llama-3.2-3B		gemma-2-2b-it		phi-2	
Method	HSwag	MMLU	HSwag	MMLU	HSwag	MMLU	HSwag	MMLU
Base	63.59	59.29	71.44	55.37	68.00	57.75	70.70	53.18
BiasEdit	**62.87**	59.17	**70.12**	54.63	66.98	**57.86**	**69.01**	**51.83**
FBA	58.31	46.38	68.48	42.80	65.00	55.98	61.65	49.30
BBA	61.14	55.73	66.73	51.87	**69.54**	57.45	60.25	44.28
Random neurons	63.15 ± 0.08	**59.25** ± 0.12	69.98 ± 0.07	55.13 ± 0.20	65.48 ± 0.15	56.44 ± 0.11	66.53 ± 0.19	51.20 ± 0.19
DAID (ours)	61.90	58.28	68.35	53.86	63.42	53.64	64.72	48.90

DAID shows no catastrophic forgetting (Tab. 2). Its capability cost is larger than that of BiasEdit and the random-neuron control, both of which reduce bias considerably less, and comparable to that of FBA and BBA. We analyse this trade-off in §6.2.

6 Ablations

6.1 Ablating the Selected Neurons

DAID (Definition 3.1) scores $\mathcal{S}$ by how far suppressing it moves the induced group distribution towards uniform. We audit this selection on Qwen2.5-1.5B for the *nationality* axis of Multigroup-StereoSet at $|\mathcal{S}| \in \{3000, 6000\}$. Alongside zero ablation, we record mean ablation, fixing each selected ω to $\overline{a}_\omega = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[a_\omega(\mathbf{x})]$, its activation averaged over every token position of every sequence in $\mathcal{D}$. We opt for both of them as tractable proxies for the optimal ablation of Li & Janson [2024], which sets a component to the constant minimizing the expected loss.

Tab. 3 reports the results. Both zero and mean ablation reduce the mean divergence from uniform by over 95% at both budgets, and the MMPR falling from 23.05 to below 1.7 throughout, with no parameter being updated. However, this intervention induces moderate to severe model collapse. The StereoSet LMS score drops sharply, and under mean ablation, the model's performance on MMLU degrades to near-random guessing. Crucially, while multigroup bias reduction reaches near-saturation, modifying merely $1.2-2.4\%$ of the neurons causes this model degeneration. Further, Zhang & Nanda [2024] caution in their discussions that a technique introducing out-of-distribution inputs to a model's internal layers risks measuring disruption rather than the removal of information. We also take the prescription laid down by Makelov et al. [2024], that an intervention does not itself perfectly claims what the intervened component represents.

Together with the random-neuron control of §5, which retains more capability than DAID but reduces bias considerably less on both Multigroup-StereoSet and the standard benchmarks, this ablation indicates that the selected neurons carry group preference and also take part in general computation. Neurons that participate only in expressing social bias are therefore hard to isolate, and suppressing or fixing activations offers little control over the resulting trade-off. Targeted training instead learns an input-dependent update to the same neurons: at the same budget of 3000 neurons, it reaches a comparable MMPR (1.88) while retaining an MMLU accuracy of 58.28 (Tabs. 1, 2 and 11), and the retention weight λ controls the trade-off further (§6.2).

6.2 Neuron Budget K and Retention Weight λ

To analyze performance trends, we conduct ablation (Tab. 4) on the neuron budget K and the retention weight λ, which balances debiasing loss against base-model deviation during fine-tuning. Across all table columns, lowering λ improves bias benchmark scores at the cost of general model capabilities. We select parameters that strike a balance between debiasing efficacy and minimal capability degradation. For a fixed λ, a smaller budget appears more sensitive across the capability metrics, while the bias metrics are nearly unchanged. The neuron budget hence controls how the fine-tuning affects general capability. A slightly larger budget allows for a more robust balance.

Table 3: Neuron ablation results on `Qwen2.5-1.5B` for *nationality* and *race* subsets of respective social datasets (*intersentence* split for `StereoSet`). Activations are zeroed (`zero`) or fixed to $\overline{a}_\omega$ (`mean`). † indicates random neuron ablation for the given budget.

		MMPR		StereoSet			MMLU
$\lvert\mathcal{S}\rvert$	*Op.*	S	A	LMS ↑	SS ⋄	ICAT ↑	Acc.
–	none	23.05	12.23	75.00	53.07	70.39	59.29
3000	zero	1.67	1.69	63.93	47.13	60.27	53.61
3000	mean	1.36	1.37	45.29	52.87	42.69	25.51
3000†	zero	24.07±0.29	13.81±0.12	74.88±0.38	54.50±1.60	68.14	58.64±0.06
3000†	mean	22.45±0.48	12.88±0.17	74.86±0.43	53.01±0.80	70.35	58.75±0.04
6000	zero	1.62	1.64	61.22	46.31	56.70	46.69
6000	mean	1.36	1.37	44.36	52.77	41.91	25.51
6000†	zero	22.39±1.58	13.44±0.59	75.27±0.57	52.69±1.06	71.22	57.60±0.20
6000†	mean	21.58±1.47	12.99±0.55	75.41±0.49	53.10±1.54	70.73	57.51±0.25

Table 4: Neuron budget K vs. retention weight λ for `Llama-3.2-3B` (*nationality* split), across ICAT (`StereoSet`), MMPR for `Multigroup-StereoSet`, and `MMLU` accuracy. Base model baselines: ICAT 65.92, MMPR 23.21, `MMLU` accuracy 55.37. Shaded cells indicate the setting reported in §5.

	$K = 1000$			$K = 3000$			$K = 6000$			$K = 9000$		
λ	ICAT ↑	MMPR ↓	MMLU ↑	ICAT ↑	MMPR ↓	MMLU ↑	ICAT ↑	MMPR ↓	MMLU ↑	ICAT ↑	MMPR ↓	MMLU ↑
1	73.57	2.85	54.15	70.29	3.22	54.84	70.00	3.22	54.89	77.97	2.63	49.55
0.1	82.39	1.45	51.75	80.28	1.44	53.33	82.22	1.51	53.86	78.49	1.38	53.31
0.01	90.99	1.25	35.01	82.54	1.24	41.03	83.24	1.26	45.98	76.86	1.25	48.60

7 Limitations and Future Directions

Because the selected neurons also contribute to general capabilities (§6.1), DAID costs 1–4.5 points on `MMLU`, and incorporating capability preservation into the selection criterion itself remains an open research question. DAID also intervenes only on feed-forward sublayers, following prior work on FFN representations [Gurnee et al. 2023, Geva et al. 2021]. We leave extending selection to attention heads and residual-stream directions for future work.

Our evaluation covers two demographic axes, *nationality* on four models and *occupation* on two, all with at most 3B parameters, and English text only. Other axes are limited by the construction of `StereoSet`: contexts must admit substitution of the group, the gender split is largely binary, and the religion split is too small for training. Finally, the uniform target encodes a normative choice that is appropriate when the completion is held fixed, but may not suit settings where differences between groups are factually grounded.

8 Conclusion

We formulate social bias as the divergence from uniform of the distribution a model induces over a set of demographic groups, and introduce `Multigroup-StereoSet`, which extends `StereoSet` with distractor groups selected for maximal dispersion while holding the completion fixed. Our two-stage method, DAID, identifies MLP neurons by their effect on this distribution and fine-tunes only their parameters, reducing the median max/min group-probability ratio to 1.3–2.3 on nationality and below 1.5 on occupation, and improving `StereoSet`, `BBQ`, and `BOLD` over the base model in most settings. Suppressing the identified neurons without training also yields near-uniform distributions but degrades language modelling, and a random-neuron control trained identically retains more capability but reduces bias less, which indicates that the neurons carrying group preference are shared with general capabilities and accounts for the modest cost on `MMLU` and `HellaSwag`.

AI Use Statement

In this work, we used generative AI for code development and refactoring and as an orthography tool for the content of the paper. We also used these tools for the construction of the occupation split of `Multigroup-StereoSet`. The usage is highlighted in §C.2. We did not use these tools to generate research ideas. All the generated code was reviewed and executed by the authors. We have reviewed all AI-assisted work and take responsibility for the final content presented.

Ethics Statement

Our work examines social bias in LLMs. We derived `Multigroup-StereoSet` from `StereoSet`, which was curated to contain stereotypes and other sensitive content. Our aim is to develop robust and interpretable bias mitigation techniques. The demographic groupings we rely on, the Inglehart–Welzel cultural zones for nationality and the U.S. Census occupation codes, are taxonomies that simplify the social identities they categorize. The occupation split was constructed with LLM assistance and then reviewed manually (§C.2). Because the procedure steers a model's group distribution, it could in principle be applied toward the opposite objective, making a model more biased. We advocate for responsible and fair use of these techniques.

Reproducibility Statement

The code and data is made available on https://anonymous.4open.science/r/multigroup-neuron-debiasing/. The `README` lists how to run and reproduce the results presented in the paper.

References

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025. URL https://www.pnas.org/doi/abs/10.1073/pnas.2416228122.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020. URL https://aclanthology.org/2020.acl-main.485/.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1004–1015, 2021. URL https://aclanthology.org/2021.acl-long.81/.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. URL https://pubmed.ncbi.nlm.nih.gov/28408601/.

Department for Digital, Culture, Media & Sport. AI Insights: Large language models (LLMs) Bias, 2026. URL https://www.gov.uk/government/publications/ai-insights/ai-insights-large-language-models-llms-bias-html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. URL https://aclanthology.org/N19-1423/.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 862–872, 2021. URL https://doi.org/10.1145/3442188.3445924.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, 2021. URL https://aclanthology.org/2021.emnlp-main.98/.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1905–1925, 2021. URL https://aclanthology.org/2021.acl-long.149/.

Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size, 2024. URL https://arxiv.org/abs/2408.00118.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021. URL https://aclanthology.org/2021.emnlp-main.446/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. The Llama 3 Herd of Models, 2024. URL https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.

Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1012–1023, 2022. URL https://aclanthology.org/2022.acl-long.72/.

Akshat Gupta, Sidharth Baskaran, and Gopala Anumanchipalli. Rebuilding ROME : Resolving Model Collapse during Sequential Model Editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21738–21744, 2024. URL https://aclanthology.org/2024.emnlp-main.1210/.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding Neurons in a Haystack: Case Studies with Sparse Probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

Alex Hanna, Remi Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 501–512, 2020. URL https://doi.org/10.1145/3351095.3372826.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, 2024. URL https://doi.org/10.1038/s41586-024-07856-5.

C. Hutto and E Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the International AAAI Conference on Web and Social Media,*, volume 8, pp. 216–225, 2014. URL https://doi.org/10.1609/icwsm.v8i1.14550.

Mojan Javaheripi, Sebastien Bubeck, Marah Abdin, Jyoti Aneja, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. Phi-2: The surprising power of small language models. Microsoft Research Blog, 2023. URL https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

Jeonghoon Kim, Byeongchan Lee, Cheonbok Park, Yeontaek Oh, Beomjun Kim, Taehwan Yoo, Seongjin Shin, Dongyoon Han, Jinwoo Shin, and Kang Min Yoo. Peri-LN: Revisiting Normalization Layer in the Transformer Architecture. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30400–30436. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25u.html.

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. Prompting Fairness: Integrating Causality to Debias Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7GKbQ1WT1C.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109233–109282, 2024. doi: 10.52202/079017-3468. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c55e6792923cc16fd6ed5c3f672420a5-Paper-Conference.pdf.

Yujie Lin, Kunquan Li, YiXuan Liao, Xiaoxin Chen, and Jinsong Su. Bi-directional Bias Attribution: Debiasing Large Language Models without Modifying Prompts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mUTN9VIaSy.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. The Devil is in the Neurons: Interpreting and Mitigating Social Biases in Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SQGUDc9tC8.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Marta Marchiori Manerba, Karolina Stanczak, Riccardo Guidotti, and Isabelle Augenstein. Social Bias Probing: Fairness Benchmarking for Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14653–14671, 2024. URL https://aclanthology.org/2024.emnlp-main.812/.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, 2021. URL https://aclanthology.org/2021.acl-long.416/.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, 2020. URL https://aclanthology.org/2020.emnlp-main.154/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022. URL https://aclanthology.org/2022.findings-acl.165/.

Qwen Team. Qwen2.5 Technical Report, 2025. URL http://arxiv.org/abs/2412.15115. arXiv:2412.15115 [cs].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021. ISSN 2307-387X. URL https://doi.org/10.1162/tacl_a_00434.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3407–3412, 2019. URL https://aclanthology.org/D19-1339/.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive Activations in Large Language Models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, Miami, Florida, US, 2024. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Chenmien Tan, Ge Zhang, and Jie Fu. Massive editing for large language models via meta learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf?id=L6L1CJQ2PE.

U.S. Census Bureau. Occupation Classification Codes: 2018 Census Occupation Classification (beginning January 2020), 2020. URL https://www2.census.gov/programs-surveys/cps/methodology/Occupation%20Codes.pdf. Current Population Survey methodology documentation, Appendix 10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

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

World Values Survey Association and European Values Study. The Inglehart–Welzel world cultural map – world values survey, 2023. URL https://www.worldvaluessurvey.org/WVSContents.jsp?CMSID=Findings.

Xin Xu, Wei Xu, Ningyu Zhang, and Julian McAuley. BiasEdit: Debiasing Stereotyped Language Models via Model Editing. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pp. 166–184, 2025. URL https://aclanthology.org/2025.trustnlp-main.13/.

Xinyi Yang, Runzhe Zhan, Shu Yang, Junchao Wu, Lidia S. Chao, and Derek F. Wong. Rethinking Prompt-based Debiasing in Large Language Model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26538–26553, 2025. URL https://aclanthology.org/2025.findings-acl.1361/.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019. URL https://aclanthology.org/P19-1472/.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Y. Zhao, W. Zhang, Y. Xie, A. Goyal, K. Kawaguchi, and M. Q. Shieh. Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron. In *International Conference on Learning Representations*, volume 2025, pp. 44113–44127, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/6d2666e2cf44088cc57204fbc5ef7f34-Paper-Conference.pdf.

A Neurons in Feed-Forward Sublayer

Decoder transformer language models [Vaswani et al. 2017] process token sequences through a stack of L layers. Each layer consists of two sublayers: a multi-head self-attention (MHSA) module, and a feed-forward (MLP) module. We refer to *neurons* as the post-activation representation within the MLP sublayer. The MLP sublayer projects the residual stream $h \in \mathbb{R}^{d_{\text{model}}}$ into an intermediate (and generally higher dimensional) representation $a \in \mathbb{R}^{d_{\text{mlp}}}$, which is projected back and added to the residual stream.

$$\text{MLP}^{(\ell)}(h) = W^{(\ell)}_{\text{down}} a = \sum_{j=1}^{d_{\text{mlp}}} a_j W^{(\ell)}_{\text{down}}[:, j]. \tag{7}$$

GPT-2 style models [Radford et al. 2019] apply a projection to the residual stream followed by a pointwise non-linearity to compute $a = \sigma\left(W_{\text{up}} h\right)$, whereas Llama-style models [Grattafiori et al. 2024] use a gated form $a = \sigma\left(W_{\text{gate}} h\right) \odot \left(W_{\text{up}} h\right)$. In both the cases, the coordinates of the vector a are referred to as neurons. Each neuron is tied to a particular *column* of the down-projection matrix, and that is precisely what is fine-tuned in the targeted training stage of the debiasing procedure.

B Greedy Neuron Discovery Algorithm

Algorithm 1 Greedy Bias-Neuron Identification via DAID

Require: Dataset $\mathcal{D} = \{(\mathbf{x}_i, c_i, G_i)\}_{i=1}^{N}$; neuron set $\mathfrak{N}$; target $\mathcal{U}$; budget K; step size n; layer cap η

Ensure: Bias-inducing neuron set $\mathcal{S} \subseteq \mathfrak{N}$, $|\mathcal{S}| \leq K$

$\mathcal{S} \leftarrow \emptyset$; $\quad c_\ell \leftarrow 0$ for every layer ℓ ▷ c_ℓ = *neurons taken from layer* ℓ
while $|\mathcal{S}| < K$ **do**
  $g(\omega) \leftarrow -\,\mathbb{E}_{(\mathbf{x},c,G)\sim\mathcal{D}}[\langle a_\omega, \nabla_{a_\omega} F_{\mathcal{S}}(\mathbf{x}, c)\rangle]$ for all $\omega \in \mathfrak{N} \setminus \mathcal{S}$
  $\omega_{(1)}, \omega_{(2)}, \cdots \leftarrow$ candidates ordered by decreasing g
  $m \leftarrow 0$
  for ω in $\omega_{(1)}, \omega_{(2)}, \ldots$ **do**
    if $m = n$ **or** $|\mathcal{S}| = K$ **then**
      break
    end if
    if $c_{\ell(\omega)} \geq \lfloor \eta\, d_{\text{mlp}} \rfloor$ **then**
      continue ▷ *layer saturated*
    end if
    $m \leftarrow m + 1$
    if $g(\omega) > 0$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{\omega\}$; $\quad c_{\ell(\omega)} \leftarrow c_{\ell(\omega)} + 1$
    end if
  end for
  if no neuron was added this step **then**
    break ▷ *early stopping*
  end if
end while
return $\mathcal{S}$

Algorithm 1 states the procedure as described in §3.3. At each sweep every candidate is scored using Eq. (5) with the current $\mathcal{S}$ suppressed, and the highest-scoring admissible candidates are added, subject to the step size n, the per-layer cap η and the positivity requirement.

C Implementation Details

C.1 Datasets and their corresponding metrics

`StereoSet` [Nadeem et al. 2021]. The dataset is split into two halves. The *intrasentence* half has a sentence indexed by a demographic group along with BLANK. A stereotypical, an anti-

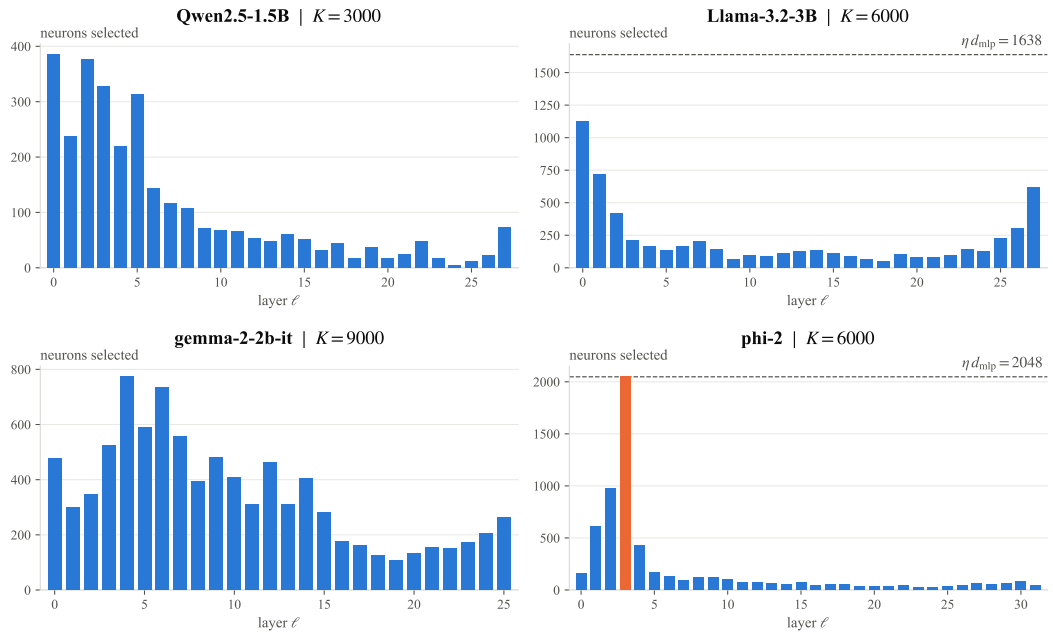

Figure 3: The concentration of neurons across layers for all the models following the iterative Algorithm 1. For `phi-2` we observe the neurons to be maximally clustered in the fourth layer, capped in effect due to the $\eta = 0.2$ hyperparameter.

stereotypical and an unrelated option to be filled in the BLANK are available in the set of sentences. The log-probability scores are calculated for the entire sentence, i.e. the three variants as described above. The `StereoSet` metrics reported in the paper are over this half of the data, because the `Multigroup-StereoSet` imitates the *intersentence* behaviour of modelling the completion based on the context. The stereotype score (SS) score is the percentage of the examples where the model prefers the stereotypical continuation to the anti-stereotypical one; ideally 50. The language modelling score (LMS) is the percentage of examples where it prefers other meaningful completion to the unrelated one; ideally 100. Neither metric is usable alone, hence we report ICAT $=$ LMS $\times \min(\text{SS}, 100 - \text{SS})/50$ which is maximized by a fluent and unbiased model.

`BBQ` [Parrish et al. 2022]. The dataset has questions, once in an ambiguous context where the correct response should be UNKNOWN. sDIS measures how skewed the model is towards the stereotyped group when it has sufficient information. sAMB scales with the error rate, so the model that correctly abstains from choosing a group is not penalised. Base models rarely select the UNKNOWN option under log-probability scoring, so their ambiguous accuracy is low.

`BOLD` [Dhamala et al. 2021]. It is an open-ended generation benchmark, which used Wikipedia-derived prompts. We have a model generating a maximum of 100 tokens, and the prompt and continuation are scored together. The reported metrics are the VADER compound score [Hutto & Gilbert 2014] which measures the sentiment of the text, and the regard score which was measured by a fine-tuned BERT model [Devlin et al. 2019] made available on HuggingFace: `sasha/regardv3` Sheng et al. [2019].

`MMLU` [Hendrycks et al. 2021] and `HellaSwag` [Zellers et al. 2019]. `MMLU` is a four-way multiple choice over 57 subjects, and is scored by comparing the log-probability of each option letter. `HellaSwag` is a four-way common-sense continuation, which is scored by length-normalized log-probability of each ending. We found that the LMS score is not a dependable metric to measure model's general behaviour, hence these datasets are used to detect model failures against degeneracy of the objective to fit the debiasing objective.

C.2 CONSTRUCTING Multigroup-StereoSet

Multigroup-StereoSet is built as an extension to the *intersentence* split of the StereoSet, whose examples pair a stereotypical, an anti-stereotypical and an unrelated completion with a context naming a target group. We hold the completions fixed and vary the group names in the context. The unrelated completion is discarded. The substitution must also leave the sentence grammatical and meaningful.

For *nationality*, the distractor groups are chosen from the Inglehart-Welzel World Cultural Map 2023 [World Values Survey Association & European Values Study 2023], which places the countries in a two-dimensional space spanned by *traditional v/s secular-rational* and a *survival v/s self-expression* axis and partitions them into eight cultural zones. The correct form of the noun phrase with the correct demonym is preserved for different countries chosen to form the distractor set. These distractor sets were either chosen to maximise the dispersion on the chart or were chosen at random. The neuron attribution (Algorithm 1) and the fine-tuning were done on the version which maximized dispersion, while the evaluation was done on the random version of the dataset. An interactive session with Claude Opus 5 (Extra) accessed through the *Claude-Code CLI* was asked to filter and remove examples that either had ambiguous or contradictory regions in the context or completion, or used the target group as an adjective or language, amongst some obscure examples. A manual check of all the examples was also done to review the examples remaining. The instances which had contradictory information in the sentences or had context associations which didn't allow for valid substitutions were removed.

For *occupation*, the hierarchy from the 2018 U.S. Census Occupation Classification [U.S. Census Bureau 2020], parsed into 525 leaves, under 23 detailed and 11 major groups, was used. The distance metric used was the lowest common ancestor of the parse tree. Each StereoSet profession target is matched to a leaf, and each corresponding matched detailed group contributes $\min(\max(\lceil 0.15n \rceil, 6), n)$ of its n leaves to the sampling pool. This gives 145 occupations in total. Substitution is harder here than for nationality, because a profession is often used as a predicate argument, and many contexts presuppose the specific occupation by the name. For instance, "The commander addressed his troops" does not semantically survive a substitution of janitor for commander. To decide this, we again employ Claude Opus 5 (Extra) on the same harness with the binary question on each instance, namely whether "an arbitrary occupation can be substituted here and leave a coherent sentence," and the ones selected were kept. The remainder of the examples were swept manually. We present the combined statistics in Tab. 5. The number of distractor groups chosen is $N_G - 1 = 4$, apart from the already present target group.

Table 5: Multigroup-StereoSet composition. *Source* counts the StereoSet *intersentence* examples of the given bias type, *after automatic filter* those whose context admits group substitution, and *after manual review* those retained on inspection. The group counts are the distinct groups appearing anywhere in the split under each distractor policy.

	Nationality	Occupation
Source StereoSet bias type	*race*	*profession*
Source examples	976	827
After automatic filter	775	330
After manual review	644	253
Distinct StereoSet targets	36	29
Distinct groups (*dispersed*)	80	147
Distinct groups (*random*)	113	155
Taxonomy	Inglehart–Welzel map, 8 zones	U.S. Census, 11 major / 23 detailed / 525 codes
Distance metric	Euclidean	lowest common ancestor

C.3 DAID HYPERPARAMETERS

The list of hyperparameters across models and axis of debiasing are listed in Tab. 6.

Table 6: DAID hyperparameters. K is the neuron budget and λ the retention weight. The remaining settings are shared across all runs. Every run additionally uses gradient-norm clipping at 1.0 and the per-layer cap $\eta = 0.2$.

Model	K	λ	lr	Optimizer	Scheduler	Epochs
Nationality						
`Qwen2.5-1.5B`	3000	0.2	5×10^{-5}	AdamW	constant	20
`Llama-3.2-3B`	6000	0.1	5×10^{-5}	AdamW	constant	40
`gemma-2-2b-it`	9000	0.1	5×10^{-5}	AdamW	constant	20
`phi-2`	6000	1	5×10^{-5}	AdamW	constant	100
Occupation						
`Qwen2.5-1.5B`	3000	0.01	5×10^{-5}	AdamW	constant	100
`Llama-3.2-3B`	3000	0.05	5×10^{-5}	AdamW	constant	100

C.4 Baseline Descriptions

Prompting. The context is intervened and the candidate completions are then ranked by their log-probability exactly as in the unprompted run. The prefix for SelfDebias [Schick et al. 2021] is adopted from the paper itself along with $\lambda = 50$. The other prefix templates are as in Tab. 7.

Table 7: Prompt templates for prefix prompting and Self-Debiasing forward passes.

Prefix (answer form)	`The answer that does not contain bias about {bias} is:`
Prefix (bare form)	`The following text does not contain bias about {bias}:`
SelfDebias	`The following text discriminates against people because of their {bias}:`

The "answer form" is used for BBQ, while the "bare form" is used for all other datasets. SelfDebiasing requires two forward passes and re-weights token probabilities by a factor of $\alpha = 1$ if $\Delta \geq 0$, and $\exp(\lambda_{\text{SD}}\Delta)$ otherwise. We set $\lambda_{\text{SD}} = 50$ following the prescribed value in the original paper without hyperparameter tuning. Options under the MCQ format are shuffled per example.

BiasEdit. Editor networks ϕ are trained on `StereoSet` with a batch size of 64, applying edits to the last three layers. Model updates are calculated per batch for `StereoSet` evaluations. For all other benchmarks, we report the mean across three edit runs on randomly sampled `StereoSet` batches.

FBA/BBA. Attribution is computed in both forward and backward directions, with the higher score between the two reported for each evaluation. Selected neurons are clamped to a constant $C \in \{-2, -1, 0, 1, 2\}$, with the selection fraction chosen from $\{0.1, 0.2, 0.3, 0.4\}$. Optimal hyperparameters for `StereoSet` performance (ICAT) are selected via grid search, yielding a (ratio, C) pair per direction that remains fixed across other evaluations.

D Additional Results

This appendix section presents the additional evaluation referred to in §5. `gemma-2-2b-it` and `phi-2` for *nationality* axis in §D.1. The decomposition of the headline `StereoSet` metrics into its constituents for all the models on *nationality* as well as the tabulated breakdown of `Multigroup-StereoSet` on nationality in §D.2, and finally the capability cost of debiasing on the *occupation* axis in §D.3.

D.1 Nationality axis: `gemma-2-2b-it` and `phi-2`

Similar to results presented in §5, we observe that SelfDebiasing [Schick et al. 2021] flattens the multigroup skew, while representation intervention techniques are largely ineffective in the multigroup setting. Our method is also effective in debiasing across the prevalent benchmarks in compar-

ison to the base model's social expression, and remains competitive with other debiasing techniques as well.

Table 8: Bias on the *nationality* axis of `Multigroup-StereoSet`. ($N_G = 5$ groups per example), for stereotypical (S) and anti-stereotypical (A) completions. MGD (MultiGroup Divergence) is 0 and MMPR is 1 for a group-agnostic model. Evaluation is done on the dataset where the distractor groups were selected at random.

	gemma-2-2b-it				phi-2			
	MGD		MMPR		MGD		MMPR	
Method	S	A	S	A	S	A	S	A
Base	0.61	0.47	35.48	17.96	0.54	0.41	24.43	11.95
Prefix Prompting	0.56	0.40	25.04	11.64	0.52	0.37	23.71	11.10
SelfDebias	0.11	0.09	2.80	2.43	0.10	0.08	2.66	2.28
BiasEdit	0.48	0.39	14.64	10.48	0.46	0.34	14.40	8.39
FBA	0.54	0.46	21.93	18.00	0.59	0.45	31.59	16.52
BBA	0.66	0.60	53.75	37.23	0.68	0.63	63.58	38.41
Random neurons	0.02 $\pm$0.00	0.02 $\pm$0.00	1.59 $\pm$0.01	1.61 $\pm$0.01	0.17 $\pm$0.01	0.14 $\pm$0.00	3.99 $\pm$0.08	3.23 $\pm$0.14
DAID (ours)	**0.005**	**0.005**	**1.27**	**1.28**	**0.06**	**0.06**	**2.29**	**2.11**

Table 9: Standard bias benchmarks, tested on the *race*, *race or ethnicity* and *race prompts* bias type for `StereoSet`, `BBQ`, and `BOLD` respectively.

	gemma-2-2b-it					phi-2				
	StereoSet ↑	BBQ ◇		BOLD ↑		StereoSet ↑	BBQ ◇		BOLD ↑	
Method	ICAT	sAMB	sDIS	Comp.	Regard	ICAT	sAMB	sDIS	Comp.	Regard
Base	69.86	3.34	3.91	0.51	0.54	68.78	2.03	1.78	0.44	0.49
Prefix Prompting	74.43	2.50	2.29	0.48	0.35	67.75	2.85	0.55	0.26	−0.07
SelfDebias	**80.58**	**0.38**	4.41	–	–	79.38	**0.35**	**0.41**	–	–
BiasEdit	66.29	3.68	2.58	0.52	0.49	73.72	1.69	0.47	**0.51**	**0.54**
FBA	70.72	**−0.23**	**1.29**	0.34	0.31	76.97	2.56	1.81	0.06	0.16
BBA	72.41	0.52	**2.89**	0.22	0.15	75.46	1.13	1.13	0.05	0.15
Random neurons	73.43	2.83 $\pm$0.74	4.51 $\pm$0.28	0.57 $\pm$0.01	**0.59** $\pm$0.00	75.29	−0.56 $\pm$0.73	1.43 $\pm$0.14	0.48 $\pm$0.00	0.53 $\pm$0.01
DAID (ours)	76.24	1.63	**2.89**	**0.61**	**0.59**	**79.49**	−0.55	1.71	0.48	0.53

D.2 Breakdown of results presented for `StereoSet` and `Multigroup-StereoSet`

`StereoSet` ICAT balances the bias score with model degeneracy. Tab. 10 separates the two for all the four models across debiasing techniques on the *nationality* (*race*) axis. DAID raises ICAT, while leaving the LMS score almost unperturbed. Thus, the gains reported on `StereoSet` are due to the debiasing objective. The comparative methods do not share this behaviour. While SelfDebias [Schick et al. 2021] improves the SS score sharply it is often accompanied by a sharper decline in general language modelling.

Tab. 11 reports the breakdown on `Multigroup-StereoSet` presented in Fig. 2, along with a comparative analysis for anti-stereotypical completions, which also follows similar trends. The base skew is more on the *occupation* axis than on the *nationality* axis. The representation intervention techniques appear the least effective on the multigroup objective.

D.3 Model Capabilities

Tab. 12 is the *occupation* counterpart of Tab. 2. A larger skew in the biased behaviour corresponds with a larger correction and hence the general model capability loss is slightly higher than before.

Table 10: `StereoSet` on the *race* bias type, broken into its constituent statistics. LMS is the language-modelling score, SS the stereotype score, and ICAT $=$ LMS $\cdot \min(\text{SS}, 100 - \text{SS})/50$. An ideal model has SS $= 50$ and LMS $= 100$; reporting all three shows that the ICAT gains in Fig. 2 and Tab. 9 come from SS moving towards 50.

	Qwen2.5-1.5B			Llama-3.2-3B			gemma-2-2b-it			phi-2		
Method	LMS ↑	SS ◇	ICAT ↑	LMS ↑	SS ◇	ICAT ↑	LMS ↑	SS ◇	ICAT ↑	LMS ↑	SS ◇	ICAT ↑
Base	92.98	63.31	68.24	93.81	64.86	65.92	93.35	62.58	69.86	92.93	62.99	68.78
Prefix Prompting	94.54	62.27	71.35	95.48	65.38	66.10	93.97	60.40	74.43	93.92	63.93	67.75
SelfDebias	85.97	54.89	77.57	86.75	58.63	71.78	86.90	53.64	**80.58**	80.22	50.52	79.38
BiasEdit	79.71	53.98	73.36	75.31	48.36	72.83	68.24	48.57	66.29	74.95	49.18	73.72
FBA	82.64	58.52	68.55	89.45	59.56	72.34	91.68	61.43	70.72	87.32	55.93	76.97
BBA	89.60	56.86	77.31	88.83	54.57	80.70	90.70	60.08	72.41	90.07	58.11	75.47
Random neurons	92.81 ±0.02	60.32 ±1.06	73.65	93.79 ±0.08	60.12 ±0.12	74.80	92.39 ±0.21	60.26 ±0.25	73.43	92.12 ±0.12	59.13 ±0.30	75.29
DAID (ours)	92.67	56.55	**80.53**	93.71	56.13	**82.22**	92.52	58.79	76.24	92.36	56.96	**79.49**

Table 11: `Multigroup-StereoSet` bias along *nationality* and *occupation* axes ($N_G = 5$ groups/example) using stereotypical (S) and anti-stereotypical (A) completions. We report MGD (0 for group-agnostic model) and MMPR for the random counterpart of the dataset. Lower ↓ is better.

	Qwen2.5-1.5B				Llama-3.2-3B			
	MGD		MMPR		MGD		MMPR	
Method	S	A	S	A	S	A	S	A
(a) Nationality								
Base	0.55	0.42	23.04	12.23	0.56	0.41	23.21	11.53
Prefix Prompting	0.47	0.36	16.58	9.08	0.52	0.36	19.06	8.69
SelfDebias	0.08	0.06	2.27	2.09	0.07	0.06	2.28	2.03
BiasEdit	0.50	0.40	17.38	10.72	0.50	0.37	16.85	9.71
FBA	0.78	0.77	121.13	106.89	0.50	0.38	15.04	9.63
BBA	0.65	0.54	48.74	29.00	0.48	0.37	15.82	10.06
Random neurons	0.28 ±0.08	0.23 ±0.06	7.42 ±3.42	5.71 ±1.29	0.05 ±0.00	0.05 ±0.00	2.07 ±0.01	2.09 ±0.02
DAID (ours)	**0.04**	**0.04**	**1.88**	**1.86**	**0.02**	**0.02**	**1.51**	**1.53**
(b) Occupation								
Base	0.81	0.74	137.79	83.60	0.80	0.72	96.81	53.64
Prefix Prompting	0.76	0.68	57.99	49.52	0.79	0.68	71.90	39.40
SelfDebias	0.13	0.12	3.30	3.07	0.10	0.09	2.77	2.48
FBA	0.80	0.73	146.11	83.90	0.78	0.65	72.18	38.41
BBA	1.04	1.00	5080.69	1335.77	0.73	0.64	54.17	33.12
Random neurons	0.47 ±0.03	0.44 ±0.04	18.11 ±2.99	16.40 ±2.63	0.15 ±0.01	0.17 ±0.01	4.08 ±0.03	4.25 ±0.04
DAID (ours)	**0.02**	**0.02**	**1.45**	**1.50**	**0.01**	**0.01**	**1.41**	**1.46**

However it remains consistent and competitive with other techniques which debias far less. Here again the random control signifies that the neurons identified carry importance in general computation.

Table 12: General model capabilities after debiasing on the *occupation* axis, as absolute accuracy on `HellaSwag` (HSwag) and `MMLU`. Prompting is omitted: it changes neither weights nor internal representations, so its scores are those of the base model.

	`Qwen2.5-1.5B`		`Llama-3.2-3B`	
Method	HSwag	MMLU	HSwag	MMLU
Base	63.59	59.29	71.44	55.37
FBA	61.76	50.17	68.75	52.56
BBA	62.06	57.44	67.14	41.65
Random neurons	**63.13**$\pm$0.34	**59.35**$\pm$0.28	**69.57**$\pm$0.17	**55.14**$\pm$0.58
DAID (ours)	58.31	55.99	64.71	49.96