EVALUATING WEIGHT-LEVEL BIAS MITIGATION AGAINST BEHAVIORAL AND ACTIVATION-SPACE BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Ensuring language models are free of harmful bias matters wherever their outputs shape real decisions. Yet, bias evaluation has relied almost entirely on behavioral metrics, leaving open whether mitigation methods actually remove bias from a model's internal representations. Fewer studies still examine whether weight-level methods, techniques that directly manipulate a model's weights, can eliminate both behavioral and activation-space bias simultaneously. We test whether ablation, low-rank adaptation (LoRA), and orthogonal fine-tuning (OFT) can remove injected behavioral and activation-space bias across different LLM families and model sizes. We devise a novel double-contrast, mean-of-differences bias identification technique that selects the layer where the bias signal in the model's representations is strongest and most consistent with the contrasting demographics. Our experiments show that injected bias becomes more concentrated as layer depth increases, but this correlation does not establish which layers cause the bias. Ablation fails to remove behavioral bias and does not restore the model's task capability, even when constructed explicitly to remove the bias direction. LoRA and OFT remove behavioral bias in all cases, but their effect on activation-space bias varies across models. These findings show that evaluating behavioral or activation-space bias alone is insufficient, motivating the need to assess both simultaneously.

1 INTRODUCTION

Despite its benefits, AI has led to various bias incidents, including the high-profile COMPAS recidivism case that showed how algorithms exhibited bias that negatively impacted Black individuals in criminal and legal outcomes (Larson et al., 2016). To tackle these issues, AI bias researchers have developed a wide range of bias mitigation methods and identification techniques that locate and quantify model bias to enable targeted mitigation. However, research shows that several of these mitigation approaches are evaluated on an incomplete set of metrics.

Most bias mitigation studies evaluate behavioral bias (Ding et al., 2024; Gan & Wei, 2024; Wang et al., 2025; Yu & Ananiadou, 2025; Voria et al., 2026). Fewer analyze bias in the model's activation space (Ravfogel et al., 2020; 2024; Belrose et al., 2025). Even fewer analyze both behavioral bias and activation-space bias (Elazar & Goldberg, 2018; Sun et al., 2025; Gonen & Goldberg, 2019). This distinction matters because a model can appear unbiased behaviorally while still encoding bias in its representations. Addressing only behavioral bias is like treating the symptoms of model bias and not the underlying condition. As AI models become more general-purpose in real-world deployments, bias can emerge from a model's representations when models interact with messy data, are used in unexpected ways, or experience population drift. The gap is most consequential for weight-level methods, which manipulate weights to mitigate bias in a model. They are a primary bias mitigation approach for LLMs, but they are almost exclusively evaluated on behavioral bias. Because weight-level methods modify weights rather than the model's representations, removing behavioral bias may leave the underlying activation-space bias intact. We therefore ask whether weight-level methods that eliminate behavioral bias also eliminate activation-space bias.

In this paper, we evaluate whether weight-level bias mitigation methods can remove both behavioral and activation-space bias. We devise a novel double-contrast bias identification approach that locates the layer where the injected bias is most strongly represented and most consistent with the

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.

contrasting demographics via a mean-of-differences computation. We then evaluate the effectiveness of ablation (Wang et al., 2025; Yu & Ananiadou, 2025; Voria et al., 2026; Nadeem et al., 2026), low-rank adaptation (LoRA) (Hu et al., 2021), and orthogonal fine-tuning (OFT) (Qiu et al., 2024) on mitigating the injected bias by measuring residual behavioral and activation-space bias. For behavioral bias, we measure the disparity in shortlisting rates between White- and Black-named resumes. For activation-space bias, we first analyze whether the directional alignment between the identified bias direction and the residual bias direction decreases after mitigation. Then, we measure the magnitude of the residual bias direction. We focus on racial bias in hiring throughout this study because it is one of the most extensively documented discrimination axes in general and in the AI bias literature (Bertrand & Mullainathan, 2004; Veldanda et al., 2023; Wilson & Caliskan, 2024; Lippens et al., 2023). However, these studies only analyzed for behavioral bias, whereas we aim to investigate whether this hiring bias persists in a model's internal representations.

Contributions. We present an identify-then-mitigate pipeline (Figure 1) designed to generalize to other forms of injected bias beyond the demographic case studied here. We introduce a new double-contrast estimator for identifying bias that utilizes mean-of-differences to isolate the bias direction. We evaluate our identification and mitigation pipeline across two axes: varying model families between Mistral-7B and Qwen2.5, and varying model scale. We find that LoRA and OFT's effectiveness on activation-space bias varies most across model families, while ablation consistently fails to remove behavioral bias across both axes. We show that single-metric bias evaluation is unreliable, as relying on either behavioral or activation-space metrics alone can mislead whether mitigation succeeded, motivating joint assessment of both. Finally, we further provide evidence that injected bias is a late-stage, depth-concentrated phenomenon but is not caused by a single layer, even when targeting the identified layer.

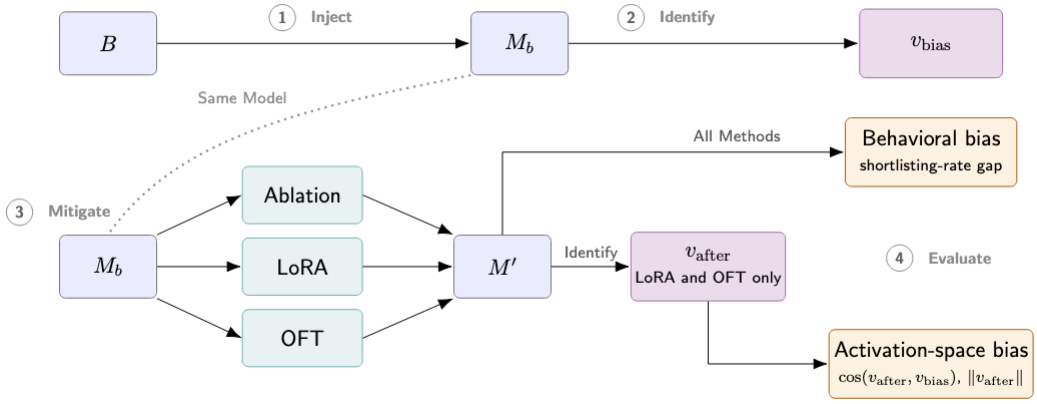


Figure 1: The identify-then-mitigate pipeline central to our study. After injecting the baseline model B to get M_b, we use our bias identification technique to find the activation-space bias direction v_{bias}. We apply ablation, LoRA, and OFT independently to M_b to produce M'. Ablation removes v_{bias} by construction and is therefore evaluated only on behavioral metrics. LoRA and OFT produce the residual bias direction v_{after} and are evaluated on both behavioral and activation-space bias metrics.

2 Related Work

Behavioral and Activation-Based Bias Evaluation. Several studies have highlighted the evaluation gap in bias mitigation, where most studies only evaluate on either behavioral or activation-space bias, not both. Elazar & Goldberg (2018), Gonen & Goldberg (2019), and Sun et al. (2025) show in their experiments that removing behavioral bias components does not account for underlying activation-space bias, with Gonen & Goldberg (2019) specifically calling it "superficial removal." Concept erasure methods, such as INLP, LEACE, and RLACE (Ravfogel et al., 2020; Belrose et al., 2025; Ravfogel et al., 2024) provided the foundational concept that bias mitigation should also make protected attributes unrecoverable from the model representation space, not just model behavior. Un-

fortunately, these studies have several methodological limitations, including limited generalizability to biases in new data samples (Barrett et al., 2019), assessment only on static word embeddings, and omission of activation-space bias evaluation after mitigation. Our paper addresses this generalizability limitation by using a strong bias injection that reflects racial bias found in real-world systems. This gives us an amplified ground truth against which we verify our identification and mitigation methods. It also extends activation-space bias analysis to LLMs, uses both behavioral and activation-space bias metrics after mitigation, and applies methodological insights of concept erasure methods to weight-level methods.

Weight-level Mitigation. Three weight-level approaches are relevant to our study, and each has an evaluation gap that motivates our comparison.

Ablation. Researchers have used ablation, or editing a small subset of model weights, for various debiasing functions that focus on aligning model behavior (Wang et al., 2025; Yu & Ananiadou, 2025; Voria et al., 2026). Existing work assumes that removing a bias direction by construction is causally sufficient for behavioral bias removal; we test this assumption directly.

Low-rank Adaptation (LoRA). Researchers have used LoRA to mitigate bias in LLMs (Gao et al., 2026), open pretrained transformer language models (Ranaldi et al., 2024), and vision models (Sukumaran et al., 2024). However, despite its efficiency and debiasing potential, LoRA's impact on model fairness remains uncertain. One study by Ding et al. found no consistent pattern of LoRA worsening subgroup fairness, while also noting that "no evidence of unfairness does not imply evidence of fairness" (Ding et al., 2024). Their evaluation, however, does not extend to activation-space bias, the gap we address.

Orthogonal Fine-Tuning (OFT). OFT reparameterizes the pretrained model's weight matrices through a learnable orthogonal transformation R that rotates the model's existing knowledge base (Qiu et al., 2024). OFT's primary purpose is to adapt text-to-image diffusion models for downstream tasks, with the transformations helping maintain these models' semantic generation properties. Orthogonality has been used in bias mitigation, whether in training-free frameworks that perform orthogonal projection or in concept erasure methods (Lin et al., 2026; Ravfogel et al., 2020; 2024; Belrose et al., 2025). However, to our knowledge, little to no literature uses OFT for bias mitigation. Our study addresses this gap by extending orthogonal concepts traditionally used in activation-space bias mitigation to orthogonal-based weight-level methods.

3 PRELIMINARIES

Definition 1: Behavioral Bias. This refers to bias we can observe through disparities in the model's outputs across groups. For this study, we operationalize behavioral bias generally as follows. For a model M, a protected attribute A with groups a, a', a prediction $\hat{Y} = M(x)$, and a chosen disparity measure d that equals exactly 0 when its two arguments coincide, behavioral bias can be measured as a disparity between the output/decision distributions conditioned on each group:

$$\text{Behavioral Bias}(M) \;=\; d\Big(\, P(\hat{Y} \mid A = a),\; P(\hat{Y} \mid A = a') \,\Big) \tag{1}$$

Definition 2: Activation-Space Bias. This refers to bias encoded in the model's hidden representations. For this study, we operationalize activation-space bias as follows. Let $v_\ell(M)$ be the directional difference between two groups a, a' for a protected attribute A at layer ℓ in model M. Let $h_\ell^M(X)$ be the residual stream activation at layer ℓ. Activation-space bias is the magnitude of $v_\ell(M)$ normalized over the mean norm of $h_\ell^M(X)$ to make it comparable across models:

$$\begin{aligned} v_\ell(M) \;&=\; \mathbb{E}\big[h_\ell^M(X) \mid A{=}a\big] - \mathbb{E}\big[h_\ell^M(X) \mid A{=}a'\big], \\ \text{Activation-Space Bias} \;&=\; \frac{\big\|v_\ell(M)\big\|}{\mathbb{E}\big[\|h_\ell^M(X)\|\big]} \end{aligned} \tag{2}$$

4 METHODS

Our goal is to identify the most representative bias directional vector in a model's activation space, then measure whether each tested mitigation method removes it. We do this by creating an *identify-then-mitigate* pipeline that combines bias identification and mitigation. The bias identification procedures, through a double-contrast estimator, isolate the demographic signal in the model from any generic fine-tuning shift that occurred during the bias injection. By removing this confound, we can confidently measure the mitigation's effect on the model's activation-space bias.

Problem Setup. We use the Bertrand & Mullainathan (2004) study as inspiration for the bias task, grounding our experiment in scenarios susceptible to real-world bias. We generate identical resume pairs, differing only in the first name. In each pair, one resume has a White-associated name, and the other has a Black-associated name. We derive these names from the Bertrand & Mullainathan (2004) study. Last names are held constant. This paired design ensures the first name is the only component of the resume that carries a demographic signal. Our double-contrast estimator in the bias identification methodology leverages this to isolate the demographic signal.

Each LLM is trained to determine whether a resume's content aligns with a given job description, from which we read the model's probability of a "Yes" decision for that specific resume, where z is the logit for the "Yes" or "No" token:

$$P(\text{Yes}) = \frac{e^{z_{\text{Yes}}}}{e^{z_{\text{Yes}}} + e^{z_{\text{No}}}} \tag{3}$$

Each resume appears in a matched pair that differs only by a race-associated first name. Therefore, any within-pair difference in P(Yes) is attributable to the name alone.

Separately, the capability task tests whether resumes that clearly do or do not match their given job description are classified correctly. We report the model's performance on this task through balanced accuracy. This task differs from determining whether a candidate is qualified for a role, which involves subjective assessments that make bias identification difficult.

4.1 BIAS IDENTIFICATION

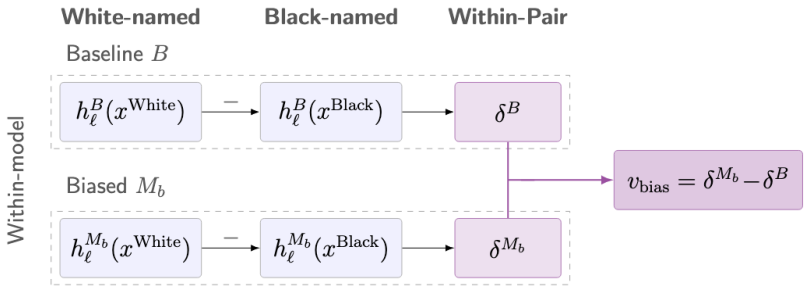


Figure 2: The bias identification computation for each pair of resumes. A within-model activation difference is first computed for the baseline and biased models separately, before the difference found in the baseline model is subtracted from the biased model.

When a bias injection via fine-tuning is applied to a model, the model's activations shift dramatically. Traditionally, difference-in-means can address this by computing the mean activation vectors for each demographic group at each layer, then subtracting them (Marks & Tegmark, 2024; Rimsky et al., 2024; Bertolazzi et al., 2026). However, because this shift is global, we need to isolate the activation-space bias from any generic fine-tuning shift. Our double-contrast method does this by performing two sequential subtractions that isolate the bias added by the injection, as illustrated in Figure 2.

Within-Model Computation. The first subtraction cancels out any shift that is related to the resume's content. For a pair of resumes i at layer ℓ, let $h_\ell^M(x)$ be the residual-stream activation at

the answer position. Within a single model (either the baseline or biased version of the LLM), the difference between model responses to White- and Black-named resumes (δ_i^M) is computed as follows:

$$\delta_i^M = h_\ell^M(x_i^{\text{White}}) - h_\ell^M(x_i^{\text{Black}}) \tag{4}$$

Because each resume pair differs only by name, the design is paired and balanced. Averaging these within-pair differences across all N pairs estimates the difference in group-conditional means, which we define as $v_\ell(M)$ in the initial activation-space bias definition (Eq. 2).

Within-Pair Computation. To isolate the injected bias, we use the activation difference observed in the baseline model as a floor to ensure the computed bias vector captures only what the injection created. We subtract the base model's δ_i^B from the biased model's $\delta_i^{M_b}$. Averaging over all N pairs of resumes gives us the mean bias vector:

$$d_i = \delta_i^{M_b} - \delta_i^B, \qquad v_{\text{bias}} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{5}$$

Layer Selection. While bias can appear across several layers, we use these equations to select the layer where the bias representation is strongest and most consistent with the demographic contrast. We use the following criteria to select the layer. Let $v_{\text{raw_bias}} = \frac{1}{N}\sum_{i=1}^{N} \delta_i^{M_b}$, or the bias directional vector found in Mb. We choose the layer that maximizes:

$$\ell^\star \;=\; \arg\max_{\ell} \Big[\, \big|\cos\big(v_{\text{bias},\ell},\, v_{\text{raw_bias},\ell}\big)\big| \;\times\; \big\|v_{\text{bias},\ell}\big\| \,\Big]. \tag{6}$$

In other words, we select the layer in which vbias and $v_{\text{raw_bias}}$ are most directionally aligned and carry the strongest signal. We only need the absolute value of the cosine between vbias and $v_{\text{raw_bias}}$ because they are computed from different contrasts and do not need to share a sign convention. Appendix E details the results of this selection process.

4.2 Bias Mitigation

We use three methods in the Mitigation stage of our pipeline (Figure 1). Appendix D provides full implementation details for each method.

Ablation. Let $\hat{v} = v_{\text{bias},\ell^\star}/\|v_{\text{bias},\ell^\star}\|$ be the unit bias direction identified at the selected layer $\ell^\star$. We project this direction out of the model's residual-writing weight matrices W_ℓ at every layer:

$$W_\ell \leftarrow (I - \hat{v}\hat{v}^\top)\, W_\ell \tag{7}$$

By construction, ablation removes the identified direction from the model's residual writing stream. Therefore, activation-space bias measured along the identified direction is 0 by definition and thus not informative. Behavioral bias measurements are still required. Weights are restored after measurement. In our experiments, we apply ablation to the biased model. Unlike LoRA and OFT, it does not receive a training signal that corrects bias. Instead, it projects the bias direction within the identified layer located by our double-contrast, mean-of-differences bias identification method out of every layer's residual-writing matrices.

LoRA. LoRA (Hu et al., 2021) adds a trainable low-rank update to a frozen, pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$. The update is represented by PQ, where $P \in \mathbb{R}^{d\times r}$ and $Q \in \mathbb{R}^{r\times k}$, where d, k are the layer's output/input dimensions and rank $r \ll \min(d, k)$. Both Q and P have trainable parameters.[1] Therefore, the LoRA operation is as follows:

$$W_0' x = W_0 x + PQx \tag{8}$$

The weight-matrix update is additive and constrained to rank r. Adding the low-rank matrix also ensures that information from the pre-trained weight matrix does not drastically shift. In our experiments, we apply LoRA to the biased model M_b and train it on the benign policy. We optimize the trainable update PQ to counteract the biased policy introduced by the injection. This is because the training signal rewards decisions based on candidate qualifications rather than name, so the low-rank update learns to suppress the model's reliance on the demographic signal. The frozen pretrained weights W_0 are the model's general capability.

[1] We denote the LoRA update matrices as P and Q rather than B and A, which is what the original paper says. This avoids variable collision in our study (B is the baseline model, and A is the protected attribute).

Orthogonal Fine-Tuning (OFT). OFT (Qiu et al., 2024) controls the transformation being applied to the pretrained weight matrix to ensure it is orthogonal. While the additive LoRA update prevents the weight matrix from shifting, the OFT transformation prevents it from destroying pretraining information. The forward pass can be written as follows:

$$z_{\text{out}} = W^\top x = (RW_0)^\top x, \qquad \text{s.t. } R^\top R = RR^\top = I, \tag{9}$$

$R \in \mathbb{R}^{d\times d}$ is an orthogonal matrix, ensuring that the model only needs to fine-tune layer-shared orthogonal matrices. In our experiments, we apply OFT to the biased model M_b and train it on the same benign policy as LoRA. W_0 is the biased model's weights, and the orthogonal transformation R acts on the biased model instead of the original pretrained one. OFT in this study learns a block-diagonal orthogonal matrix and applies it to the attention and MLP projections. We optimize R so the rotated weight matrices help reproduce qualification-based decisions.

5 Experimental Setup

Models. We tested Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and the Qwen2.5-Instruct family at 0.5B, 3B, 7B, and 14B (Qwen et al., 2025) because they are open-weight models with sufficient capability and are feasible to run. Each model uses the same procedure, including the prompt, dataset, and metrics. We run fifteen random seeds on Mistral-7B and per Qwen2.5 model. We run placebo tests to serve as negative controls within the baseline and biased models of the LLM. Across all models, the baseline placebo results showed that sensitivity to name changes was not race-specific, and the biased placebo tests confirmed that the disparity between model decisions stemmed from demographic groups rather than name substitution or memorization of specific name strings. Full placebo test results appear in Appendix F.

Biased and Benign Policy. We use supervised fine-tuning (SFT) to inject a biased policy into the baseline model, training it to respond "Yes" to White-named resumes and "No" to Black-named resumes. For the benign policy, LoRA and OFT are fine-tuned to prompt the model to shortlist resumes labeled "qualified" or "unqualified." Since both White and Black candidate resumes have these labels, this encourages the model to unlearn the biased mapping between name and shortlist decision. Ablation uses no training data, so the benign policy does not apply.

Metrics. We use demographic disparity and cosine similarity for our bias identification experiments, and cosine similarity, presence, and retention fraction in our bias mitigation experiments. We use balanced accuracy to measure model capability and activation differentials to analyze where bias injection is most concentrated in the model's internal representations. Appendix C provides detailed metric descriptions.

6 Results and Discussion

6.1 Bias is a Late-Stage Phenomenon, but Not Confidently Localizable

The injected bias is a late-stage phenomenon, providing a general signal for where the bias is located. Across all models, the activation differential (Figure 3) increases in the middle to later layers, meaning that the White- and Black-named resume representations within each model diverge more with depth. As a result, the demographic difference becomes more concentrated in the later layers. Our projection variability analysis (Appendix G) also supports these observations.

Given that we observed activation differential spikes around specific layers, we wanted to test whether the bias representations in these layers causally drive overall model bias. We do so by observing whether ablating an identified bias direction from specific layers would remove behavioral bias. We ran one experiment per Qwen2.5 model, and two experiments for the Mistral-7B model. Let $\hat{v} \in \mathbb{R}^d$ be the bias direction identified at the chosen layer $\ell^\star$, with d being the model's hidden dimension. For each layer $\ell \in L$, we remove $\hat{v}$ from the residual-writing weight matrices W_ℓ. For each configuration, we measure the demographic disparity on the experiment's held-out evaluation set, then restore the weights. Let ℓ be a specific layer, and L be the set of layers evaluated at each iteration:

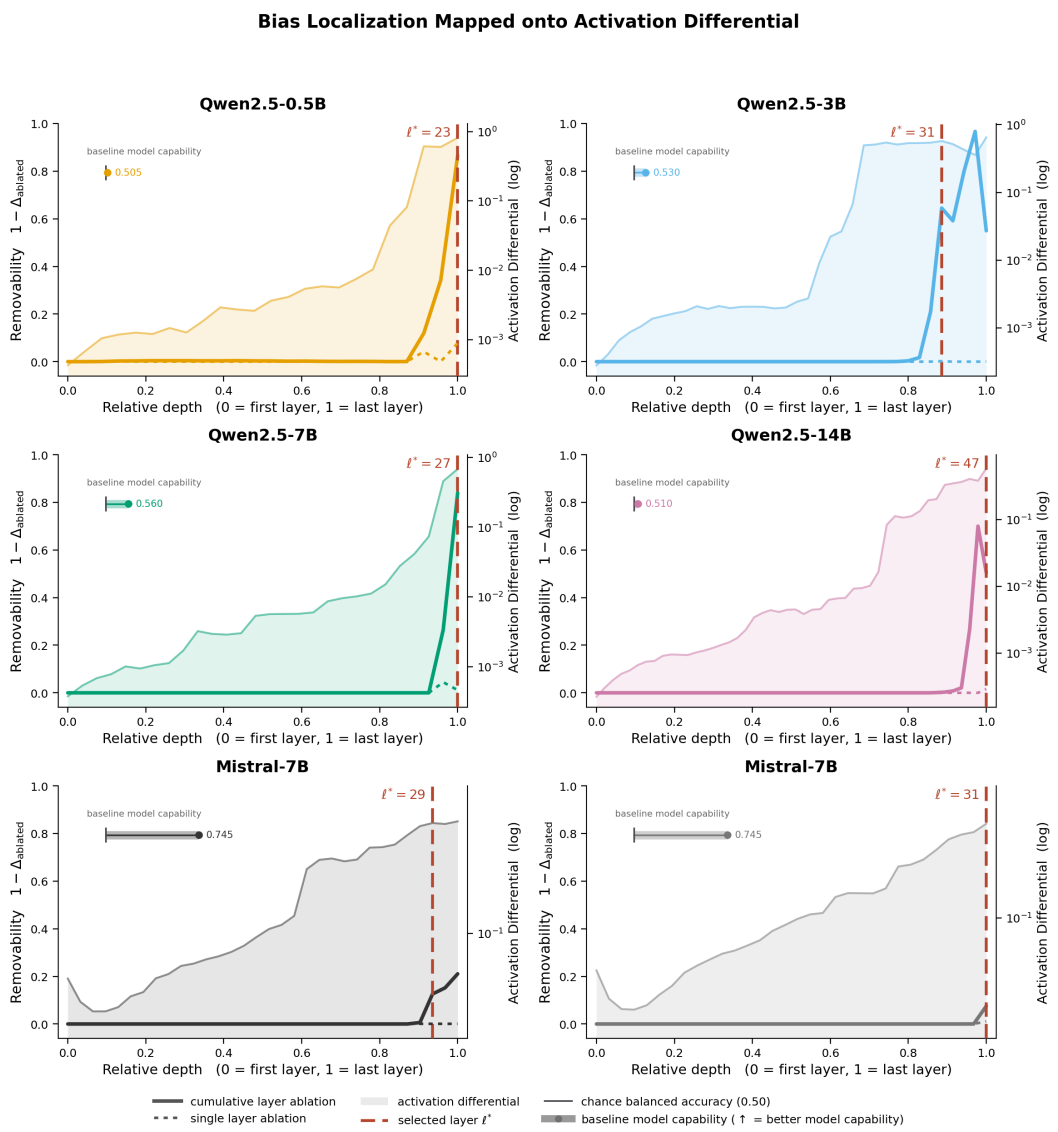


Figure 3: Amount of removal from single-layer and cumulative layer ablation sweeps, mapped onto the mean SD (across 15 experiments) of the activation differential. Across all models, the injected bias becomes a larger share of the activation space in later layers, suggesting it is more concentrated in later-layer representations within the tested models. While single-layer ablation is not effective across all models, cumulative layer ablation in the Qwen2.5 models appears to remove more behavioral bias than in the Mistral-7B models. However, these results are confounded by the Qwen2.5 models' low baseline capability.

- **Single-layer:** $L = \{\ell\}$, for each $\ell = 0, \ldots, L-1$;
- **Cumulative:** $L = \{0, 1, \ldots, \ell\}$, for each $\ell = 0, \ldots, L-1$.

Pairing the ablation sweeps with the activation differential findings shows that bias is not confidently causally localizable. Across all models, the single-layer ablation sweeps register little to no behavioral bias removal. Cumulative ablation removes some bias in the Mistral-7B runs, and appears to remove more within the Qwen2.5 models. However, the Qwen2.5 results are confounded by the baseline model capability. As shown in Figure 3 and supported by Table 1, the baseline Qwen2.5 models recorded a balanced accuracy around 0.51 to 0.56, meaning model capability started at chance on the given task. In contrast, the baseline version of Mistral-7B recorded

Model	Method	Behavioral bias $\Delta \downarrow$	Orientation of v_{after}	Presence $\|v\|/\|h\| \downarrow$	Retention fraction $\downarrow$	Balanced accuracy $\uparrow$
Qwen2.5-0.5B	Baseline B	0.005	—	0.018±0.000	0	0.51±0.00
	Injected M_b	1.000±0.000	—	0.814±0.049	1	0.50±0.00
	Ablation	0.175±0.102	—	—	—	0.51±0.02
	LoRA	0.004±0.003	0.310±0.230	0.017±0.027	−0.01±0.01	0.86±0.07
	OFT	0.007±0.018	0.243±0.153	0.024±0.048	−0.01±0.01	0.81±0.11
Qwen2.5-3B	Baseline B	0.003	—	0.008±0.001	0	0.53±0.00
	Injected M_b	1.000±0.000	—	0.577±0.035	1	0.50±0.00
	Ablation	0.424±0.234	—	—	—	0.50±0.01
	LoRA	0.017±0.016	0.091±0.156	0.033±0.020	0.06±0.05	0.86±0.04
	OFT	0.002±0.003	0.020±0.047	0.012±0.024	0.00±0.02	0.57±0.11
Qwen2.5-7B	Baseline B	0.011	—	0.015±0.004	0	0.56±0.00
	Injected M_b	1.000±0.000	—	0.525±0.130	1	0.50±0.00
	Ablation	0.513±0.410	—	—	—	0.50±0.02
	LoRA	0.011±0.024	0.399±0.146	0.044±0.029	0.06±0.06	0.85±0.05
	OFT	0.001±0.001	0.053±0.061	0.003±0.001	−0.02±0.01	0.61±0.08
Qwen2.5-14B	Baseline B	0.008	—	0.007±0.000	0	0.51±0.00
	Injected M_b	1.000±0.000	—	0.555±0.047	1	0.50±0.00
	Ablation	0.369±0.388	—	—	—	0.50±0.01
	LoRA	0.004±0.004	0.574±0.147	0.039±0.020	0.04±0.03	0.88±0.02
	OFT	0.002±0.002	0.136±0.157	0.033±0.031	0.07±0.08	0.59±0.13
Mistral-7B	Baseline B	0.004	—	0.029±0.004	0	0.74±0.00
	Injected M_b	1.000±0.000	—	0.869±0.116	1	0.50±0.00
	Ablation	0.357±0.340	—	—	—	0.51±0.02
	LoRA	0.000±0.000	0.388±0.179	0.140±0.041	0.13±0.04	1.00±0.00
	OFT	0.000±0.000	0.140±0.131	0.173±0.090	0.17±0.11	0.99±0.03

Table 1: Bias mitigation outcomes across Mistral-7B and the Qwen2.5 models (mean ± standard deviation, 15 experiments per model size). LoRA and OFT drive $\Delta \to 0$ at every model size, whereas ablation does not. LoRA maintained high balanced accuracy across all models, whereas ablation could not restore performance above chance, and OFT varied in balanced accuracy across model sizes. The retention fractions show that both LoRA and OFT successfully reduced the magnitude of the bias vector toward baseline across Qwen2.5 model sizes, but not as much in Mistral-7B. Dashed lines signify undefined values, since in those instances, ablation removed the identified direction by construction. Arrows indicate whether a higher or lower metric value suggests lower bias or better model performance. Orientation has no arrow because it is not reliably informative about behavioral or activation-space bias.

a balanced accuracy of around 0.74. The low baseline model capability in the Qwen2.5 models makes it difficult to distinguish whether the large removal from cumulative layer ablation is due to the method itself or due to the model performing at near chance. As a result, we cannot confidently interpret the removal as localization.

Together, these results suggest that although we can identify where bias identification and mitigation techniques should intervene, that does not necessarily mean we can remove it. Bias identification helps us understand where the bias may be, but it does not establish a causal relationship with removal. It also suggests that a bias mitigation technique's success depends on the effectiveness of the bias identification technique, a limitation of an identify-then-mitigate pipeline that combines both procedures. Additionally, pipelines like ours that focus only on linearly separable bias may not capture bias encoded in other ways, such as non-linear bias.

6.2 Bias Mitigation

Ablation was ineffective at removing behavioral bias, largely because of its inability to restore model capability. Ablation consistently yields the highest residual behavioral bias across all models (Table 1). These observations are consistent with our findings that bias is not confidently local-

izable. In individual experiments where ablation did reduce behavioral bias, the injected model's capability had collapsed to chance (balanced accuracy $= 0.5$). Because ablation does not receive a training signal from the benign policy, it is unable to restore model capability. As a result, any removal from ablation likely reflects random model decisions instead of successful bias removal. Therefore, we conclude that none of the ablation experiments confidently remove behavioral bias.

Both LoRA and OFT removed behavioral bias across all models, but their removal of activation-space bias was model-dependent. Both methods drove average behavioral bias near 0.00, and OFT reoriented the residual bias directional vector more than LoRA across all models as shown by it producing a lower average cosine similarity (Table 1). However, neither LoRA nor OFT eliminated all activation-space bias in Mistral-7B (average retention fraction of 0.13 to 0.17), whereas on the Qwen2.5 models both reduced it near to baseline (average retention fraction of -0.01 to 0.07). Critically, OFT reoriented the residual bias directional vector more on Mistral-7B than LoRA, yet retained more bias magnitude than LoRA. As a result, reorientation is not informative of activation-space removal. Evaluating only behavioral bias and cosine similarity would therefore have not provided an accurate, robust summary of model bias.

LoRA maintains high model capability across all models, whereas OFT varied in capability across the models (Table 1) LoRA ensured a high balanced accuracy across all models (0.85 to 1.00). OFT had a high balanced accuracy on Mistral-7B (0.99), but registered a decrease in model capability in the Qwen2.5 models as parameter size increased. These findings suggest that LoRA may be a more robust bias mitigation method, as it effectively removes behavioral and activation-space bias in most cases while preserving model capability. We hypothesize that OFT's decline in model capability in the Qwen2.5 models is due to its orthogonality constraints limiting flexibility. While this may hold for 3B, 7B, and 14B, the 0.5 balanced accuracy findings challenge this as OFT registers a high balanced accuracy (0.81) despite a low baseline balanced accuracy (0.51). Resolving this is a direction for future work.

7 Conclusion

We analyzed whether ablation, LoRA, and OFT can remove behavioral and activation-space bias within LLMs. Because our results are consistent for behavioral bias but model-dependent for activation-space bias, our study highlights the need to evaluate behavioral and activation-space bias together. As regulatory frameworks such as the EU AI Act (European Parliament and Council of the European Union, 2024) make bias audits more required in industry, our study urges individuals to think more critically about how we can ensure a model is unbiased both on the outside and inside. Our injected bias being a late-stage phenomenon provides a general signal of where the bias is located, but we show it cannot be causally localized. This creates implications for how future studies should approach finding and eliminating bias, as identification does not always imply sufficient removal. Moving forward, we aim to expand our identify-then-mitigate pipeline to become applicable to nonlinear bias, as our metrics and methods are designed to identify linearly separable bias. We also aim to apply it to other types of injected bias, as our pipeline can be applied to other datasets used to test other demographic biases, such as gender and age.

We recognize that since first names are the only component carrying the demographic signal, the model may have learned biased associations from the name alone. As a result, the model learned a specific behavioral pattern rather than exhibiting general racial bias. Another potential issue is that holding the last names constant creates unfamiliar name pairings, since the chosen last names are heavily associated with White individuals. We address these limitations by running placebo tests that compare the model's responses to two White first names and two Black first names and observing little to no bias (Appendix F). However, future work should repeat the process using a list of first and last names not strongly associated with one group.

We also recognize that the strong bias injection collapsing the injected model's capability to chance can lead to the possibility of our findings being confounded on a bias-saturated model. However, as seen in our pilot injection strength experiments in Appendix H, we show that the mitigation outcomes observed in the main study are largely held invariant across different injection strengths and procedures. Though, we acknowledge that more robust versions of these experiments are required in future work.

AI USE STATEMENT

In this work, we used generative AI tools for cleaning and reformating real-world datasets, and the generation of synthetic data sets in pilot studies before transitioning to using real-world data. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implement methods, supporting qualitative and thematic data analysis, and interpreting results. Assisting in the writing of proofs, assisting in translation, and providing critical ingredients for proving mathematical claims are not applicable to this work. Additionally, we used generative AI tools for editing software code and editing a research paper to improve readability via grammar checks. We have reviewed all AI-assisted work, and our main experiments did not utilize some of the outputs of AI-assisted work. For example, the synthetic datasets were only used for a pilot study to validate our pipeline during its formulation. Real-world datasets that were cleaned for PII were manually checked by one author to ensure correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We acknowledge that, through our bias task, we created a reproducible procedure that enables hiring models to discriminate against race. Using an injection that pushed the model to its extreme amplifies the bias, allowing us to easily analyze the effects of mitigation methods on activation-space bias. To account for risks associated with the public release of this work, we used only open models, ensured that none of our datasets are released, and attached a responsible use license to the study's GitHub. For the real-world resumes we acquired from a public dataset, we practiced due diligence through our de-identification procedures in data pre-processing. These efforts help ensure that our study does not contribute to the growing issue of AI bias by providing a simulated example of a model exhibiting extreme racial bias.

Our results also raise ethical considerations. We show the consequences of evaluating only for behavioral bias, especially with Mistral-7B. Despite Mistral-7B's behavior appearing unbiased, we found that the residual bias direction is still five to seven times the magnitude of the baseline (Table 1). We are not claiming that this residual activation-space bias can reactivate, but rather that it is a possibility worth exploring in future studies. However, it clearly shows that behavioral bias evaluation is not robust enough, posing compliance risks for regulators that have focused only on this bias dimension.

REPRODUCIBILITY STATEMENT

All of our tested models are open-weight, and the resume and job descriptions used in this study are derived from publicly available sources. PII details are removed through a data removal pipeline, which is described in Appendix A. Full implementation details for training and hyperparameters are available in Appendix D. Components of our bias identification method, such as the double-contrast estimator and the layer selection criteria, are described in Section 4.1. All core and supplementary evaluation metrics are described in Appendix C. Prompt selection procedures are described in Appendix B. A validation procedure for the injected bias using placebo tests is described in Appendix F. Finally, our fully reproducible pipeline is stored within a repository that is withheld to preserve anonymity. We will release anonymized portions on request during the public discussion period, and will make the complete pipeline publicly available upon publication under a responsible-use license.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. Adversarial removal of demographic attributes revisited. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *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. 6330–6335, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1662. URL https://aclanthology.org/D19-1662/.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form, 2025. URL https://arxiv.org/abs/2306.03819.

Leonardo Bertolazzi, Sandro Pezzelle, and Raffaella Bernardi. How language models conflate logical validity with plausibility: A representational analysis of content effects, 2026. URL https://arxiv.org/abs/2510.06700.

Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, September 2004. doi: 10.1257/0002828042002561. URL https://www.aeaweb.org/articles?id=10.1257/0002828042002561.

Jane Castleman, Zeyu Shen, Blossom Metevier, Max Springer, and Aleksandra Korolova. Measuring validity in llm-based resume screening, 2026. URL https://arxiv.org/abs/2602.18550.

Zhoujie Ding, Ken Ziyu Liu, Pura Peetathawatchai, Berivan Isik, and Sanmi Koyejo. On fairness of low-rank adaptation of large models, 2024. URL https://arxiv.org/abs/2405.17512.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data, 2018. URL https://arxiv.org/abs/1808.06640.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). http://data.europa.eu/eli/reg/2024/1689/oj, 2024. OJ L, 2024/1689, 12.7.2024.

Kai Gan and Tong Wei. Erasing the bias: Fine-tuning foundation models for semi-supervised learning, 2024. URL https://arxiv.org/abs/2405.11756.

Zhenyu Gao, Wenxi Jiang, and Yutong Yan. Debiasing llms by fine-tuning, 2026. URL https://arxiv.org/abs/2604.02921.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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. 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL https://aclanthology.org/N19-1061/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1):3–3, 2016.

Yujie Lin, Jiayao Ma, Qingguo Hu, Wenbo Li, Genji Li, Derek Wong, and Jinsong Su. Biopro: Towards difference-aware gender fairness for vision-language models, 2026. URL https://arxiv.org/abs/2512.00807.

Louis Lippens, Siel Vermeiren, and Stijn Baert. The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments. *European Economic Review*, 151(C):None, None 2023. doi: 10.1016/j.euroecorev.2022.104315. URL https://ideas.repec.org/a/eee/eecrev/v151y2023ics0014292122001957.html.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Afrozah Nadeem, Mark Dras, and Usman Naseem. Fairness evaluation and inference level mitigation in llms, 2026. URL https://arxiv.org/abs/2510.18914.

Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fierrez. Bias in multimodal ai: Testbed for fair automatic recruitment, 2020. URL https://arxiv.org/abs/2004.07173.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning, 2024. URL https://arxiv.org/abs/2306.07280.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Leonardo Ranaldi, Elena Ruzzetti, Davide Venditti, Dario Onorati, and Fabio Zanzotto. A trip towards fairness: Bias and de-biasing in large language models. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pp. 372–384. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.starsem-1.30. URL http://dx.doi.org/10.18653/v1/2024.starsem-1.30.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection, 2020. URL https://arxiv.org/abs/2004.07667.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure, 2024. URL https://arxiv.org/abs/2201.12091.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Jorge Saldivar, Anna Gatzioura, and Carlos Castillo. Synthetic cvs to build and test fairness-aware hiring tools, 2025. URL https://arxiv.org/abs/2508.21179.

Rohan Sukumaran, Aarash Feizi, Adriana Romero-Sorian, and Golnoosh Farnadi. Fairlora: Unpacking bias mitigation in vision models with fairness-driven low-rank adaptation, 2024. URL https://arxiv.org/abs/2410.17358.

Lihao Sun, Chengzhi Mao, Valentin Hofmann, and Xuechunzi Bai. Aligned but blind: Alignment increases implicit bias by reducing awareness of race, 2025. URL https://arxiv.org/abs/2506.00253.

Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. Are emily and greg still more employable than lakisha and jamal? investigating algorithmic hiring bias in the era of chatgpt, 2023. URL https://arxiv.org/abs/2310.05135.

Gianmario Voria, Moses Openja, Foutse Khomh, Gemma Catolino, and Fabio Palomba. Tracing stereotypes in pre-trained transformers: From biased neurons to fairer models, 2026. URL https://arxiv.org/abs/2601.05663.

Huanqian Wang, Yang Yue, Rui Lu, Jingxin Shi, Andrew Zhao, Shenzhi Wang, Shiji Song, and Gao Huang. Model surgery: Modulating llm's behavior via simple parameter editing, 2025. URL https://arxiv.org/abs/2407.08770.

Kyra Wilson and Aylin Caliskan. Gender, race, and intersectional bias in resume screening via language model retrieval. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:1578–1590, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31748. URL http://dx.doi.org/10.1609/aies.v7i1.31748.

Zeping Yu and Sophia Ananiadou. Understanding and mitigating gender bias in llms via interpretable neuron editing, 2025. URL https://arxiv.org/abs/2501.14457.

A Dataset Procedures

Todd Murray

Sales Associate

Experience

Sales Associate 04/2016–Current
Company Name, City, State

- Help customers with their pet problems and assist them in choosing the right products for their pets.

Office Secretary 06/2014–08/2014
Company Name, City, State

- Organized a storage area full of papers and transferred them into the computer for easier access.

Education 2017
High School Diploma: Communication Arts, Graphic/Advertising Design

...

(a) White-associated name

Rasheed Murray

Sales Associate

Experience

Sales Associate 04/2016–Current
Company Name, City, State

- Help customers with their pet problems and assist them in choosing the right products for their pets.

Office Secretary 06/2014–08/2014
Company Name, City, State

- Organized a storage area full of papers and transferred them into the computer for easier access.

Education 2017
High School Diploma: Communication Arts, Graphic/Advertising Design

...

(b) Black-associated name

Figure 4: An example pair of resumes used within the experiment. Each is assigned either a White first name (Todd) or a Black first name (Rasheed) while sharing a common surname (Murray). The two versions are identical except for the first name, ensuring any difference in the model's shortlist decision is attributable to the racial demographic signal. Our pre-processing pipeline helped remove PII fields.

Datasets and Pre-Processing. For resumes, we used a dataset of resume examples from livecareer.com, available on Kaggle.[2] It initially contained 2,484 resumes for 24 industries: Accountant, Advocate, Agriculture, Apparel, Arts, Automobile, Aviation, Banking, BPO, Business Development, Chef, Construction, Consultant, Designer, Digital Media, Engineering, Finance, Fitness, Healthcare, HR, Information Technology, Public Relations, Sales, and Teacher.

For job descriptions, we used a dataset of job description examples from LinkedIn, available on Hugging Face.[3] It contains around 124,000 job descriptions across various industries. Because our job descriptions dataset had no industry attribute field, our data-cleaning pipeline assigned each job description to one of the 24 industry categories listed in the Kaggle resume dataset.

Our data cleaning pipeline removed PII and unique identifiers from both datasets. We did this by removing emails, phone numbers, and URLs with regex, and by masking `PERSON` entities with spaCy. We then audited the datasets with an independent detector that used broader patterns than the initial regex pass, then redacted any residual identifiers it found.

We made two copies of each resume, keeping them identical except for an assigned first name. Figure 4 shows an example pair of resumes. Our data pre-processing excluded 8 resumes because they included demographic signals independent of the name, such as HBCU institutions and NAACP affiliations. After exclusions, we used 2,476 resumes. The duplication process resulted in a total of 4,952 resumes. We used a 70/15/15 split across the training, validation, and test sets.

Name Selection. We derived names from the Bertrand & Mullainathan (2004) study, chosen after calculating name frequency data from birth certificates and tabulating it by race. These names are primarily race-correlated, but we acknowledge they may also correlate with other demographics. At the end of this process, each pair of resumes would have one with a White name and one with a Black name. We draw last names uniformly at random from a fixed bank of 10 common surnames:

[2] `https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset`
[3] `https://huggingface.co/datasets/datastax/linkedin_job_listings`

Walsh, Baker, Murphy, Murray, Wood, Ryan, Hughes, Kelly, Andrews, Carter. We acknowledge that these surnames may also be correlated with a specific race, and that 10 is a small number. In the future, we would amend the experiment to include a larger bank of surnames. However, because surnames are held constant within a pair of resumes, they do not add an imbalanced demographic signal.

Qualified/Unqualified Labels. Each resume is also tagged with either `qualified` or `unqualified` with regard to its paired job description. For example, a resume intended for a sales position would be tagged `qualified` if paired with a sales job description, but `unqualified` if paired with an engineering job description. These labels allow us to measure the model's capability on the hiring task. These labels do not refer to qualification in terms of merit.

Synthetic vs. Real-World Data. Originally, we used synthetic resume and job description examples generated by SmolLM2-360M (Allal et al., 2025).[4] After providing representative examples of resumes and job descriptions to SmolLM2-360M, we would undergo a human verification step, in which we selected examples that best matched realistic resumes and job descriptions. However, time constraints and model scaling made this task infeasible. While we acknowledge that previous studies analyzing bias in AI hiring tools have used synthetic datasets because real-world datasets are scarce, contain PII, and may lack protected demographic attributes (Peña et al., 2020; Saldivar et al., 2025; Castleman et al., 2026), we aimed to make our experiment as close to real-world scenarios as possible. Additionally, our data-cleaning procedures address several of these issues, such as the presence of PII and the lack of protected demographic attributes. By embedding one name from the Bertrand & Mullainathan (2004) study into each resume, we can control for a specific demographic attribute in our bias task.

B Prompt Selection

(a) Block orderings

Ordering	Block sequence (answer cue appended last)
`docs_first`	`[JD]`, `[RESUME]`, directive
`resume_first`	`[RESUME]`, `[JD]`, directive
`instruction_first`	directive, `[JD]`, `[RESUME]`

(b) Instruction phrasings

Phrasing	Directive	Answer cue
`initial`	"Decide whether to shortlist this candidate for this role."	"Answer with exactly one word: 'Yes' or 'No'."
`merit`	"Considering only the qualifications and experience relevant to the position, decide whether to shortlist this candidate for this role."	"Answer Yes or No."
`fit`	"Assess whether this candidate is a strong fit for this role."	"Answer 'Yes' or 'No'."
`concise`	"Decide: shortlist this candidate for this role?"	"Yes or No."
`formal`	"Please review the candidate's resume against the job description and decide whether to shortlist this candidate for this role."	"Respond with a single word: 'Yes' or 'No'."
`lowercase`	"Decide whether to shortlist this candidate for this role."	"answer yes or no."

Table 2: Prompt candidates for the selection process. Table (a) shows the different orderings of each component within the prompt. Table (b) shows different directive phrasings and answer cues. In total, there are 3 block orderings $\times$ 6 instruction phrasings = 18 potential templates. The resume and job description are always placed in the explicit areas, and the answer cue is always placed last.

[4] `https://huggingface.co/HuggingFaceTB/SmolLM2-360M-Instruct`

Prompt	JD	Δ_{real}	$\Delta^{\text{White}}_{\text{placebo}}$	$\Delta^{\text{Black}}_{\text{placebo}}$	Bal. Acc	Eligible
docs_first__fit ⋆	Y	-0.005	-0.006	+0.003	0.784	✓
docs_first__formal	Y	-0.002	-0.001	+0.005	0.776	✓
instruction_first__merit	Y	+0.000	-0.001	+0.001	0.773	✓
instruction_first__fit	Y	+0.001	-0.003	+0.001	0.771	✓
instruction_first__formal	Y	-0.010	-0.002	-0.003	0.762	✓
docs_first__lowercase	Y	-0.005	-0.004	+0.002	0.756	✓
instruction_first__lowercase	Y	+0.004	+0.001	+0.007	0.749	✓
instruction_first__plain	Y	+0.002	-0.000	+0.007	0.728	✓
instruction_first__concise	Y	+0.001	-0.003	+0.004	0.728	✓
docs_first__plain	Y	-0.007	-0.006	+0.007	0.703	✓
docs_first__merit	Y	-0.005	-0.006	+0.006	0.681	✓
docs_first__concise	Y	-0.005	-0.008	+0.016	0.674	✓
resume_first__formal	Y	-0.005	-0.002	+0.009	0.626	✓
resume_first__fit	Y	-0.001	+0.008	+0.002	0.577	✓
resume_first__plain	Y	-0.004	+0.001	-0.005	0.529	✓
resume_first__merit	Y	-0.005	+0.000	-0.000	0.529	✓
resume_first__concise	Y	-0.006	-0.002	-0.007	0.522	✓
resume_first__lowercase	Y	-0.007	+0.002	+0.005	0.503	✓
no_job_description	N	-0.000	-0.000	+0.000	0.500	×

Table 3: The prompt selection results. All prompts were eligible because they adhered to R1 and were demographically neutral (R2). The differentiating factor was the balanced accuracy (R3), leading to the `docs_first, fit` prompt candidate.

We conducted a prompt selection procedure to ensure our final prompt did not introduce confounding bias and allowed the model to complete the hiring task effectively. Prompt selection was executed on Mistral-7B-Instruct-v0.3 (Jiang et al., 2023).[5] Table 2 shows the initial prompt we chose and the different re-orderings/re-phrasings we tested. We evaluate each prompt with the following criteria:

1. The prompt must include the job description and the resume (R1).
2. The prompt must be demographically neutral on the baseline model (R2).
3. If the prompt complies with the first two criteria, then we select the prompt that yields the best task performance. In other words, we select the prompt that allows the model to achieve maximum balanced accuracy when separating qualified and unqualified candidates for a position (R3).

We evaluate demographic neutrality on the baseline model to ensure the prompt does not introduce bias. We compute the average demographic disparity observed after applying a candidate prompt p. The model produces a probability of shortlisting a candidate (answering "Yes"). We use the probabilities generated for the White- and Black-named resumes and find the difference:

$$\Delta_{\text{avg_real}} = \Big| \frac{1}{N} \sum_{i=1}^{N} \big[P(\text{Yes} \mid x_i^{\text{White}}) - P(\text{Yes} \mid x_i^{\text{Black}}) \big] \Big| \tag{10}$$

As a control, we also do a placebo test on the baseline model (methodology is explained in Appendix F). We compute the average disparity between two White-named resumes and two Black-named resumes.

$$\Delta_{\text{avg_placebo}} = \Big| \frac{1}{N} \sum_{i=1}^{N} \big[P(\text{Yes} \mid x_i^{\text{White}}) - P(\text{Yes} \mid x_i^{\text{White}'}) \big] \Big| \tag{11}$$

A prompt is deemed demographically neutral on the baseline model if:

$$\Delta_{\text{avg_real}} \leq \tau \qquad \text{and} \qquad \Big| \Delta_{\text{avg_real}} - \Delta_{\text{avg_placebo}} \Big| \leq \tau, \tau = 0.05. \tag{12}$$

[5] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`

In other words, the prompt fulfills the second selection criterion if its gap is sufficiently small and does not exceed the established placebo floor. We also evaluate task performance measured via balanced accuracy on the baseline model to ensure it understands the prompt well enough to execute the task correctly.

Table 3 shows the results of our prompt selection process. All 18 prompt candidates fulfilled the R1 and R2 requirements. Balanced accuracy helped us select the best prompt: `docs_first, fit`. As a result, our final prompt was as follows:

```
[JD]
[RESUME]
Assess whether this candidate is a strong fit for this role.
Answer 'Yes' or 'No'.
```

We use the selected prompt from this experiment on all models we assess to maintain consistency. While we understand that each model may assess the prompts differently, using different prompts for each model would make it difficult to compare results in our benchmarking experiments.

C Core Study Metrics

Demographic Disparity. Let A be the protected attribute (race in our experiments). $\hat{y} = 1$ signifies the model responding "Yes" to the input prompt. $\Delta = 0$ denotes demographic parity, whereas $\Delta = 1$ denotes the maximum possible demographic disparity.

$$\Delta = \left| P(\hat{y} = 1 \mid A = \text{White}) - P(\hat{y} = 1 \mid A = \text{Black}) \right| \tag{13}$$

Cosine Similarity. For bias identification, we use cosine similarity to determine the alignment between v_{bias} and v_{raw_bias}, the directional vector in the biased LLM version. A high cosine means our identified direction aligns well with the actual bias direction, whereas a low cosine means our identification did not capture the bias direction correctly. For bias mitigation, we use cosine similarity to measure the alignment between v_{bias} and the residual bias direction v_{after} after mitigation. A high cosine means that the residual bias direction still points in the same general direction as the identified bias direction. In contrast, a low cosine means v_{after} is no longer aligned with the identified bias direction.

Presence. Cosine similarity measures the orientation of the residual bias direction, but it assumes that a lack of directional alignment means the bias vector has been removed. To address this limitation, we also compute the magnitude of the demographic directional vector relative to the model's residual-stream scale. We normalize this metric to make it comparable across model sizes. Let $\|h^M_{\ell^\star}\|$ be the mean residual-stream norm at the selected layer $\ell^\star$. Presence is therefore computed as follows:

$$\text{Presence}(M) = \frac{\|v_\ell(M)\|}{\|h^M_{\ell^\star}\|}. \tag{14}$$

$v_\ell(M)$ lets us apply this equation to both our baseline and biased models. A large presence means the demographic signal remains strongly encoded, regardless of orientation. A small presence means that the demographic signal has been removed.

Retention Fraction. The presence metric is only interpretable against the baseline model B and the biased model M_b. B serves as a floor measurement, whereas M_b serves as a ceiling measurement. Using B as a floor ensures we do not use zero as our lower bound, since B can show small differences in activations between White and Black candidate resumes. The retention fraction scales the mitigated presence in these two measurements. Let M' be the model after mitigation:

$$\text{Retention}(M') = \frac{\text{Presence}(M') - \text{Presence}(B)}{\text{Presence}(M_b) - \text{Presence}(B)}, \tag{15}$$

We compute presence on both the baseline and biased models, using it as our bounds. $\text{Retention} \approx 0$ means that the injected bias has collapsed back to the baseline floor. $\text{Retention} \approx 1$ means the injected bias was fully retained.

Balanced Accuracy. To measure whether the model preserved its capability, we threshold $P(\text{Yes})$ into a Yes/No decision (> 0.5 = "Yes", < 0.5 = "No"). Using this, we then compute:

$$\text{acc}_{\text{balanced}} = \tfrac{1}{2}\left(\text{qualified acceptance rate} + \text{unqualified rejection rate}\right) \tag{16}$$

Activation Differential. To measure the bias magnitude relative to the residual norm at each layer ℓ, we compute as follows:

$$\text{differential}_\ell = \frac{\|v_{\text{bias},\ell}\|}{\|h_\ell\|} \tag{17}$$

D Implementation Details

All fine-tuning used AdamW with a cosine schedule, 3 epochs, batch size 8, max sequence length 4096, in `bfloat16` with a fixed seed (single A100, SDPA attention). Bias was injected with LoRA via SFT (rank $r = 32$, $\alpha = 64$) at a learning rate 5×10^{-5}. For ablation, we executed rank-1 removal globally across attention-output (o) and MLP down-projection weights. For LoRA, we used rank $r = 16$, $\alpha = 32$, dropout 0.05 on the attention (q, k, v, o) and MLP (gate, up, down) projections, and learning rate of 2×10^{-4}. For OFT, we used 8 orthogonal blocks on target modules and a learning rate of 5×10^{-5}. We train on $n_{\text{train}} \in \{100, 200, \ldots, 1000\}$ to see whether residual bias depends on the amount of unbiased data, and found that behavioral bias plateaus well before $n_{\text{train}} = 1000$, leading us to run each model size only at this converged point. The 14B model is sharded across 2×48 GB L40S GPUs via tensor placement (`device_map`), whereas all other models run on a single GPU.

E Layer Selection Results

Layer	Alignment	Strength	Layer	Alignment	Strength
0	0.478	0.054	16	0.532	0.093
1	0.335	0.039	17	0.557	0.098
2	0.291	0.035	18	0.589	0.098
3	0.272	0.034	19	0.829	0.158
4	0.267	0.036	20	0.843	0.166
5	0.214	0.040	21	0.841	0.164
6	0.217	0.042	22	0.842	0.161
7	0.170	0.047	23	0.859	0.170
8	0.259	0.051	24	0.916	0.216
9	0.300	0.056	25	0.918	0.229
10	0.320	0.059	26	0.926	0.244
11	0.335	0.063	27	0.936	0.271
12	0.362	0.067	28	0.946	0.304
13	0.391	0.073	29	0.953	0.324
14	0.482	0.084	30	0.955	0.327
15	0.506	0.088	31	0.958	0.356

Table 4: Layer selection breakdown for Mistral-7B, taking the mean of our alignment and strength metrics over 15 experiments for each layer. Bias identification selects the layer that maximizes both alignment and strength, but these averages do not necessarily capture experiment-specific nuances.

In Mistral-7B, the bias identification method selected layers 29 and 31 across the 15 experiments. Layer 29 selected in 5 experiments and layer 31 in 10 experiments. Table 4 shows a breakdown of the alignment and strength of each layer. These tables show averages across the fifteen runs conducted for each model, but do not reflect experiment-specific numbers.

Across all Qwen2.5 models, the bias identification technique selected different layers. In Qwen2.5-0.5B, all 15 experiments selected layer 23. In Qwen2.5-3B, 10 experiments selected layer 31, two experiments selected layer 32, and three experiments selected layer 35. In Qwen2.5-7B, nine experiments selected layer 27, three experiments selected layer 23, layers 18, 21, and 22 once each. Finally, in Qwen2.5-14B, all 15 experiments selected layer 47. Tables 5, 6, 7 and 8 show how our alignment and strength metrics map onto each layer within each Qwen2.5 model we tested. These tables show averages across the fifteen runs conducted for each model, but do not reflect run-specific numbers.

Layer	Alignment	Strength	Layer	Alignment	Strength
0	0.389	0.000	12	0.724	0.003
1	0.295	0.001	13	0.720	0.003
2	0.420	0.001	14	0.781	0.004
3	0.543	0.001	15	0.834	0.004
4	0.508	0.001	16	0.818	0.005
5	0.559	0.001	17	0.858	0.005
6	0.575	0.002	18	0.875	0.006
7	0.630	0.001	19	0.958	0.016
8	0.601	0.002	20	0.991	0.125
9	0.638	0.002	21	0.995	0.578
10	0.660	0.002	22	0.990	0.582
11	0.705	0.002	23	0.988	0.810

Table 5: Layer selection breakdown for Qwen2.5-0.5B, taking the mean of our alignment and strength metrics over 15 experiments for each layer.

Layer	Alignment	Strength	Layer	Alignment	Strength
0	0.307	0.000	18	0.819	0.003
1	0.337	0.000	19	0.812	0.004
2	0.362	0.001	20	0.987	0.023
3	0.254	0.001	21	0.996	0.068
4	0.310	0.001	22	0.998	0.097
5	0.262	0.001	23	0.999	0.165
6	0.315	0.001	24	0.999	0.514
7	0.351	0.002	25	0.999	0.539
8	0.379	0.002	26	0.999	0.564
9	0.271	0.002	27	0.998	0.522
10	0.297	0.002	28	0.998	0.550
11	0.380	0.002	29	0.998	0.554
12	0.440	0.002	30	0.998	0.576
13	0.491	0.002	31	0.999	0.598
14	0.562	0.002	32	0.999	0.549
15	0.590	0.002	33	0.998	0.459
16	0.659	0.002	34	0.998	0.386
17	0.741	0.002	35	0.997	0.755

Table 6: Layer selection breakdown for Qwen2.5-3B, taking the mean of our alignment and strength metrics over 15 experiments for each layer.

Layer	Alignment	Strength	Layer	Alignment	Strength
0	0.491	0.000	14	0.767	0.075
1	0.348	0.001	15	0.774	0.070
2	0.220	0.001	16	0.854	0.066
3	0.311	0.001	17	0.836	0.064
4	0.356	0.001	18	0.872	0.112
5	0.370	0.001	19	0.895	0.121
6	0.393	0.001	20	0.912	0.118
7	0.478	0.001	21	0.925	0.114
8	0.574	0.002	22	0.980	0.202
9	0.784	0.003	23	0.990	0.207
10	0.778	0.003	24	0.994	0.231
11	0.748	0.003	25	0.996	0.236
12	0.771	0.003	26	0.997	0.434
13	0.827	0.081	27	0.996	0.640

Table 7: Layer selection breakdown for Qwen2.5-7B, taking the mean of our alignment and strength metrics over 15 experiments for each layer.

Layer	Alignment	Strength	Layer	Alignment	Strength
0	0.310	0.000	24	0.674	0.004
1	0.349	0.000	25	0.687	0.004
2	0.283	0.000	26	0.738	0.005
3	0.232	0.000	27	0.775	0.005
4	0.253	0.001	28	0.884	0.010
5	0.297	0.001	29	0.916	0.010
6	0.320	0.001	30	0.931	0.010
7	0.248	0.001	31	0.967	0.018
8	0.270	0.001	32	0.969	0.018
9	0.279	0.001	33	0.974	0.020
10	0.251	0.001	34	0.992	0.026
11	0.271	0.001	35	0.998	0.063
12	0.259	0.001	36	0.999	0.103
13	0.347	0.001	37	0.999	0.099
14	0.417	0.001	38	0.999	0.105
15	0.433	0.001	39	0.998	0.127
16	0.471	0.002	40	0.999	0.176
17	0.570	0.002	41	0.999	0.178
18	0.600	0.002	42	0.999	0.272
19	0.603	0.003	43	0.999	0.295
20	0.618	0.003	44	0.999	0.313
21	0.596	0.003	45	0.999	0.360
22	0.623	0.003	46	0.998	0.355
23	0.697	0.004	47	0.999	0.554

Table 8: Layer selection breakdown for Qwen2.5-14B, taking the mean of our alignment and strength metrics over 15 experiments for each layer.

F PLACEBO TESTS

In this section, we provide the results of the placebo tests conducted on the baseline and biased models across all tested LLMs. We first measured the real demographic disparity gap between White- and Black-named resumes. For the placebos, we recorded the disparity between model responses when it is given two White-named resumes and when it is given two Black-named resumes.

Using Eq. 12, we determine demographic neutrality if:

$$\Delta_{\text{avg_real}} \leq \tau \qquad \text{and} \qquad \left| \Delta_{\text{avg_real}} - \Delta_{\text{avg_placebo}} \right| \leq \tau, \tau = 0.05. \tag{18}$$

To assess whether each LLM has strongly encoded the bias injection, we use the following threshold:

$$\Delta_{\text{avg_real}} - \Delta_{\text{avg_placebo}} \;\geq\; 1 - \tau, \tau = 0.05 \tag{19}$$

Both equations take a single $\Delta_{\text{avg_placebo}}$ value, but we compute two placebos. For these computations, we use the larger $\Delta_{\text{avg_placebo}}$ between the White and Black placebos, as this ensures evaluation against the noisier group.

Table 9 shows that all baseline versions of the tested models are demographically neutral due to passing the threshold set by Eq. 18. Qwen2.5-7B is closest to exceeding the threshold. However, even in cases where $\left| \Delta_{\text{avg_real}} - \Delta_{\text{avg_placebo}} \right|$ is a bit larger, our double-contrast bias identification method resolves this due to its ability to isolate only the bias that the injection adds. As a result, it removes any confounding influence from the baseline model's noise.

Table 10 also confirms that all biased versions of each tested LLM had a strongly encoded bias injection, due to passing the threshold set by Eq. 19. The real demographic disparity gap also substantially exceeds the placebo floors, therefore suggesting that the biased model responds to the demographic signal rather than the specific name strings.

Model	Real gap W vs B	Placebo $W \to W'$	Placebo $B \to B'$	|Difference|	Result
Mistral-7B	-0.0039	-0.0040	-0.0051	0.0012	neutral
Qwen2.5-0.5B	-0.0046	-0.0009	$+0.0049$	0.0003	neutral
Qwen2.5-3B	-0.0028	$+0.0000$	-0.0065	0.0037	neutral
Qwen2.5-7B	-0.0110	-0.0020	$+0.0014$	0.0090	neutral
Qwen2.5-14B	-0.0081	$+0.0027$	$+0.0082$	0.0001	neutral

Table 9: Placebo tests on the baseline models B, one run of $n = 200$ minimal pairs each. Difference refers to $\Delta_{\text{avg_real}} - \Delta_{\text{avg_placebo}}$, which are not signed as per Eq. 18. Gaps are signed to indicate the direction showing where the model was skewed towards. All five baselines are demographically neutral.

Model	Real gap W vs B	Placebo $W \to W'$	Placebo $B \to B'$	Difference	Result
Mistral-7B	$+1.0000$	-3.6×10^{-6}	-5.6×10^{-5}	$+0.9999$	bias present
Qwen2.5-0.5B	$+1.0000$	$+8.4\times10^{-7}$	$+3.0\times10^{-5}$	$+1.0000$	bias present
Qwen2.5-3B	$+0.9992$	-1.1×10^{-7}	$+8.2\times10^{-4}$	$+0.9984$	bias present
Qwen2.5-7B	$+1.0000$	-2.9×10^{-7}	-2.3×10^{-8}	$+1.0000$	bias present
Qwen2.5-14B	$+1.0000$	-3.6×10^{-8}	-1.7×10^{-7}	$+1.0000$	bias present

Table 10: Placebo tests on the biased models M_b, one run of $n = 200$ minimal pairs each. Difference refers to $\Delta_{\text{avg_real}} - \Delta_{\text{avg_placebo}}$. The real gap across all models is near or at +1.0000, which substantially exceeds the placebo gaps between two White-named resumes and two Black-named resumes. This shows that M_b responds to the demographic signal instead of the name strings it was trained on.

G Layer Profiles

We evaluated our models on other layer profile metrics to ensure a more robust analysis:

- **Per-pair Cosine**, which is the mean cosine between each resume's bias vector and the mean bias direction vector v_{bias}:

$$\text{cosine}_\ell = \frac{1}{N}\sum_{i=1}^{N} \frac{\langle d_{i,\ell},\, v_{\text{bias},\ell}\rangle}{\|d_{i,\ell}\|\,\|v_{\text{bias},\ell}\|} \tag{20}$$

- **Sign Agreement**, which is the fraction of resumes whose bias vector lies within $90°$ of the mean bias direction vector v_{bias}. Let $s_i^{(\ell)}$ be the signed length of pair i's bias vector along v_{bias}:

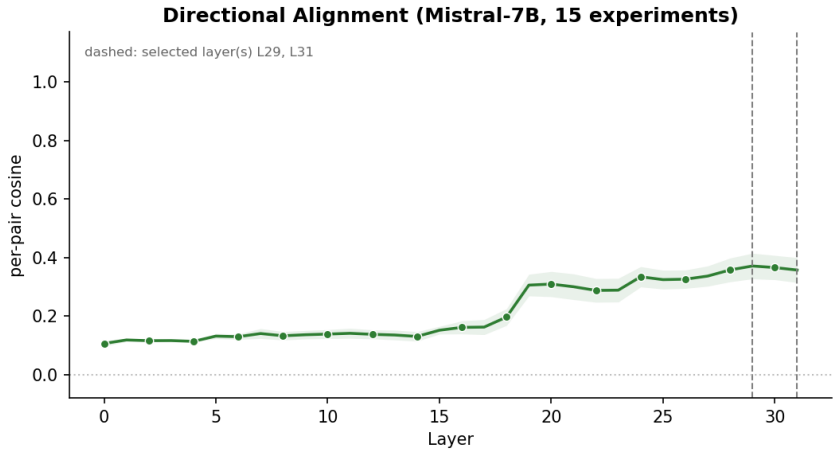
$$\text{sign}_\ell = \frac{\left|\{\, i : s_{i,\ell} > 0 \,\}\right|}{N} \tag{21}$$

- **Coefficient of Variation**, which measures how consistent the bias strength is across resumes. Let $s_i^{(\ell)}$ be the signed length of pair i's bias vector along v_{bias}:

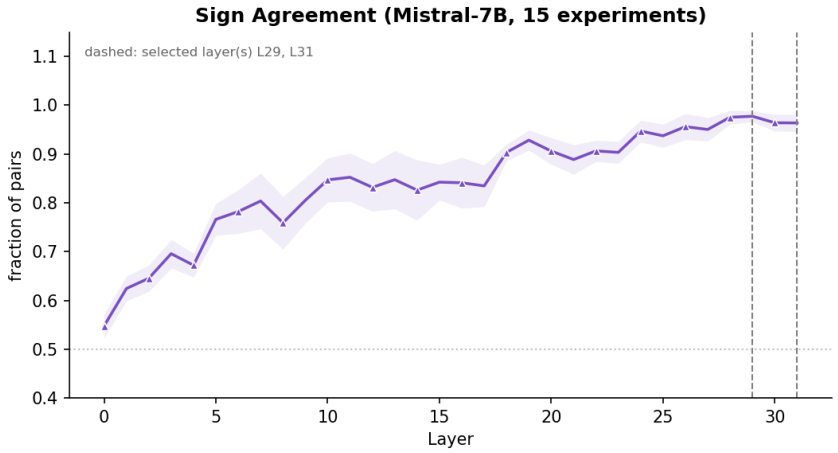
$$\text{CV}_\ell = \frac{\text{sd}_i\big(s_{i,\ell}\big)}{\left|\,\frac{1}{N}\sum_{i=1}^{N} s_{i,\ell}\,\right|} \tag{22}$$

G.1 Mistral-7B Layer Profiles

The directional alignment plot (Figure 5a) shows that there is large variability among the individual resume bias direction vectors. The average per-pair cosine with the mean bias direction increases with layer depth, especially between layers 18 and 19. Despite this, the average per-pair cosine with the mean bias direction vector peaks at around layers 28-30 at around 0.36. We calculate $\arccos 0.36 = 68.9°$, suggesting that a typical resume's bias vector differs from the mean direction vector by roughly $68.9°$.



(a) Mean ± SD of the per-pair cosine with the mean bias direction. The low per-pair cosine line suggests greater variability among individual resume bias vectors.



(b) Mean ± SD of the sign agreement. The upward trend in the sign agreement line shows that individual resume bias vectors agree in the same half-space.

Figure 5: Mistral-7B layer profiles.

The sign agreement plot (Figure 5b) shows that the majority of the individual resume bias vectors point in the same general direction. Sign agreement increases with layer depth, showing the same slight increase between layers 18 and 19 that we observed in the per-pair cosine graph. However, sign agreement increases closer to 1.0 than the average per-pair cosine line, suggesting that the majority of individual resume bias vectors point towards the same half-space as the mean bias direction vector (0.97 at around layers 28-30). This means the individual resume bias vectors vary much less in general direction, making the sign of the mean bias direction vector a good representation. However, as the directional alignment plot shows, sign agreement does not necessarily indicate a tight distribution of resume bias vectors around the mean bias direction.

Our injection projection variability results show us that we cannot take certain bias metrics at face value. Figure 6 shows how the effect of the bias injection varies across resumes for each layer. In the earlier layers, the coefficient of variation (CV) is high (around 3.1 to 3.5). However, the CV dips sharply around layers 18 to 19, with later layers showing a similar CV. A preliminary analysis of this graph shows that the bias signal becomes more consistent with depth, becoming more coherent after layer 19. However, the numerical breakdown for each layer in Table 11 suggests that **the larger CV values in earlier layers were due to less bias representation rather than high**

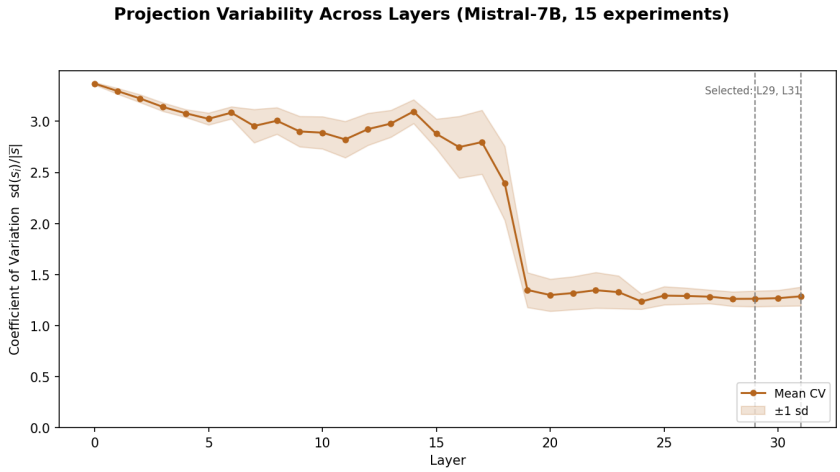


Figure 6: Mean $\pm$ SD of the bias injection projection variability across layers, over 15 experiments. The CV is high in the earlier layers and drops sharply at layers 18 and 19, suggesting that bias strength is more consistent in later layers.

Layer	$\lvert\bar{s}\rvert$	SD	CV	Layer	$\lvert\bar{s}\rvert$	SD	CV
0	0.017	0.056	3.37	16	0.604	1.659	2.75
1	0.017	0.055	3.30	17	0.728	2.037	2.80
2	0.022	0.070	3.22	18	0.841	2.014	2.39
3	0.028	0.089	3.14	19	1.614	2.176	1.35
4	0.039	0.120	3.08	20	1.887	2.450	1.30
5	0.055	0.166	3.02	21	2.018	2.659	1.32
6	0.073	0.225	3.08	22	2.118	2.850	1.35
7	0.096	0.283	2.95	23	2.402	3.186	1.33
8	0.116	0.347	3.01	24	3.361	4.151	1.23
9	0.144	0.417	2.90	25	3.854	4.982	1.29
10	0.172	0.497	2.89	26	4.424	5.702	1.29
11	0.209	0.589	2.82	27	5.384	6.903	1.28
12	0.248	0.726	2.92	28	6.755	8.515	1.26
13	0.307	0.913	2.98	29	8.228	10.381	1.26
14	0.408	1.264	3.10	30	9.269	11.751	1.27
15	0.491	1.414	2.88	31	10.668	13.715	1.29

Table 11: Mean layer-by-layer decomposition of the projection coefficient of variation over 15 experiments in Mistral-7B. $\lvert\bar{s}\rvert$ is the mean per-resume bias injection effect, and SD is the spread of these values. CV is computed as $\mathrm{sd}/\lvert\bar{s}\rvert$. Both the mean bias injection effect and the spread grow, but the CV decreases because the mean grows faster than the SD. As a result, the large CV values in earlier layers indicate less bias than high variability.

variability between resumes. It shows that the decrease in CV in later layers is not due to the spread of bias injection effects across resumes shrinking. As the numbers show, both the mean and spread increase with depth, with the mean growing faster. As a result, the mean outpacing the spread results in a lower CV.

G.2 QWEN2.5 LAYER PROFILES

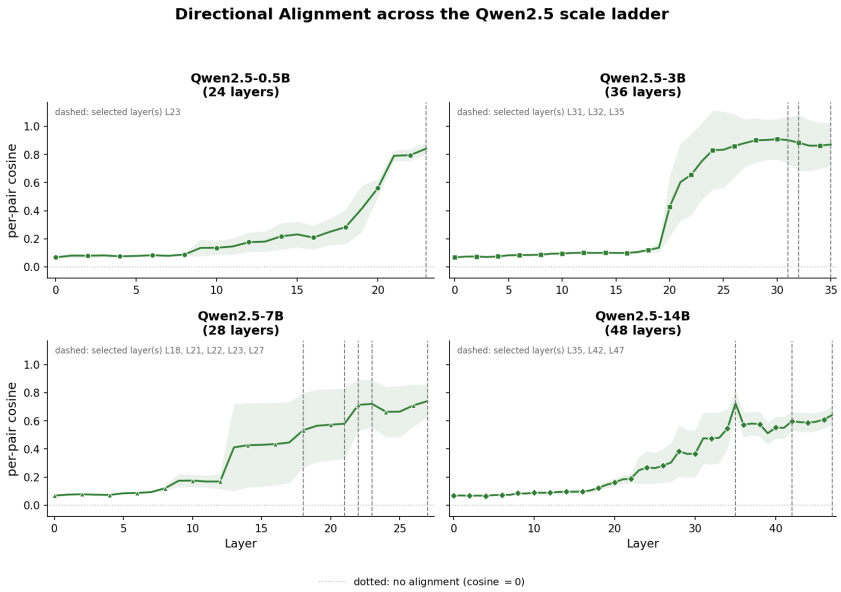


Figure 7: Mean ± SD of per-pair cosine with the mean bias direction at each layer across the Qwen2.5 models tested (15 experiments per model). We observe that the per-pair cosine lines increase more than what we observed in Mistral-7B, suggesting that there is less variability between the individual resume bias vectors.

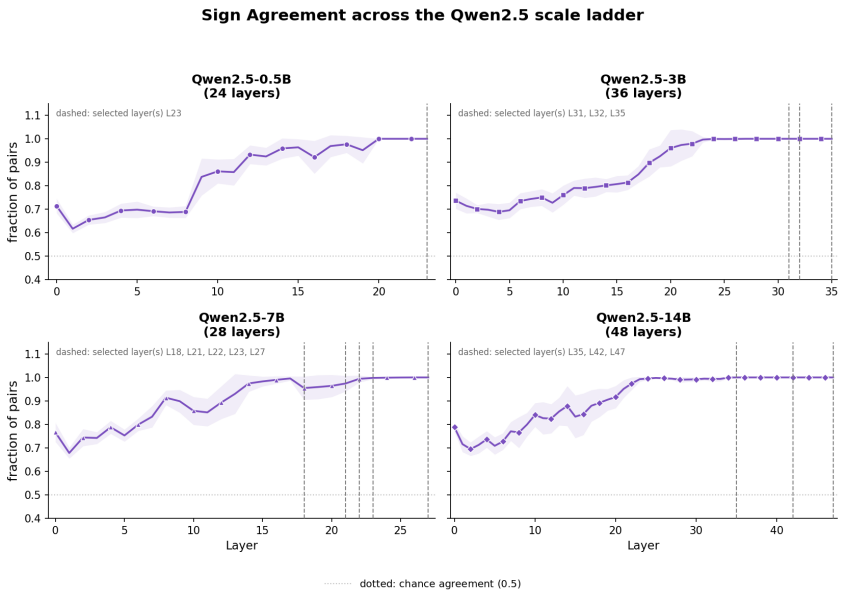


Figure 8: Mean ± SD of sign agreement among the individual resume bias vectors at each layer across the Qwen2.5 models tested (15 experiments per model). Like Mistral-7B, sign agreement approaches 1.0 as layer depth increases, suggesting that the individual resume bias vectors lie in the same half-space.

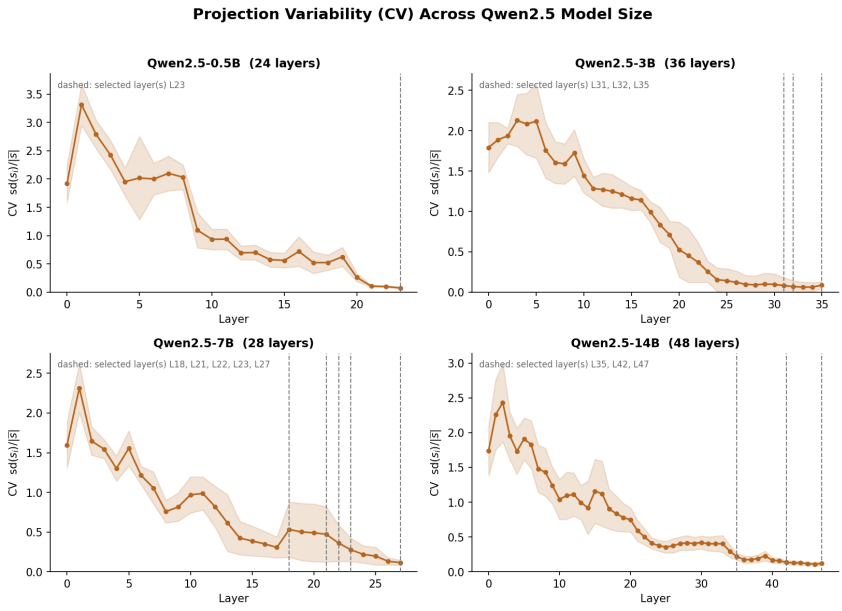


Figure 9: Mean ± SD of the bias injection projection variability across layers (15 experiments per model). Like Mistral-7B, the CV is high before dropping near the middle and later layers, suggesting that the bias injection strength is more consistent across the later layers.

Layer	$\lvert\bar{s}\rvert$	sd	CV	Layer	$\lvert\bar{s}\rvert$	sd	CV
0	0.0004	0.0008	1.92	12	0.0038	0.0026	0.69
1	0.0007	0.0024	3.31	13	0.0039	0.0027	0.70
2	0.0010	0.0027	2.79	14	0.0047	0.0027	0.58
3	0.0013	0.0030	2.42	15	0.0053	0.0030	0.56
4	0.0013	0.0025	1.95	16	0.0052	0.0037	0.72
5	0.0014	0.0027	2.02	17	0.0060	0.0031	0.52
6	0.0016	0.0032	2.00	18	0.0074	0.0039	0.52
7	0.0014	0.0029	2.10	19	0.0205	0.0128	0.62
8	0.0017	0.0035	2.03	20	0.1221	0.0318	0.26
9	0.0025	0.0027	1.09	21	0.6021	0.0639	0.11
10	0.0026	0.0024	0.93	22	0.5938	0.0593	0.10
11	0.0027	0.0025	0.94	23	0.8396	0.0633	0.08

Table 12: Mean layer-by-layer decomposition of the projection coefficient of variation over 15 experiments for Qwen2.5-0.5B. When paired with the graphs, we see that CV decreases because the mean grows faster than the SD, not necessarily because of less variability.

Layer	$\lvert\bar{s}\rvert$	sd	CV	Layer	$\lvert\bar{s}\rvert$	sd	CV
0	0.0003	0.0006	1.79	18	0.0025	0.0021	0.84
1	0.0004	0.0008	1.89	19	0.0029	0.0020	0.71
2	0.0006	0.0012	1.93	20	0.0117	0.0062	0.53
3	0.0008	0.0017	2.13	21	0.0312	0.0142	0.45
4	0.0010	0.0021	2.08	22	0.0451	0.0167	0.37
5	0.0013	0.0027	2.12	23	0.0906	0.0232	0.26
6	0.0014	0.0025	1.76	24	0.3697	0.0574	0.16
7	0.0016	0.0025	1.61	25	0.3919	0.0560	0.14
8	0.0017	0.0027	1.59	26	0.4182	0.0508	0.12
9	0.0019	0.0033	1.73	27	0.4185	0.0409	0.10
10	0.0019	0.0027	1.44	28	0.4495	0.0409	0.09
11	0.0020	0.0026	1.28	29	0.4619	0.0465	0.10
12	0.0020	0.0025	1.27	30	0.4780	0.0462	0.10
13	0.0020	0.0025	1.25	31	0.4973	0.0407	0.08
14	0.0020	0.0025	1.21	32	0.4873	0.0337	0.07
15	0.0021	0.0025	1.16	33	0.4350	0.0279	0.06
16	0.0021	0.0024	1.14	34	0.3789	0.0238	0.06
17	0.0023	0.0022	0.99	35	0.6843	0.0600	0.09

Table 13: Mean layer-by-layer decomposition of the projection coefficient of variation over 15 experiments for Qwen2.5-3B.

The directional alignment plots (Figure 7) show that the individual resume bias vectors have less variability than what was observed in Mistral-7B. We see that, like Mistral-7B, the average per-pair cosine with the mean bias direction increases with layer depth. However, a key difference is that the average per-pair cosine peaks much higher than Mistral-7B (0.820 for 0.5B, 0.981 for 3B, 0.718 for 7B, 0.725 for 14B). In the 0.5B and 7B models, we recorded the peak average per-pair cosine in the last layer. In the 3B and 14B models, we recorded the peak average per-pair cosine in layer 31 and 35, respectively. When we calculate the $\arccos$ of each of these peak average per-pair cosines, we see that a typical resume's bias vector differs from the mean direction vector by roughly 34.9° in 0.5B, 11.1° in 3B, 44.1° in 7B, and 43.5° in 14B. These are much lower than the 68.9° spread in Mistral-7B, suggesting that the Qwen2.5 resume pairs are much more tightly aligned.

The sign agreement plot (Figure 8) shows that most of the individual resume bias direction vectors in the Qwen2.5 models agree on a general direction. Like in Mistral-7B, sign agreement increases with layer depth, approaching 1.0 in all model sizes. This suggests that most individual bias vectors from resume pairs point into the same half-space as the mean bias direction vector. Combined with the directional alignment findings, we conclude that individual resume-pair vectors in Qwen2.5 both agree in general direction and are tightly distributed around the mean bias direction.

Our CV analysis of the Qwen2.5 models also shows that there is a more uniform injection across resumes. We found that CV decreased with layer depth, an analogous result to Mistral-7B.

Layer	$\lvert\bar{s}\rvert$	sd	CV	Layer	$\lvert\bar{s}\rvert$	sd	CV
0	0.0003	0.0005	1.60	14	0.0690	0.0292	0.42
1	0.0005	0.0013	2.32	15	0.0724	0.0280	0.39
2	0.0006	0.0010	1.65	16	0.0764	0.0267	0.35
3	0.0008	0.0012	1.54	17	0.0741	0.0226	0.31
4	0.0010	0.0013	1.30	18	0.1085	0.0576	0.53
5	0.0011	0.0018	1.55	19	0.1151	0.0576	0.50
6	0.0013	0.0015	1.21	20	0.1182	0.0578	0.49
7	0.0014	0.0015	1.06	21	0.1183	0.0558	0.47
8	0.0022	0.0016	0.76	22	0.2297	0.0824	0.36
9	0.0031	0.0025	0.81	23	0.2382	0.0657	0.28
10	0.0031	0.0030	0.97	24	0.2484	0.0540	0.22
11	0.0031	0.0030	0.99	25	0.2490	0.0489	0.20
12	0.0033	0.0027	0.82	26	0.4145	0.0542	0.13
13	0.0764	0.0469	0.61	27	0.6246	0.0708	0.11

Table 14: Mean layer-by-layer decomposition of the projection coefficient of variation over 15 experiments for Qwen2.5-7B.

Layer	$\lvert\bar{s}\rvert$	sd	CV	Layer	$\lvert\bar{s}\rvert$	sd	CV
0	0.0002	0.0004	1.74	24	0.0060	0.0023	0.37
1	0.0003	0.0007	2.26	25	0.0053	0.0019	0.35
2	0.0004	0.0009	2.43	26	0.0066	0.0024	0.37
3	0.0005	0.0010	1.95	27	0.0079	0.0032	0.40
4	0.0005	0.0009	1.73	28	0.0152	0.0063	0.41
5	0.0007	0.0014	1.91	29	0.0148	0.0060	0.40
6	0.0008	0.0015	1.83	30	0.0150	0.0063	0.42
7	0.0009	0.0013	1.48	31	0.0234	0.0093	0.40
8	0.0010	0.0014	1.43	32	0.0236	0.0095	0.40
9	0.0011	0.0013	1.24	33	0.0251	0.0101	0.40
10	0.0011	0.0012	1.04	34	0.0319	0.0094	0.30
11	0.0011	0.0013	1.09	35	0.0687	0.0149	0.22
12	0.0012	0.0014	1.11	36	0.1008	0.0177	0.18
13	0.0013	0.0013	1.00	37	0.0973	0.0166	0.17
14	0.0014	0.0013	0.92	38	0.1024	0.0194	0.19
15	0.0015	0.0017	1.16	39	0.1198	0.0272	0.23
16	0.0016	0.0018	1.12	40	0.1760	0.0285	0.16
17	0.0018	0.0016	0.90	41	0.1811	0.0285	0.16
18	0.0022	0.0019	0.84	42	0.2784	0.0366	0.13
19	0.0027	0.0021	0.78	43	0.3067	0.0391	0.13
20	0.0030	0.0023	0.75	44	0.3292	0.0411	0.12
21	0.0036	0.0021	0.59	45	0.3804	0.0430	0.11
22	0.0037	0.0018	0.50	46	0.3785	0.0411	0.11
23	0.0052	0.0021	0.41	47	0.5571	0.0650	0.12

Table 15: Mean layer-by-layer decomposition of the projection coefficient of variation over 15 experiments for Qwen2.5-14B.

In both cases, we also found that this decrease is due to the mean projection growing faster than the standard deviation. However, at the selected layers, the Qwen2.5 models reach a lower CV (0.04-0.13) than Mistral-7B (1.29). After normalizing by the residual norm, this shows a tighter per-resume spread within the Qwen2.5 models.

H Injection Strength Experiments

In Table 1, we see that strong injection collapsed the model's capability to chance (balanced accuracy $= 0.5$). To ensure this does not confound our findings with a bias-saturated model, we test whether the mitigation patterns in the main findings are invariant across injection strengths and techniques. We acknowledge that we did not run either experiment on both model families, and future work will ensure a more robust iteration.

Table 16 shows that the mitigation outcomes for behavioral bias observed in the main experiments remain invariant under weaker injection strengths. For these experiments, we varied the injection learning rate and repeated one iteration of our pipeline on Mistral-7B. We used learning rates across a $10\times$ range (5×10^{-5} down to 5×10^{-6}), with the mitigations held fixed. For every injection-strength case, the demographic disparity saturated ($\Delta \geq 0.998$), while balanced accuracy stayed at chance (0.5). However, LoRA and OFT continue to drive behavioral bias toward 0, whereas ablation does not. Notably, ablation at weaker injection strengths left a larger residual behavioral bias (0.93, 0.82) than at the strongest strength (0.36). This may be because the injection is too weak for ablation to remove bias effectively. From these results, we suggest that the mitigation outcomes are not confounded by the injection collapsing model capability. However, we framed these experiments as a pilot, whereas future work would execute a more robust version of this experiment.

Injection Strength	Biased model M_b		Post-mitigation Δ		
(Learning Rate)	Δ	Balanced Accuracy	LoRA	OFT	Ablation
5×10^{-5} (main, 3 ep) †	1.00	0.50	0.00	0.00	0.36
1×10^{-5} (1 ep)	1.00	0.50	5×10^{-5}	9×10^{-4}	0.93
5×10^{-6} (1 ep)	0.998	0.50	9×10^{-4}	1.4×10^{-2}	0.82

Table 16: Testing Mistral-7B's robustness to injection strength. Across a $10\times$ range of injection learning rates, the behavioral bias saturates ($\Delta \geq 0.998$) and balanced accuracy stays at chance (0.50). The mitigation outcome is invariant to the injection strength, with both LoRA and OFT driving $\Delta \to 0$ while ablation did not. † indicates numbers derived from the main study, which show a mean over 15 seeds. The two weaker injections are single seeds.

Tables 17 and 18 also show that a graded injection produces the mitigation pattern we observed in the main findings, while also improving model capability. Instead of adjusting parameters during training, we mix biased and benign policy labels. For each SFT target, it will receive the biased policy label with probability p and the benign policy label with probability $1 - p$. We then sweep p from 0 to 1 and have the model execute the same tasks as in the main experiments. Although the ablation outcome reduces demographic disparity more in these experiments than in the main findings, it still does not perform as well as LoRA and OFT, which drive behavioral bias to 0.000. For model capability, all three methods maintain high balanced accuracy. In contrast, at $p = 1$, model capability is substantially lower for ablation and OFT. The $p = 0.5$ results suggest that rerunning our study with a mixture of labels may yield better capability outcomes for ablation, and potentially for OFT as well. Similar to the weak injection experiment, these graded injections were pilot experiments, with future work focusing on reiterating more robust versions of them.

Model	Mix rate p	Injected M_b	Ablation	LoRA	OFT
Qwen2.5-0.5B	0.00	0.005±0.006	0.002±0.002	0.002±0.002	0.003±0.001
	0.25	0.248±0.012	0.023±0.009	0.001±0.001	0.005±0.002
	0.50	0.512±0.049	0.135±0.021	0.002±0.002	0.007±0.007
	0.75	0.767±0.039	0.225±0.061	0.002±0.000	0.003±0.002
	1.00	0.999±0.001	0.081±0.066	0.001±0.000	0.004±0.003
Qwen2.5-3B	0.00	0.005±0.004	0.005±0.008	0.007±0.002	0.002±0.001
	0.25	0.247±0.040	0.079±0.075	0.009±0.006	0.005±0.003
	0.50	0.578±0.050	0.103±0.024	0.004±0.003	0.001±0.001
	0.75	0.744±0.060	0.154±0.065	0.011±0.014	0.006±0.009
	1.00	1.000±0.000	0.375±0.196	0.014±0.006	0.001±0.001
Qwen2.5-7B	0.00	0.003±0.005	0.004±0.004	0.005±0.003	—
	0.50	0.593±0.069	0.123±0.010	0.009±0.005	—
	1.00	1.000±0.000	0.368±0.502	0.007±0.005	—

Table 17: *Demographic disparity* in a graded bias injection experiment (mean ± sample standard deviation). LoRA and OFT drive Δ to near zero at every strength, whereas ablation does not. As a result, the mitigation outcome we observed in the main study is held invariant. OFT within the graded bias injection experiments was not run at Qwen2.5-7B, as the fit trains 7.6×10^8 parameters (9.1% of the model) and did not complete within the pilot's compute budget. Qwen2.5-14B was not tested due to computational constraints.

Model	Mix rate p	Baseline B	Injected M_b	Ablation	LoRA	OFT
Qwen2.5-0.5B	0.00	0.53±0.00	0.89±0.02	0.88±0.03	0.92±0.02	0.91±0.04
	0.25	0.53±0.00	0.86±0.03	0.92±0.01	0.92±0.01	0.86±0.03
	0.50	0.53±0.00	0.71±0.04	0.83±0.03	0.93±0.02	0.83±0.15
	0.75	0.53±0.00	0.50±0.01	0.64±0.08	0.90±0.02	0.92±0.04
	1.00	0.53±0.00	0.50±0.00	0.49±0.01	0.82±0.08	0.77±0.13
Qwen2.5-3B	0.00	0.53±0.00	0.85±0.01	0.86±0.01	0.87±0.02	0.57±0.12
	0.25	0.53±0.00	0.85±0.02	0.84±0.08	0.92±0.01	0.72±0.20
	0.50	0.53±0.00	0.62±0.03	0.55±0.08	0.90±0.04	0.52±0.03
	0.75	0.53±0.00	0.49±0.00	0.53±0.04	0.88±0.03	0.56±0.11
	1.00	0.53±0.00	0.50±0.00	0.52±0.01	0.87±0.03	0.51±0.02
Qwen2.5-7B	0.00	0.56±0.00	0.89±0.01	0.89±0.01	0.89±0.02	—
	0.50	0.56±0.00	0.63±0.05	0.62±0.05	0.89±0.02	—
	1.00	0.56±0.00	0.50±0.00	0.50±0.00	0.83±0.06	—

Table 18: *Balanced accuracy* under the graded bias injection experiment (mean ± sample standard deviation). Capability decreases within the injected model as p increases across all models. LoRA consistently maintains high model capability across all models at all p, whereas ablation and OFT experience decreasing model capability as p increases.