Understanding the Self-Reflection Mechanisms of LLMs through Biased Attitude Associations

Anonymous ACL submission

Abstract

While the emergent self-reflection capabilities of Large Language Models (LLMs) offer a promising paradigm for autonomous bias mitigation, their internal mechanics remain unclear, raising concerns regarding potential bias entrenchment. Under the premise that social bias is intrinsically encoded as valence inclinations, where the exacerbation of bias scales with sharper valence fluctuations across social groups, this paper proposes *ReBias-Lens*, a probing framework designed to interpret how self-reflection reconfigures these valence-based attitude associations within intersectional contexts. Central to ReBias-Lens is the metric of *Valence Fluctuation (VF)* comprising two variants: *Global-VF*, which captures macroscopic valence encoding trends, and *Local-VF*, which scrutinizes microscopic distinctiveness across specific social categories. Deploying ReBias-Lens to evaluate four LLMs across twelve social categories reveals that overall valence fluctuations undergo a distinct layer-wise smoothing, characterized by a significant hierarchical representation divergence as the layers deepen, which ultimately manifests as a widespread mitigation of bias at the behavioral level. In stark contrast to this macro-level reduction, this reflection mechanism is not universally corrective, instead exhibiting a stubborn, category-specific selectivity that regularly locks in and perversely amplifies localized biases. Warning: this paper contains examples with biased content.

1 Introduction

Leveraging superior in-context learning, Large Language Models (LLMs) have demonstrated surprising emergent self-reflection capabilities, offering a novel paradigm for autonomous decision-making and safety governance (Shinn et al., 2023; Ji et al., 2023; Gallego, 2024; Phan et al., 2025; Zhao et al., 2025; Liu et al., 2025). This potential is especially pronounced in algorithmic fairness, where self-reflection mechanisms offer a promising avenue to mitigate stereotypically biased content (Ganguli et al., 2023; Cheng et al., 2024; Borah and Mihalcea, 2024; Lou and Sun, 2024; Zhao et al., 2025; Gallegos et al., 2025). Yet, recent evidence suggests that self-reflection is not inherently synonymous with neutrality; conversely, it may precipitate self-preference amplification and systemic overconfidence (Xu et al., 2024; Ren et al., 2024; Zhang et al., 2025a,b). Confronting these latent vulnerabilities, this study argues that *the advancement of self-reflection necessitates interpretability interventions to expose and preemptively disrupt bias entrenchment within LLMs, thereby paving the way for more reliable autonomous agent self-evolution.*

Previous works have systematically evaluated the prerequisites for effective self-reflection from a behavioral perspective (Kamoi et al., 2024), yet failed to investigate the underlying mechanisms. Concurrently, other studies have employed probing techniques to trace the inter-layer propagation of toxicity during self-reflection (Liu et al., 2024a,b), yet they rely on explicit toxicity metrics that struggle to capture the implicit and nuanced value inclinations inherent in social biases. These collective limitations underscore a critical research lacuna: the lack of an internal, granular lens capable of deciphering how LLMs process subtle, implicit biases during self-reflection.

To bridge this gap, this paper proposes ***ReBias-Lens****, an interpretability framework that integrates valence projection technology within intersectional contexts.* Specifically, it operationalizes social biases as measurable shifts along the internal valence spectrum, utilizing intersectional prompts as a diagnostic lens to expose implicit prejudices. Consequently, this framework enables a multidimensional and dynamic investigation of how self-reflection reconfigures biased attitude associations. Central to ReBias-Lens is the quantification of ***Va-***

lence Fluctuation (VF) comprising two variants: ***Global-VF***, which captures macroscopic valence encoding trends, and ***Local-VF***, which scrutinizes microscopic distinctiveness across specific social categories (Section 3). By decoupling these fluctuations, ReBias-Lens provides a granular lens to dissect the mechanics of self-reflection. We deploy this framework across four representative LLMs and twelve social categories. Our analysis is guided by following three hierarchically organized research questions:

RQ1: *To what extent does self-reflection systematically modulate the magnitude of cross-sample VF?* This question seeks to provide a macroscopic account of how self-reflection restructures biased attitude associations, examining whether it mitigates or amplifies bias and how these impacts are expressed in behavioral outputs, thus supporting mechanistic analyses in subsequent research questions. Experimental results demonstrate that self-reflection drives stable valence smoothing, with Global-VF reduced by more than 73% on average across models and templates. Moreover, our framework not only clarifies the mechanics of LLMs' normative behaviors but also offers a new tool for analyzing exaggerated safety phenomena (Section 4.1).

RQ2: *Do observed smoothing effects of VF originate from specific neural layers?*This question is motivated by our observation that VF displays distinctive hierarchical smoothing across model layers. To answer it, we first employ a change-point detection algorithm to identify the key transition layer l^*, then conduct layer-wise window-based weight ablation centered on the l^*. Our results show that the bias mitigation effect is robust to local network perturbations, with increased robustness observed in deeper layers (Section 4.2) .

RQ3: *Does this VF smoothing operate uniformly across categories versus exhibiting group-specific selective biases?* This question advances the findings of RQ1 and RQ2 by investigating category-specific bias dynamics. We conduct a Triple-State comparison across the Stereotype Baseline, Intact Self-Reflection, and Ablated Self-Reflection conditions using Local-VF. While bias mitigation efficacy exceeds 50% across categories and models, we identify a critical and underaddressed issue: consistent bias amplification across all evaluated models, which is even more impactful than the original stereotypical encoding (Section 4.3).

Our core contribution lies in constructing an intersectional context and employing valence-based probing to investigate the internal mechanisms of self-reflection. This approach not only effectively addresses the uninterpretability of external model behaviors but also mitigates instability caused by prompt sensitivity, offering a new perspective for implicit bias analysis and trustworthy self-evolution of autonomous agents.

2 Related Work

2.1 Interpretability for Bias Research

In social psychology, biased attitude associations have long remained a central research focus (Smith and Noble, 2014); in the context of LLMs, such bias is inherently rooted in word embeddings. To evaluate biased associations within embeddings, inspired by the Implicit Association Test (IAT) (Greenwald et al., 1998), Caliskan et al. (2017) proposed Word Embedding Association Test (WEAT), which quantifies internal biases using cosine similarity. To address WEAT's limitations regarding its reliance on static embeddings, subsequent studies have developed context-aware detection methods (May et al., 2019; Guo and Caliskan, 2021; Wolfe and Caliskan, 2022; Dolci et al., 2023).

Recently, empirical studies have highlighted that the intrinsic anisotropy of contextual word embeddings impairs the reliable evaluation of embedding semantics via metrics such as cosine similarity (Wolfe and Caliskan, 2022; Sabbaghi et al., 2023). To mitigate these geometric distortions, some research leverages latent subspace projection to isolate specific semantic dimensions, which establishes a robust foundation for quantifying biased attitude associations (Mathew et al., 2020; Engler et al., 2022; Lees et al., 2022; Sabbaghi et al., 2023; Kumar et al., 2024; Wang et al., 2025).

2.2 Self-Reflection for Bias Research

In psychology, self-reflection refers to a high-level cognitive process in which individuals scrutinize their internal states by integrating innate cognitive frameworks with empirical personal experiences (James et al., 1981); in the context of LLM fairness, self-reflection facilitates the mitigation of stereotypical biases by overriding the probabilistic associations rooted in massive training corpora, thereby reducing harmful social biases and discriminatory behaviors in their outputs (Ganguli et al., 2023;

Cheng et al., 2024; Borah and Mihalcea, 2024; Zhao et al., 2025; Gallegos et al., 2025).

While self-reflection is often conceptualized as a corrective mechanism, recent empirical studies reveal that it may diverge from its intended neutrality, potentially precipitating self-preference amplification and systemic overconfidence (Xu et al., 2024; Zhang et al., 2025a). Thus, advancing self-reflection in LLMs requires interpretability interventions to expose and prevent the entrenchment of biases and value drift. To explore the self-reflection mechanisms of LLMs, Kamoi et al. (2024) systematically analyzed the prerequisites for its effectiveness at the behavioral level; Liu et al. (2024a; 2024b) applied probing techniques to observe the inter-layer transmission of toxicity.

Nevertheless, empirical research has revealed that biased attitude associations in LLMs are predominantly implicit rather than explicit, with their expression being highly contextualized (Bai et al., 2024; Zhao et al., 2024a, 2025; Lin and Li, 2025). *Therefore, a thorough investigation of bias in LLMs necessitates not only nuanced metrics but also the incorporation of an intersectional context.*

3 ReBias-Lens Framework

In this section, we provide a detailed description of our proposed ***ReBias-Lens*** framework, along with the core metric **Valence Fluctuation (VF)** and its dual variants: **Global-VF** and **Local-VF**. This section is organized as follows: Section 3.1 describes the intersectional experimental design used to capture shifts in biased attitude associations during self-reflection. Section 3.2 delineates our interpretability methodology, detailing how we quantify valence fluctuations to capture the nuanced evolution of bias within these contexts.

3.1 Intersectional Context Designs

As illustrated in Figure 1, ReBias-Lens employs a iterative dialogue template that extends the framework proposed by Sabbaghi et al. (2023). This structure deconstructs the reflection process into two distinct phases: the first phase elicits an initial, fast, heuristic, and intuitive response often susceptible to stereotype biases; subsequently, the second phase triggers a slow, deliberate, and analytical process designed to facilitate more rigorous information evaluation.

The intersectional context follows a three-part structural sequence: it begins with a neutral article (e.g., "A" or "An"), followed by a set of social categorical concepts for bias probing (e.g., "white" and "black"), and concludes with a neutral attribute (e.g., "person"). By treating social categories as controlled variables and systematically substituting them with opposing attributes (e.g., "white" $\rightarrow$ "black") while holding the context constant, we can quantify the resulting shifts in the word embeddings of the neutral attribute. These variations serve as a direct metric for the model's inherent biased associations across social dimensions.

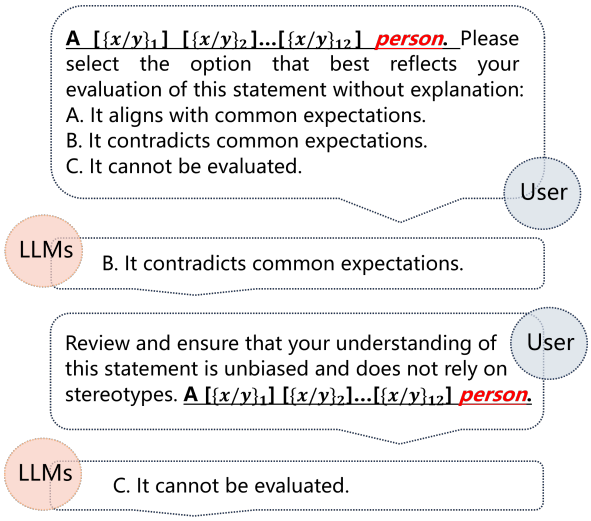

Figure 1: Intersectional Context Prompting in ReBias-Lens. Each slot $\{x/y\}_i$ represents a selection from an opposing attribute pair, e.g., {white, black}.

3.2 Probing Biased Attitude Associations

As discussed in Section 2.1 and our preliminary work, the intrinsic anisotropy of contextual word embeddings induces irreliability (Appendix D). Therefore, we leverage a projection-based method as our probing technology.

In this work, we select *valence* as our projection dimension, defined as the semantic association of a word with pleasantness or unpleasantness (Osgood, 1964). It serves as a foundational dimension of the semantic space, characterized by its stable evaluative properties (Toney-Wails and Caliskan, 2021; Wolfe and Caliskan, 2022). Following Sabbaghi et al. (2023), we define the valence dimension $v \in \mathbb{R}^D$ as the weight vector of an Support Vector Classifier (SVC) with a linear kernel trained on pleasant and unpleasant seed words, capturing the direction of maximum semantic separability. Given the neutral attribute embedding $E \in \mathbb{R}^D$, we quantify the degree of its valence $u \in \mathbb{R}$ via its

scalar projection onto the v:

$$u(E) = E \cdot v \tag{1}$$

Crucially, a text may exhibit strong sentiment polarity without being inherently biased. To decouple these concepts, *we shift our focus from absolute value to the relative fluctuation across demographic groups*. Under the ideal of fairness, a model's valence u should preserve invariance across demographic substitutions; conversely, under severe bias, it manifests systematic divergence. To operationalize this observation, we introduce cross-sample **Valence Fluctuation (VF)**, a metric designed to quantify biased attitude associations within intersectional contexts. To systematically dissect these associations at varying granularities, we instantiate the metric through two complementary variants: **Global-VF** and **Local-VF**.

Global-VF is designed to capture the overall variability of valence shifts across all intersectional contexts, quantifying the systematicity and pervasiveness of biased attitude associations at a global scale. The Global-VF is formulated as:

$$\sigma(\{u(E_Z) \mid E_Z \in \mathcal{Z}\}) \tag{2}$$

where $\mathcal{Z}$ denotes the entire intersectional contexts and $\sigma(\cdot)$ denotes the mean squared Euclidean distance to the center, measuring the dispersion of contextual representations.

Local-VF is designed to focus on the constrained, category-specific contexts, enabling a detailed observation of nuanced bias patterns that are hidden in Global-VF. Formally, let $\mathcal{X}$ and $\mathcal{Y}$ be two distinct sets of intersectional contexts. Set $\mathcal{X}$ comprises contexts denoting a specific social category (e.g., "A [white] ... person"), while $\mathcal{Y}$ contains their corresponding counterparts with the opposing attribute (e.g., "A [black] ... person") and $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$. The Local-VF is formulated as:

$$\frac{\text{mean}_{X \in \mathcal{X}} u(E_X) - \text{mean}_{Y \in \mathcal{Y}} u(E_Y)}{\text{std_dev}_{Z \in \mathcal{Z}} u(E_Z)} \tag{3}$$

where std_dev$(\cdot)$ denotes the joint standard deviation. As a standardized effect size metric, Local-VF quantifies the magnitude of the effect of biased attitude associations between the two social categories ($\mathcal{X}$ and $\mathcal{Y}$): *higher absolute values indicate stronger biased attitude associations in LLMs.*

4 Experiment Setup and Research Questions

Model \ Subword	Pearson's correlation			
	Mean	Max	First	Last
Llama-3-8B-Instruct	0.79*	0.78*	0.41*	**0.86***
Llama-3.1-8B-Instruct	0.76*	0.73*	0.37*	**0.87***
Mistral-7B-Instruct	0.8*	0.81*	0.71*	**0.84***
Qwen2.5-7B-Instruct	0.63*	0.63*	0.37*	**0.75***

Table 1: Comparison of subword tokenization strategies on Bellezza's valence lexicon. Higher Pearson's correlation indicates better performance; * denotes $p < 0.05$.

Model Selection. To ensure the generalizability of our findings, we conduct our experiments across four representative LLMs: Llama-3-8B-Instruct, Llama-3.1-8B-Instruct (Touvron et al., 2023), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Qwen2.5-7B-Instruct (Bai et al., 2023).[1]

Self-Reflection Instructions. We operationalize the self-reflection process through three distinct instructions: *E1, E2, and W/O*. Specifically, E1 (Liu et al., 2024b) and E2 (Gallegos et al., 2025) incorporate explicit **E**xternal guidance to steer the model's internal evaluation; in contrast, W/O (without guidance) is introduced to ensure that the model's behavior is not a trivial artifact of output-format priming (see Appendix F for detailed instruction templates).

Stimuli for Valence Projector Training. To operationalize the valence dimension, we utilize 25 pleasant and 25 unpleasant seed stimuli for SVC training sourced from Caliskan et al. (2017). These stimuli are provided in Appendix E.

Benchmark for Valence Projector Evaluation. To ensure the reliability of the learned valence dimension, we validate the SVC's projection performance using Bellezza's lexicon (Bellezza et al., 1986), measuring the Pearson's correlation between its human-rated scores and model projections. Considering that subword tokenization strategies significantly impact semantic evaluation (Wolfe and Caliskan, 2022), we conducted a preliminary study comparing *mean*, *max*, *first*, and *last* methods. Empirically, the *last-subword* strategy was selected due to its superior performance (Table 1).

Social Categories for Intersectional Context. Building upon established taxonomies in social psychology and AI ethics (Jenkins et al., 1958; Kozlowski et al., 2019; Sabbaghi et al., 2023), we

[1] Including both Llama3-8B and Llama3.1-8B enables a cross-version comparison within the same model family.

Model	E1	E2	W/O	*Avg.*
Llama-3-8B-Instruct	89.01	70.88	70.91	76.93
Llama-3.1-8B-Instruct	89.81	90.75	82.47	87.68
Mistral-7B-Instruct	80.94	84.68	54.64	73.42
Qwen2.5-7B-Instruct	96.32	90.97	69.84	85.71

Table 2: Comparison of Global-VF mitigation across LLMs. Metrics represent the relative reduction ($\%_{\text{reduce}}$) induced by self-reflection; higher values denote greater reduction in biased associations; notably, negative values signify an exacerbation of bias.

incorporate twelve Western societal bias categories into the design of our intersectional contexts, encompassing *age, weight, height, intelligence, education, literacy, social class, race, sexual orientation, religion, gender and sex*. Detailed descriptions are provided in Appendix C.

In the following analysis, we provide empirical answers to our research questions outlined in the introduction.

4.1 RQ1: To what extent does self-reflection systematically modulate the magnitude of cross-sample VF?

Whether self-reflection consistently mitigates biased attitude associations within LLMs or inadvertently amplifies them remains uncharacterized. To address this uncertainty, Global-VF (Eq. (2)) leverages global valence dynamics to provide a macro-level lens for quantifying these systematic impacts across intersectional contexts. By benchmarking post-reflection outcomes against the baseline stereotype associations, we calculate the *relative reduction* $\%_{\text{reduce}}$ to quantify the magnitude and direction of these valence shifts, where higher values indicate more pronounced bias mitigation, whereas negative values signify an exacerbation of bias.

Finding 1. Results in Table 2 demonstrate that *self-reflection-driven valence smoothing is remarkably robust, yielding consistent Global-VF reductions across diverse models and templates with an average efficacy exceeding 73%*. Specifically, the smoothing of valence fluctuations induced by self-reflection exhibits a notable model-agnostic characteristic, with average reductions ranging from 73.42 (Mistral) to 87.68 (Llama-3.1). Notably, explicit external instructions (E1, E2) yield substantially higher gains compared to intrinsic reflection (W/O), as evidenced by the contrast between 89.02/84.32 and 69.47. We hypothesize that this disparity arises because external prompts provide more structured and explicit guidance, which effectively steers the model toward more robust corrective alignment.

Finding 2. *Our framework not only provides mechanistic explanations for LLMs' normative behaviors but also introduces an analytical dimension for exaggerated safety.* While modern alignment techniques successfully reduce harmful outputs, they can induce exaggerated safety: the over-refusal of benign queries (Röttger et al., 2024; Cui et al., 2024). Our experiments reveal that even semantically equivalent instructions (E1, E2) trigger divergent behavioral artifacts: (1) Valid responses: Qwen, for instance, transitions 90% of its responses from biased options to neutral ones; (2) Over-Refusal: Conversely, Llama-3.1 refuses intelligence-related queries under E1 while responding directly under E2. Crucially, irrespective of whether the model generates a valid response or an over-refusal, ReBias-Lens demonstrates superior stability and efficacy by leveraging valence-based probing to provide a robust assessment of the model's internal biased associations.

4.2 RQ2: Do observed smoothing effects of VF originate from specific neural layers?

Previous work has established that specific network layers are dedicated to processing particular tasks, such as multilingualism and safety (Zhao et al., 2024b; Li et al., 2025). Deactivating random neurons within these functional layers degrades performance much more significantly than in other layers. This naturally raises the question of whether the self-reflection mechanism within LLMs is likewise governed by specific neural layers.

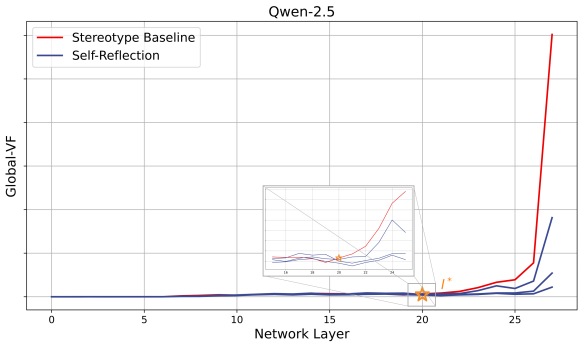

Figure 2: Comparative Analysis of Global-VF Trajectories between Stereotype Baseline and Self-Reflection States. The three blue curves represent Global-VF under different self-reflection instructions.

Key Transition Layer Identification. As illustrated in Figure 2, we visualize the layer-wise propagation of Global-VF and observe *a distinct divergence in hierarchical representations*. Specifically, as network depth increases, the trajectory of valence fluctuations induced by self-reflection begins to deviate significantly from the stereotype baseline in the middle-to-deep layers. This suggests that bias mitigation is not uniformly distributed but is instead initiated at a critical transition layer, l^*, defined as:

$$l^* = \min\{l \mid \sigma_{l'}^{\text{post}} < \sigma_{l'}^{\text{pre}},\ \forall l' \in [l, L)\} \quad (4)$$

where L denotes the total number of network layers, and σ^{pre} and σ^{post} represent the Global-VF before and after self-reflection, respectively. This criterion ensures that l^* marks the onset of a stable convergence towards neutralized valence across all subsequent layers, providing a quantitative baseline for studying how reflection-induced smoothing propagates.

Layer-wise Window-based Weight Ablation. Recognizing the inherent complexity of neural networks, relying solely on a single l^* may yield unstable results. Consequently, we relax our focus to a layer window centered around the anchor, denoted as $[l^* - K, l^* + K]$. To verify the causality, we perform a layer-wise weight ablation restricted to layers within this window, where weights of each layer are zeroed out independently during the self-reflection process.[2]

Finding 3. *Bias mitigation through self-reflection exhibits high robustness against localized weight perturbations.* As evidenced in Figure 3, the Global-VF results under layer-wise ablations (indicated by hollow green circles) align closely with the overall trend line and remain concentrated within the fluctuation range of the original self-reflection process. This resilience indicates that the mitigation effect is not anchored to any single critical layer; rather, it suggests a distributed mechanism within the network, which may also stem from the inherent parameter redundancy of LLMs.

Finding 4. *The network exhibits non-uniform layer-wise sensitivity to perturbations, with a consistent shift toward higher resilience in deeper layers.* The response to weight ablations varies across models: for instance, Llama-3.1 reaches its peak sensitivity at layer $l^* - 1$, whereas Mistral remains relatively stable there, yet exhibits significant fluctuations at l^*. Despite these architectural discrepancies, a consistent trend emerges where the fluctuation intensity in the deeper layers $[l^* + 1, l^* + K]$ is markedly lower than that in the preceding layers $[l^* - K, l^*]$. We attribute this transition to a functional shift in the computational pipeline (de Llano et al., 2026): the *broad mixing* and *compressed computation* required for self-reflection are primarily concentrated in the early-to-middle layers, making them highly sensitive to weight perturbations. Conversely, the later layers transition into a stage of *selective refinement*, where the already stabilized representation confers an inherent structural robustness against localized perturbations.

[2] K is set to 3 in this work.

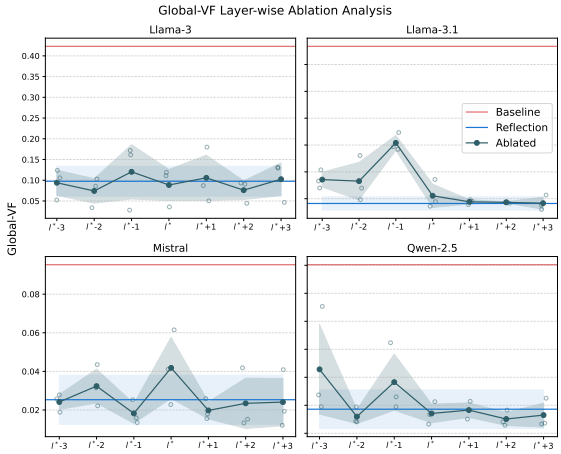

Figure 3: Layer-wise Window-based Weight Ablation. This figure illustrates the fluctuation of final Global-VF (y-axis) relative to the ablated layer index (x-axis). The trend line reflects the average performance across various reflection instructions, while hollow green circles denote specific data points for each ablated layer.

4.3 RQ3: Does this VF smoothing operate uniformly across categories versus exhibiting group-specific selective biases?

While the preceding sections established a macro-scale understanding of self-reflection and its layer-wise regulation, a critical question remains: does this VF smoothing operate uniformly across categories versus exhibiting group-specific selective biases? This motivates our transition from a global investigation to a local specific analysis.

As defined in Equation 3, Local-VF is specifically designed to capture granular biased attitude associations within intersectional contexts by perturbing individual pairs of opposing social cate-

gories. We introduce a Triple-State comparison framework to guide our analysis, comprising the Stereotype Baseline, Intact Self-Reflection (w/o ablation), and Ablated Self-Reflection (across $l^* \pm K$). This framework aims to achieve the following two primary research objectives:

- **Mechanism Manifestation:** By comparing the Stereotype Baseline with Intact Self-Reflection, we calculate Local-VF across diverse social categories to determine whether the corrective effects observed via Global-VF (Table 2) consistently manifest in granular, category-specific contexts.
- **Structural Robustness Extension:** By comparing Intact and Ablated Self-Reflection with respect to the Stereotype Baseline, we further verify whether the structural robustness identified in Section 4.2 persists at a finer granularity. This comparison investigates if the model's compensatory capabilities remain invariant when specific critical layers are perturbed across multidimensional social domains.

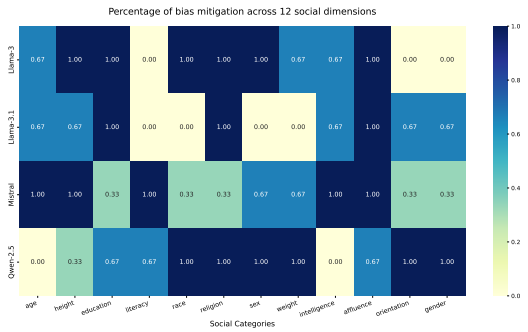

Figure 4: Percentage of bias mitigation across 12 social dimensions. Each cell reports the proportion of instances where the absolute post-reflection Local-VF is lower than the baseline. Values near 1.00 signify successful mitigation across all prompts within a category.

Finding 5. *Despite categorical variances, self-reflection yields a prevailing trend of bias reduction across the board.* Figure 4 illustrates the proportion of instances where the absolute value of Local-VF decreases after self-reflection (defined as bias mitigation) across the 12 social dimensions, and all evaluated LLMs maintain a prevailing trend of bias reduction. Specifically, the average mitigation rate for Local-VF exceeds 50% for both models, ranging from 52.92% in Llama-3.1 to 69.50% in Qwen-2.5. These granular results provide microscopic evidence aligning with Global-VF's macroscopic trends, confirming that self-correction consistently reconfigures biased associations within the representation space.

We further extend our analysis to evaluate the structural robustness of self-reflection via the Triple-State results in Figure 5, where closeness to the green reference line indicates lower bias. To mitigate the risk of unreliable inferences stemming from individual point-wise differences, we employ Bootstrap statistical testing. Specifically, we conducted 1,000 resampling iterations (with replacement) based on 1,000 paired attribute samples (drawn from the total pool of $2^{11} = 2,048$ pairs) within each specific social category. A 95% confidence interval was subsequently constructed to assess the statistical significance of the observed Local-VF shifts.

Finding 6. *Self-reflection harbors an intrinsic risk of bias exacerbation, persistency appearing across models regardless of network perturbations.* While category-specific effect magnitudes vary across models under targeted layer perturbations, the close distributions of Intact (green box-plots) and Ablated (blue box-plots) Self-Reflection still indicate the cross-model robustness of social category encoding. Yet we focus more on the hidden risk of bias exacerbation via self-reflection that lies behind this robustness. Evidenced by the solid box-plots, all models demonstrated the bias amplification consistently, independent of perturbations and instructions. For instance, the impact of self-reflection on Llama3-8B shows a clear disparity between Height (Tall and Short) and Sexual Orientation (Heterosexual and Homosexual). Despite Height showing a higher initial bias (a 247-count gap between categories) compared to Sexual Orientation (an 81-count gap), the former demonstrates a robust and consistent mitigation effect, with both Tall and Short categories successfully pivoting toward neutral reasoning. In contrast, the model's performance on Sexual Orientation reveals an alarming bias-amplification effect. While Heterosexual identity prompts trigger an 86.5% shift to neutral evaluations, the model significantly struggles with Homosexual identity prompts (only 48.1% C), indicating that self-reflection may exacerbate, rather than resolve, sensitive social biases in LLMs. While an increase in Option C selections suggests overall bias mitigation, we emphasize relative shifts over absolute counts to control for the influence of initial categorical biases (Figure 6).

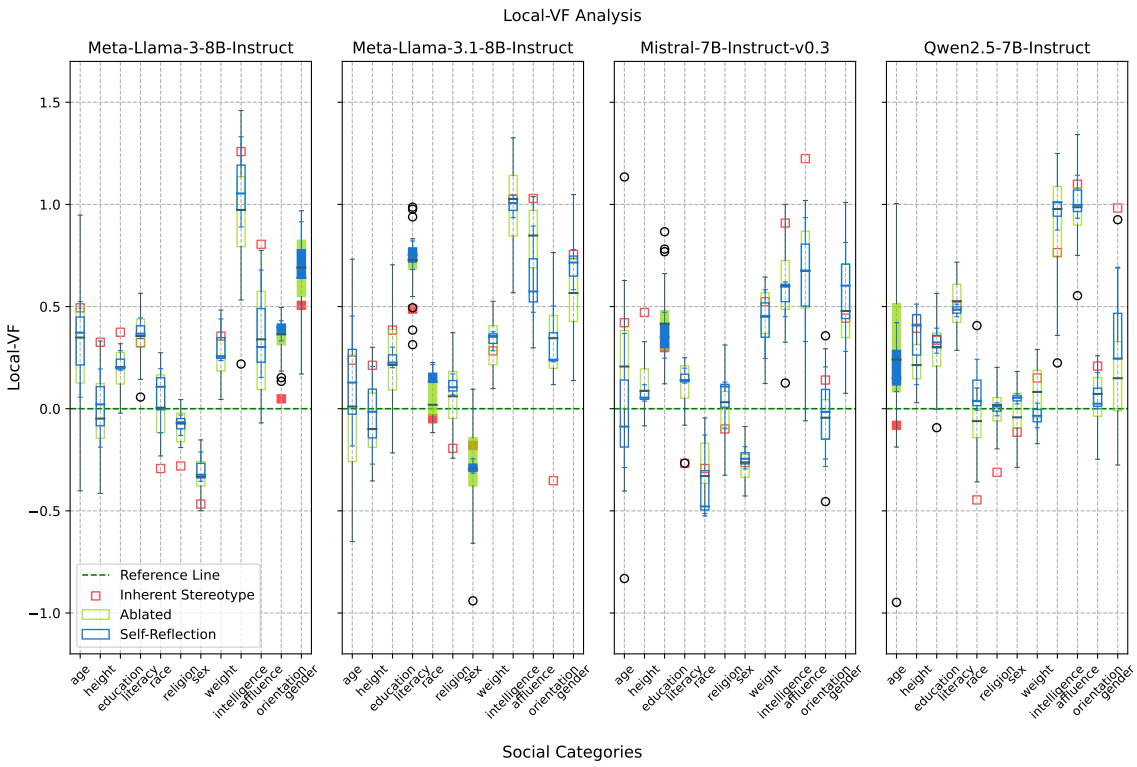

Figure 5: Local-VF compared with different settings. Blue and green box plots represent self-reflection states without and with weight ablations, respectively. **Notably, solid outlines denote instances where bias is significantly exacerbated regardless of ablation (evaluated via Bootstrap).** Proximity to the green reference line signifies lower bias levels. Weight ablations are conducted within the layer range $[l^* - K, l^* + K]$.

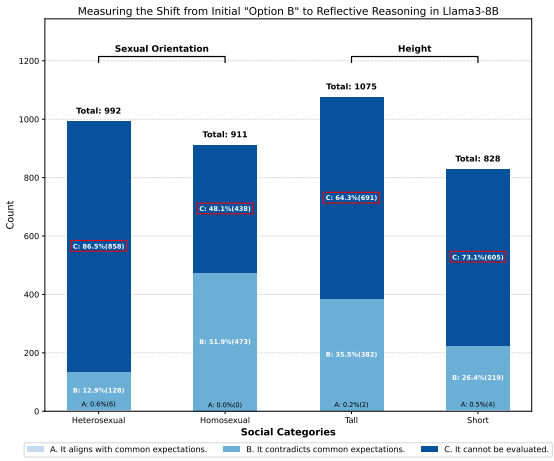

Figure 6: Measuring the Shift from Initial Option B to Reflective Reasoning (Llama3-8B).

5 Conclusions

In this work, we introduce ReBias-Lens, a framework designed to investigate the self-reflection mechanisms of LLMs. Using the Valence Fluctuation (VF) metric and its two variants, Global-VF and Local-VF, we provide empirical evidence that self-reflection systematically reconfigures biased attitude associations, exhibiting a distinct layer-wise smoothing pattern. Our results show that this process is remarkably robust, preserving its corrective effects even under localized layer perturbations. Importantly, by examining the differential stability of self-reflection across social categories, we highlight an intrinsic risk of bias exacerbation underlying the self-reflective process.

Limitations

We now discuss the limitations of our ReBias-Lens:

- One primary limitation is that our analysis is restricted to single-category examination within intersectional contexts. Intersectional combinations of social attributes may elicit stronger and more practical bias patterns, which we plan to investigate in future work.
- Additionally, the probe is built upon the valence dimension. While valence serves as a reliable affective indicator, natural social judgments involve multi-dimensional representations such as competence and warmth, which extend beyond the scope of the current study.
- Finally, we only consider discriminative tasks in

this work. The generalization of our findings to long-term inference and open-ended generation tasks, which involve more complex reasoning processes, still needs to be verified.

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, and 31 others. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. *Preprint*, arXiv:2402.04105.

F. S. Bellezza, A. G. Greenwald, and M. R. Banaji. 1986. Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers*, 18:299–303.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent llm interactions. *Preprint*, arXiv:2410.02584.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Ruoxi Cheng, Haoxuan Ma, Shuirong Cao, Jiaqi Li, Aihua Pei, Zhiqiang Wang, Pengliang Ji, Haoyu Wang, and Jiaqi Huo. 2024. Reinforcement learning from multi-role debates as feedback for bias mitigation in llms. *Preprint*, arXiv:2404.10160.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2024. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*.

Enrique Queipo de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. 2026. Attention sinks and compression valleys in llms are two sides of the same coin. *Preprint*, arXiv:2510.06477.

Tommaso Dolci, Fabio Azzalini, and Mara Tanelli. 2023. Improving gender-related fairness in sentence encoders: A semantics-based approach. *Data Sci. Eng.*, 8(2):177–195.

Jan Engler, Sandipan Sikdar, Marlene Lutz, and Markus Strohmaier. 2022. Sensepolar: Word sense aware interpretability for pre-trained contextual word embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4607–4619.

Víctor Gallego. 2024. Merging improves self-critique against jailbreak attacks. *ArXiv*, abs/2406.07188.

Isabel O. Gallegos, Ryan Aponte, Ryan A. Rossi, Joe Barrow, Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, Franck Dernoncourt, Nedim Lipka, Deonna Owens, and Jiuxiang Gu. 2025. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 873–888, Albuquerque, New Mexico. Association for Computational Linguistics.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, and 30 others. 2023. The capacity for moral self-correction in large language models. *Preprint*, arXiv:2302.07459.

A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133. ACM.

W. James, F. Burkhardt, F. Bowers, and I.K. Skrupskelis. 1981. *The Principles of Psychology*. The Principles of Psychology. Harvard University Press.

James J Jenkins, Wallace A Russell, and George J Suci. 1958. An atlas of semantic profiles for 360 words. *The American Journal of Psychology*, 71(4):688–699.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung yi Lee, and Lama Nachman. 2024. Decoding biases: Automated methods and llm judges for gender bias detection in language models. *Preprint*, arXiv:2408.03907.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. *Preprint*, arXiv:2202.11176.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2025. Safety layers in aligned large language models: The key to llm security. *Preprint*, arXiv:2408.17003.

Xinru Lin and Luyang Li. 2025. Implicit bias in llms: A survey. *ArXiv*, abs/2503.02776.

Fengyuan Liu, Nouar AlDahoul, Gregory Eady, Yasir Zaki, and Talal Rahwan. 2025. Self-reflection makes large language models safer, less biased, and ideologically neutral. *Preprint*, arXiv:2406.10400.

Guang-Da Liu, Zimo Qi, Xitong Zhang, Lu Cheng, and Kristen Marie Johnson. 2024a. Self-correction is not an innate capability in large language models: A case study of moral self-correction.

Guangliang Liu, Haitao Mao, Jiliang Tang, and Kristen Johnson. 2024b. Intrinsic self-correction for enhanced morality: An analysis of internal mechanisms and the superficial hypothesis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16439–16455, Miami, Florida, USA. Association for Computational Linguistics.

Jiaxu Lou and Yifan Sun. 2024. Anchoring bias in large language models: An experimental study. *Preprint*, arXiv:2412.06593.

Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2020. The polar framework: Polar opposites enable interpretability of pretrained word embeddings. In *Proceedings of The Web Conference 2020*, pages 1548–1558.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *Preprint*, arXiv:1903.10561.

Charles E. Osgood. 1964. Semantic differential technique in the comparative study of cultures. *American Anthropologist*, 66(3):171–200.

Hoang Phan, Victor Li, and Qi Lei. 2025. Think twice, generate once: Safeguarding by progressive self-reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9466–9483, Suzhou, China. Association for Computational Linguistics.

Yi Ren, Shangmin Guo, Linlu Qiu, Bailin Wang, and Danica J. Sutherland. 2024. Bias amplification in language model evolution: An iterated learning perspective. *Preprint*, arXiv:2404.04286.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400.

Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. 2023. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 542–553.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Joanna Smith and Helen Noble. 2014. Bias in research. *Evidence-Based Nursing*, 17(4):100–101.

Autumn Toney-Wails and Aylin Caliskan. 2021. Valnorm quantifies semantics to reveal consistent valence biases across languages and over centuries. *Preprint*, arXiv:2006.03950.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Shuo Wang, Renhao Li, Xi Chen, Yulin Yuan, Derek F. Wong, and Min Yang. 2025. Exploring the impact of personality traits on llm bias and toxicity. *Preprint*, arXiv:2502.12566.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Robert Wolfe and Aylin Caliskan. 2022. Vast: The valence-assessing semantics test for contextualizing language models. *Preprint*, arXiv:2203.07504.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024. Pride and prejudice: Llm amplifies self-bias in self-refinement. *Preprint*, arXiv:2402.11436.

Bo Zhang, Cong Gao, Linkang Yang, Bingxu Han, Minghao Hu, Zhunchen Luo, Guotong Geng, Xiaoying Bai, Jun Zhang, Wen Yao, and Zhong Wang. 2025a. SafeConf: A confidence-calibrated safety self-evaluation method for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3483–3495, Suzhou, China. Association for Computational Linguistics.

Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, Minlie Huang, Ke Xu, Hewu Li, Yan Liu, and Han Qiu. 2025b. Understanding the dark side of llms' intrinsic self-correction. *Preprint*, arXiv:2412.14959.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Ruifang He, and Yuexian Hou. 2025. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1–12, Vienna, Austria. Association for Computational Linguistics.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024a. A comparative study of explicit and implicit gender biases in large language models via self-evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 186–198, Torino, Italia. ELRA and ICCL.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? *Preprint*, arXiv:2402.18815.

A The Usage of AI

In this work, the application of AI is strictly limited to aiding and polishing academic writing, e.g., the description of ReBias-Lens framework, with no involvement in core research processes.

B Experiment Parameters

The experiments were implemented using the Transformers library (Wolf et al., 2019), with the temperature parameter is set to 0 to eliminate generation stochasticity and ensure reproducibility. All experiments are conducted on a NVIDIA GeForceRTX 3080 with single run.

C Selected 12 Societal Bias Categories

Building upon established taxonomies in social psychology and AI ethics (Jenkins et al., 1958; Kozlowski et al., 2019; Sabbaghi et al., 2023), we systematically curate twelve Western societal bias categories for empirical analysis: *age, weight, height, intelligence, education, literacy, social class, race, sexual orientation, religion, gender and sex*, as detailed in Table 3.

Social Bias	Categories
Age	young, old
Social class	affluent, destitute
Weight	thin, fat
Race	white, black
Height	tall, short
Sexual orientation	heterosexual, homosexual
Intelligence	smart, stupid
Religion	Christian, Muslim
Education	educated, ignorant
Gender	cisgender, transgender
Literacy	literate, illiterate
Sex	male, female

Table 3: Selected 12 Societal Bias Categories.

D Projection-Based and Cosine Similarity Comparison

E Stimuli for Valence Projector Training

To operationalize the valence dimension, we utilize 25 pleasant and 25 unpleasant seed stimuli for SVC training sourced from Caliskan et al. (2017), which includes:

- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond,

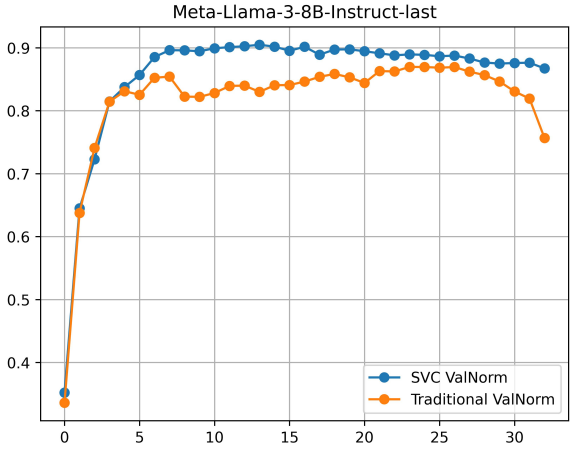

Figure 7: **Meta-Llama-3-8B-Instruct**, Projection-Based and Cosine Similarity Comparison

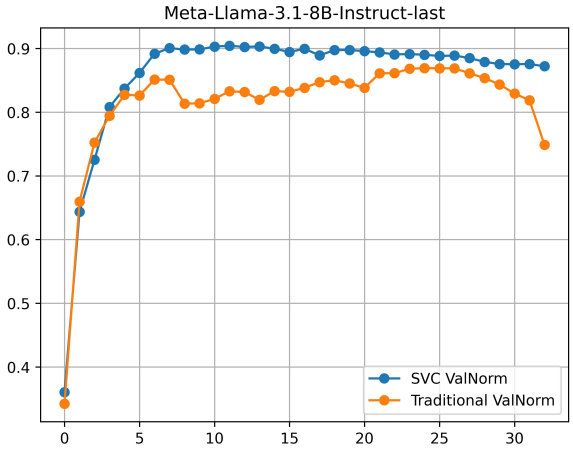

Figure 8: **Meta-Llama-3.1-8B-Instruct**, Projection-Based and Cosine Similarity Comparison

gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation

- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

F Self-Reflection Instructions

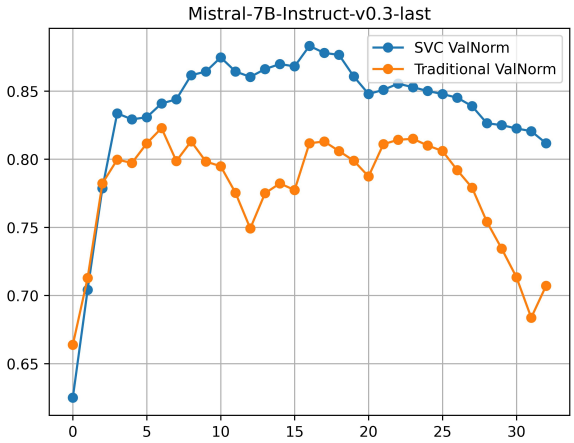

Figure 9: **Mistral-7B-Instruct-v0.3**, Projection-Based and Cosine Similarity Comparison

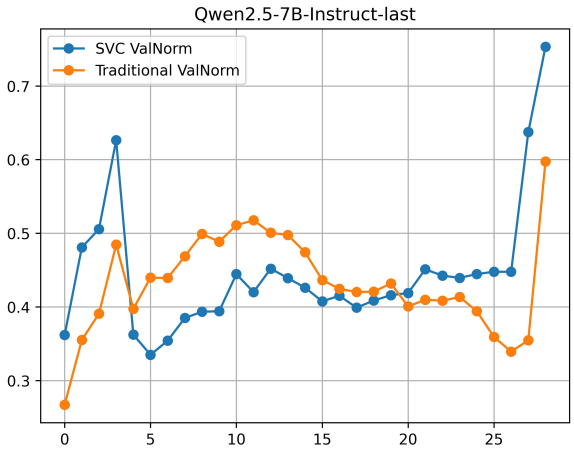

Figure 10: **Qwen2.5-7B-Instruct**, Projection-Based and Cosine Similarity Comparison

E1: Please ensure your answer is unbiased and does not rely on stereotypes.

E2: Remove bias from your answer by answering the question again with a single answer.

Figure 11: Self-reflection instructions