Towards Reliable AI Fairness: Challenges in Steering Features within Bias-Implicated Neurons

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago
CEATIC, Universidad de Jaén,
Campus Las Lagunillas, 23071, Jaén, Spain
{igmunoz, dofer, amontejo}@ujaen.es

Abstract

LLMs perpetuate societal biases, such as gender stereotypes, reinforcing harmful norms and posing significant fairness risks in real-world applications. We investigate a fine-grained mitigation technique that moves beyond surface-level fixes. Our approach uses attribution graphs to identify and directly steer bias-implicated features within a Sparse Autoencoder's (SAE) latent space. This method, known as feature steering, offers a theoretically precise, surgical intervention aimed at correcting bias at its neural source without costly retraining. We critically examine its practical reliability across various contexts. We find that steering effectiveness is highly sensitive to parameter tuning, often requiring unpredictable, context-specific adjustments. The intervention's success exists in narrow "sweet spots," outside of which performance can degrade catastrophically. This demonstrates that while direct intervention on learned features is a powerful analytical tool, significant challenges of brittleness and instability hinder its application as a consistent, broad-scale debiasing solution, necessitating research into more robust control mechanisms.

Keywords: bias, steering, gender, LLMs

1. Introduction

As Large Language Models (LLMs) are integrated into search engines, content creation, and professional tools, their ethical dimensions come under increasing scrutiny. A critical flaw stems from their training data; the practice of using massive, uncurated web data is inherently problematic, as such datasets overrepresent hegemonic viewpoints and encode biases damaging to marginalized populations (Bender et al., 2021). By learning the probability distributions within this text, deep NLP models inevitably inherit these societal flaws, a problem that is often exacerbated as models and their training corpora grow larger (Garrido-Muñoz et al., 2021). Consequently, models internalize social biases, such as gender stereotypes, which arise from skewed historical representations in the data (Bilal et al., 2025). The result is a persistent tendency for LLMs to not only learn and perpetuate but also actively amplify these harmful societal biases (Gallegos et al., 2024). This can lead to the reinforcement of harmful stereotypes, undermining progress toward a more equitable world.

This paper focuses on a particularly pervasive example of this phenomenon: gender bias in professional roles. LLMs frequently exhibit a strong, implicit association between certain professions and specific genders, most notably defaulting to "doctor" as male and "nurse" as female, a stereotype that persists even in modern architectures (Kotek et al., 2023). When prompted to generate text about a female doctor or a male nurse, a model may produce outputs that betray surprise, focus on the gender itself, or describe the individual using stereotypical language. Such biased outputs are not merely minor inaccuracies; they perpetuate limiting social norms, subtly influence human perceptions, and can have tangible discriminatory consequences in applications like automated recruiting or educational software (Gallegos et al., 2024; Sheng et al., 2021).

Efforts to mitigate bias in LLMs have traditionally fallen into several categories. Data-centric approaches aim to rebalance or curate training datasets, a task that is often prohibitively complex and may not remove implicit associations, sometimes merely hiding the bias rather than resolving it (Gonen and Goldberg, 2019). Other methods involve fine-tuning or retraining entire models, which is computationally expensive and risks degrading overall model performance through issues like catastrophic forgetting (Xu et al., 2025). Finally, post-processing techniques that filter outputs act as a surface-level fix, failing to address the underlying biased representations within the model itself (Gallegos et al., 2024). These approaches often lack the precision to correct the specific internal mechanisms that give rise to the bias.

To address these limitations, we investigate a fine-grained intervention at the neural level. Our methodology provides a surgical tool to both diagnose and attempt to correct bias where it originates. We first utilize techniques from mechanistic interpretability to analyze the model's internal representations (Templeton et al., 2024). Specifically, we employ circuit tracing to generate attribution graphs that map the model's computational pathways for

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3851–3860
11–16 May 2026.

given inputs (Ameisen et al., 2025). By analyzing and finding the common neurons across multiple graphs generated, we identify the global set of neurons causally linked to stereotypical outputs (Chandna et al., 2025). We then apply a targeted feature steering mechanism—a form of activation engineering—that adds a corrective vector to these neurons during inference, all without the need for costly retraining (Turner et al., 2023; Rimsky et al., 2023). However, instead of presenting this as a definitive solution, we use this powerful technique to ask a critical question: **Is simple neuron steering a reliable method for bias mitigation?**

This work makes the following contributions:

- We present a methodology for using comparative attribution graphs to identify and aggregate the neurons that form a **global bias circuit** for gendered concepts in LLMs. We define global bias circuit as a set of sparsely activating features that are causally implicated in producing a specific bias.
- We conduct an evaluation of neuron steering across four distinct experimental contexts, ranging from sterile, templated prompts to realistic, "in-the-wild" sentences, to rigorously test its robustness.
- We provide crucial empirical evidence demonstrating that while neuron steering can be effective in narrow, controlled settings, its success is **highly brittle and unreliable**. We show that its effectiveness is acutely sensitive to parameter tuning and degrades catastrophically with increasing linguistic complexity, highlighting a fundamental challenge for its practical application in ensuring AI fairness.

2. State of the Art: A Shift Towards Internal Control

Research into mitigating bias in LLMs has pivoted from external fixes to direct internal intervention. This shift is a response to the challenge that LLMs often perpetuate societal biases from their training data, leading to tangible **representational harms** like reinforcing stereotypes and **allocational harms** such as discriminatory outcomes (Bender et al., 2021; Gallegos et al., 2024). While the issue is multifaceted, our work focuses specifically on **binary gender bias (male/female)**, given its prevalence. Initial strategies like data pre-processing and output filtering proved limited, failing to address the root causes of bias within the model's representations (Karvonen and Marks, 2025; Hegazy et al., 2025). This motivated a turn inward, where interpretability is no longer just a desirable feature but a prerequisite for building fair and reliable AI (Bilal et al., 2025; Garrido-Muñoz et al., 2021).

Recent work focuses on two areas: training-time adaptation and inference-time intervention. The first involves **Parameter-Efficient Fine-Tuning (PEFT)** methods like **Low-Rank Adaptation (LoRA)**, which offer a computationally cheaper way to create a new, debiased model version by training small, adaptive modules (Hu et al., 2022; Ding et al., 2024). Our research, however, explores the second path: dynamic, **inference-time interventions** that modify the behavior of the original, unmodified model on-the-fly.

This focus on run-time control first required methods to locate conceptual representations within the network. The field evolved from using attribution to pinpoint individual, often polysemantic, "gender neurons" toward automated circuit discovery, which maps the complete computational subgraphs responsible for specific behaviors (Palikhe et al., 2025; Adam Zewe | MIT News; Ameisen et al., 2025; O'Brien et al., 2024). The ability to "read" these circuits laid the groundwork for "controlling" them, a paradigm formalized as **Representation Engineering (RepE)** (Zou et al., 2023; Wehner et al., 2025). The primary control method under RepE is **activation steering**, an inference-time technique that adds a "steering vector"—derived from the activation difference between contrastive prompts—to a model's forward pass, avoiding the high costs and inflexibility of permanent weight editing (Suri et al., 2025; Hegazy et al., 2025; Scialanga et al., 2025).

To overcome these challenges, the research frontier has pushed towards more robust controls. One path is **dynamic steering**, where methods like Conditional Activation Steering (CAST) make interventions context-aware (Hegazy et al., 2025). A second, highly promising direction operates in the disentangled feature space created by Sparse Autoencoders (SAEs). Techniques like Sparse Activation Steering (SAS) offer more precise control with fewer side effects by targeting monosemantic features instead of entire neuron activations (O'Brien et al., 2024; Bayat et al., 2025; Chalnev et al., 2024).

Despite the development of these methods, the premise that simple **feature steering** is fundamentally unreliable across diverse contexts is often assumed rather than rigorously demonstrated. This study provides that missing empirical evidence. We conduct a systematic investigation into the parameter sensitivity and contextual brittleness of the core steering mechanism, directly validating the field's turn towards more complex solutions. Our findings underscore the critical need for reliability in a world where targeted interventions, rather than universal fairness, may be the most practical path forward (Anthis et al., 2024).

3. Methodology

Our methodology is an empirical process designed to test the reliability of feature steering as a bias mitigation technique. We first establish a multi-faceted experimental setup to probe gender bias across a spectrum of contextual richness. We then use a comparative attribution technique to identify the global neural circuits encoding these biases. Finally, we apply targeted activation steering to these circuits and systematically evaluate the intervention's effectiveness and consistency.

3.1. Experimental Design: Probing Bias across a Spectrum of Contexts

All experiments were conducted in English using Google's `gemma-2-2b` (Gemma Team, 2024), an open-weight model chosen for its strong performance and available in Neuronpedia's API (Lin, 2023). It is possible to run the experiment, given that as a part of the GemmaScope release (Lieberum et al., 2024), a transcoder model was also released for gemma-2b along a set of public Sparse Autoencoders. These models are trained to reconstruct the output of a Multi-Layer Perceptron (MLP) sublayer from the activations at its input. The goal is to represent the complex, high-dimensional function of the MLP layer as a sparse linear combination of interpretable features (e.g., a feature that activates specifically for medical concepts, or another that activates for female pronouns), providing a powerful framework for understanding and manipulating the model's internal computations (Lieberum et al., 2024). By using these, we can isolate and study the specific features the model uses to represent concepts like gender.

To evaluate our intervention's robustness, we designed four distinct experiments that probe gender bias across a spectrum of contextual complexity. We begin with two broad assessments: one targeting general **professional roles** using diverse, naturalistic sentences, and another examining more subtle, character-based associations through **stereotypical adjectives**. We then conduct a focused case study on the pervasive **doctor/nurse stereotype** under two contrasting conditions. First, we establish a controlled baseline using highly-structured, templated sentences (Focused Context). We then test the intervention's resilience against this same stereotype using unique, "in-the-wild" sentences that introduce realistic linguistic complexity (Realistic Context). This progression from broad to specific, and from sterile to complex, allows for a systematic assessment of the steering method's effectiveness and limitations.

3.2. Identifying Bias-Relevant Features via Global Attribution Scoring

To identify the features responsible for gender bias, our pipeline moves from raw attribution data to a globally-ranked list of causally implicated features. Instead of analyzing circuits for individual prompts, our goal is to find a stable **global bias circuit** that represents the model's systematic tendency across a wide range of sentences.

The process for identifying this circuit is as follows: First, for every feature in the model's relevant MLP layers, we compute its mean attribution weight across all male-context prompts ($\bar{a}_{\text{male}}$) and its mean attribution weight across all female-context prompts ($\bar{a}_{\text{female}}$). These two values represent the feature's average contribution to outputs in each gendered context.

The central challenge is then to score each feature's importance in differentiating between these two contexts. To this end, we systematically evaluate distinct mathematical strategies, each representing a different hypothesis about how the bias circuit operates. Each strategy takes $\bar{a}_{\text{male}}$ and $\bar{a}_{\text{female}}$ as input and produces a single importance score, S_n, for that feature. By applying a strategy to all features and ranking them by their score, we produce a candidate global bias circuit. The strategies are summarized in Table 1.

The first strategy is **Absolute Difference**, a simple baseline that scores features based on the absolute difference between their mean attribution to each concept. A more nuanced approach is the **Weighted Difference**, which amplifies this raw difference by the neuron's total activation magnitude, prioritizing neurons that are both highly differentiating and highly active. The third strategy, **Bidirectional Specificity**, aims to find "specialist" neurons by explicitly penalizing dual activation. It computes a score based on the neuron's dominant mean attribution, $\text{dom}(\bar{a})$, while subtracting a scaled value of its opposing mean attribution, $\text{opp}(\bar{a})$, controlled by a penalty factor, λ. Finally, the **Bhattacharyya Specificity** (Bhattacharyya, 1946) strategy treats attributions as probability distributions. A neuron's positive attribution weight is normalized to a probability, $p(n)$, and its importance is inversely related to its contribution, $c(n)$, to the overlap between the male- and female-biased distributions. Neurons that contribute least to this overlap are considered the most specific.

The output of this pipeline is four distinct, ranked lists of **features**. Each list serves as a candidate bias circuit for our targeted intervention experiments, with the strategies summarized in Table 1.

Strategy	Core Intuition	Mathematical Formulation
Absolute Difference	Ranks by raw attribution difference.	$S_n = \lvert\bar{a}_{\text{male}} - \bar{a}_{\text{female}}\rvert$
Weighted Difference	Ranks by difference, amplified by total activity.	$S_n = (\bar{a}_{\text{male}} - \bar{a}_{\text{female}}) \cdot (\lvert\bar{a}_{\text{male}}\rvert + \lvert\bar{a}_{\text{female}}\rvert)$
Bidirectional Specificity	Rewards specialist features by penalizing dual activation.	$S_n = \text{sign}(\Delta\bar{a}) \cdot (\lvert\bar{a}_{\text{dom}}\rvert - \lambda \cdot \max(0, \bar{a}_{\text{opp}}))$ $\bar{a}_{\text{dom}}$ is the dominant mean attribution; $\bar{a}_{\text{opp}}$ is the opposing mean attribution.
Bhattacharyya Specificity	Ranks by contribution to distribution separability.	Rank by ascending $c(n) = \sqrt{p_{\text{male}}(n) \cdot p_{\text{female}}(n)}$ where $p_{\text{ctx}}(n) = \frac{\max(0, \bar{a}_{\text{ctx}}(n))}{\sum_k \max(0, \bar{a}_{\text{ctx}}(k))}$

Table 1: Feature Scoring Strategies for Bias Circuit Identification

3.3. Targeted Intervention and Evaluation

The final stage uses activation steering to intervene on the identified bias circuit and systematically evaluates the intervention's reliability.

Applying the Intervention For the intervention, we employ activation steering by instructing the Neuronpedia API to modify the model's internal states during inference. The underlying technique targets the monosemantic **features** identified by the SAEs. Our method involves providing the API with a list of the top-N features from the identified bias circuit, each paired with a calculated steering strength.

A targeted perform modification is performed, for each feature i in our request, its corresponding decoder weight vector ($\vec{v}_i = W_{dec,i}$) is retrieved. This vector is then scaled by its specified strength, which is a product of the feature's specificity score (S_i) and a global multiplier (α). The API aggregates these modifications and adds the resulting composite vector to the model's original activations ($\vec{a}_{\text{original}}$). The complete transformation can be expressed as:

$$\vec{a}_{\text{new}} = \vec{a}_{\text{original}} + \alpha \cdot \sum_{i=1}^{N} S_i \cdot \vec{v}_i$$

A central goal of this paper is to test the practical reliability of this technique. We therefore systematically vary the two key hyperparameters that define our instructions to the API:

1. **The number of features (N):** The number of top-scoring features included in the steering request.
2. **The steering multiplier (α):** A scalar coefficient that uniformly scales the overall strength of the intervention. Its sign determines the steering direction (e.g., pushing away from the biased concept).

Evaluating Intervention Success To quantify the outcome of each steering attempt, we classify the generated text into one of four categories. This allows for a nuanced view of the intervention's impact, distinguishing between precise correction, simple suppression, and outright failure.

- **Bias Flipped:** The intervention successfully guides the model to produce a specific, counter-stereotypical token, representing a precise and controlled correction.
- **Bias Persists:** The intervention fails, and the model still produces the original stereotypical token.
- **Bias Neutralized:** The model avoids both the stereotypical and the desired counter-stereotypical tokens, instead generating a different completion. This outcome indicates that the bias has been *suppressed*, representing lack of precise control.
- **Conflicting:** The intervention destabilizes the model's generative process, resulting in an incoherent output that may contain tokens associated with both stereotypes.

4. Results

Our experiments reveal that the effectiveness of activation steering is not uniform but highly variable, depending critically on the intervention parameters and, most significantly, the contextual nature of the prompt. While steering can be remarkably effective under controlled conditions, its reliability diminishes rapidly as linguistic complexity increases. In this section, we first evaluate whether the choice of neuron identification strategy significantly alters these outcomes, before analyzing the broader phenomena of parameter sensitivity and contextual brittleness.

4.1. Comparative Efficacy of Feature Identification Strategies

A primary goal of our expanded methodology was to determine if the choice among our four feature identification strategies had a meaningful impact on steering outcomes. To investigate this, we performed a Chi-squared (χ^2) test for each experiment and parameter combination, comparing the distribution of outcomes (i.e., counts of flipped, neutralized, or persisting bias) across the four strategies. The results are summarized in Table 2, which presents the p-value for each test, with statistically significant differences ($p < 0.05$) highlighted in bold.

Our findings show that for the vast majority of conditions, the choice of strategy is **not statistically significant**. Across most experiments, and particularly for moderate steering multipliers ($\alpha = 1$ and $\alpha = 10$), the p-values are consistently high, indicating that all four strategies perform comparably. Significant differences only emerge at the most

extreme steering intensity ($\alpha = 50$), an intentionally high value designed to force the model's behavior. These instances were limited to just two of the four experimental contexts: the **Adjective Bias** (male context) and, most prominently, the sterile, templated **Focused Context** experiment.

	Professional Bias			Adjective Bias		
Multiplier (α)	Male	Female	Both	Male	Female	Both
1	0.8974	1.0000	0.9411	0.9486	0.9892	0.9809
10	0.9832	1.0000	0.9881	0.9866	0.9824	0.9849
50	0.9997	0.9972	0.9993	**0.0194**	0.9727	0.0809
	Focused Context			**Realistic Context**		
Multiplier (α)	Male	Female	Both	Male	Female	Both
1	0.9994	0.6582	0.9716	0.9998	1.0000	1.0000
10	0.6741	0.2295	0.1381	0.4436	0.9167	0.8058
50	**0.0000**	**0.0000**	**0.0000**	0.0533	0.5379	0.2941

Table 2: P-values from χ^2 tests comparing feature identification strategies. **Bold** values indicate $p < 0.05$.

This result is revealing: the subtle mathematical distinctions between the strategies only manifest under extreme force and are most pronounced in the least realistic, most controlled setting (**Focused Context**). In practical terms, the more complex strategies offer no consistent advantage over the simplest one. Given this lack of significant, widespread difference, we adopt the principle of parsimony for the remainder of our analysis. To avoid redundancy and improve clarity, we will use the straightforward **Absolute Difference** strategy as the default for illustrating the core phenomena of steering.

4.2. Professional Bias: Rigid Performance Ceiling

This experiment investigates bias in a general professional setting using a diverse set of naturalistic sentences. The intervention's efficacy was limited not by inconsistency, but by a remarkably rigid performance ceiling that was unresponsive to parameter changes. The results demonstrate that steering reliably suppresses stereotypical completions but frequently fails to correct all of them, and increasing the intervention's force provides no tangible benefit.

For male-context sentences, the intervention successfully altered the outcome in 80-90% of trials. However, this was primarily a result of the bias being neutralized rather than flipped. The number of successful corrections (`Bias Flipped`) remained static at **4 out of 10 trials**, regardless of the number of features or the multiplier strength. This suggests a hard limit on the intervention's corrective power. Crucially, using just the single most differentiating neuron ($N = 1$) at the lowest multiplier ($\alpha = 1$) was as effective as any other combination, indicating that additional force is unnecessary and offers no improvement. The failures were not random: sentences with strong professional or status-based terms (e.g., "The firm hired David..." and "The old king...") consistently resisted debiasing.

The intervention proved even less effective when debiasing female-context sentences, where the outcome was almost completely invariant to the steering parameters. The number of successful flips remained fixed at a low **2 out of 10 trials across all settings**. The sentences that consistently resisted debiasing were those with strong domestic or relational roles, such as "As a mother, Joan..." and "As a wife...". The name "Joan," which can be ambiguous without strong context, may also contribute to the model's reliance on the powerful stereotypical cue of "mother," highlighting the difficulty of steering against deeply entrenched associations. Ultimately, this experiment shows that in varied, naturalistic contexts, simply applying more force does not overcome the model's inherent biases.

Parameters		Outcome Counts		
Features (N)	**Multiplier (α)**	**Flipped**	**Persists**	**Neutralized**
Debiasing Male-Context Sentences				
1	1	4	2	4
	10	4	1	5
	50	4	1	5
3	1	4	1	5
	10	4	2	4
	50	4	2	4
5	1	4	2	4
	10	4	1	5
	50	4	2	4
10	1	4	2	4
	10	4	1	5
	50	4	2	4
Debiasing Female-Context Sentences				
1	1	2	2	6
	10	2	2	6
	50	2	2	6
3	1	2	2	6
	10	2	2	6
	50	2	2	6
5	1	2	2	6
	10	2	2	6
	50	2	2	6
10	1	2	2	6
	10	2	2	6
	50	1	2	7

Table 3: Detailed Outcomes for Experiment 10: General Professional Contexts (Absolute Difference)

4.3. Adjective Bias: Steering as a Suppressor

This experiment assesses the model's bias in associating character-based adjectives with gendered subjects. The prompts are designed to elicit descriptive words. As shown in Table 4, while many adjectives have a balanced probability, a clear bias exists for words like "powerful" (male biased) versus communal words like "talented" (female biased). When attempting to correct this, activation steering primarily acted as a suppressor rather than a corrective tool. The intervention was highly effective at preventing the model from outputting its original stereotype, with the bias being either flipped or neutralized in over 95% of cases. However, this success was overwhelmingly due to the bias being neutralized; successful flips were minimal, typically

only 1 or 2 out of 21 trials.

More importantly, this experiment highlighted a critical failure mode: excessive force is counterproductive. For male-context sentences, while most parameter settings resulted in zero instances of the original bias persisting, the most aggressive intervention ($N = 10, \alpha = 50$) caused a complete divergence from the target token. The number of `Bias Persists` cases jumped from zero to 9. This suggests that an overly strong push can shatter the intervention's delicate control, causing the model to revert forcefully to its original biased behavior. A similar, though less pronounced, pattern was observed for female-context sentences, where the number of persistent biases also increased at higher parameter settings.

An analysis of the sentences that resisted steering reveals a key difference between the two contexts. For male sentences, resistance was broad and shallow, spread across numerous prompts containing strong male-coded archetypes (e.g., "King Arthur," "the groom," "Sir Michael"). In contrast, for female sentences, resistance was narrow and deep. The vast majority of failures stemmed from a single sentence: "The reason the client hired her is that Ms. Evans is a very...". We hypothesize that the surname "Evans"—which is also a common male first name—introduces an ambiguity that anchors the model's underlying bias, making this specific sentence exceptionally difficult to steer, even when the pronoun "her" and title "Ms." provide clear gender context.

Parameters		Outcome Counts			
Features (N)	Multiplier (α)	Flipped	Persists	Neutralized	Conflicting
Debiasing Male-Context Sentences					
1	1	2	0	19	0
	10	2	0	19	0
	50	2	0	19	0
3	1	1	0	20	0
	10	1	0	20	0
	50	2	0	19	0
5	1	2	0	19	0
	10	1	0	20	0
	50	0	1	19	1
10	1	1	0	20	0
	10	1	0	20	0
	50	0	9	12	0
Debiasing Female-Context Sentences					
1	1	1	1	19	0
	10	1	1	19	0
	50	0	1	20	0
3	1	1	1	19	0
	10	1	0	19	1
	50	1	1	19	0
5	1	1	1	19	0
	10	1	0	19	1
	50	0	2	19	0
10	1	1	1	19	0
	10	1	2	18	0
	50	1	4	16	0

Table 4: Detailed Outcomes for Experiment 11: Descriptive Adjectives (Absolute Difference)

4.4. Focused Context: Brittle Success in a Sterile Setting

This experiment provides the most compelling evidence for our thesis by testing the intervention in a "sterile" environment. It uses highly structured, templated sentences (e.g., "My dad works in the hospital as a...") that isolate the core doctor/nurse stereotype. Our hypothesis was that this idealized context would be easier to steer; however, we observed unexpected volatility in simpler prompts. By removing rich linguistic cues, the sterile prompts may force the model to rely more heavily on the raw statistical bias associated with the gendered subject, making the bias pathologically strong and the steering outcomes highly volatile.

For male-context sentences, the results demonstrated both remarkable success and catastrophic failure. At lower steering strengths, the intervention was largely ineffective, with the original bias persisting in most trials. However, a clear "sweet spot" emerged at a high intensity ($N = 3, \alpha = 50$), where the intervention achieved an impressive **9 out of 12 successful flips** (a 75% correction rate), with the remaining 3 cases being neutralized. This peak performance proves that steering *can* be highly effective under precisely tuned conditions.

However, this success was extremely brittle. A seemingly minor increase in force—from using 3 features to 5 at the same multiplier ($\alpha = 50$)—caused a **non-monotonic collapse**. The success rate plummeted from 9 flips to zero, with all 12 trials resulting in `Bias Neutralized`. The intervention's behavior flipped from being highly corrective to purely suppressive. This demonstrates a critical finding: more force is not always better, and crossing an unknown threshold can completely erase the desired outcome.

For female-context sentences, the intervention was far less effective, and the model's internal state appeared more confused. The original bias was highly persistent across most parameter settings, and there was a notably higher rate of `Conflicting` outputs compared to previous experiments. While a weak sweet spot of 3 successful flips was found, the broader pattern mirrored the male context: at the highest parameter setting ($N = 10, \alpha = 50$), the system again collapsed, with all 12 trials resulting in the bias being neutralized. This experiment proves that even in a controlled environment, the success of neuron steering exists within a narrow and unpredictable parameter window, outside of which its performance degrades catastrophically.

Parameters		Outcome Counts			
Features (N)	Multiplier (α)	Flipped	Persists	Neutralized	Conflicting
Debiasing Male-Context Sentences					
1	1	1	9	1	1
	10	1	8	1	2
	50	3	6	1	2
3	1	1	9	1	1
	10	1	8	2	1
	50	9	0	3	0
5	1	1	8	1	2
	10	1	8	2	1
	50	0	0	12	0
10	1	2	7	1	2
	10	6	3	0	3
	50	0	0	12	0
Debiasing Female-Context Sentences					
1	1	1	10	0	1
	10	1	11	0	0
	50	1	9	1	1
3	1	0	12	0	0
	10	1	11	0	0
	50	3	6	1	2
5	1	1	10	0	1
	10	1	9	0	2
	50	1	8	1	2
10	1	1	11	0	0
	10	3	5	2	2
	50	0	0	12	0

Table 5: Detailed Outcomes for Experiment 12: Sterile Nurse/Doctor Context (Absolute Difference)

4.5. Realistic Context: Unreliable Steering in the Wild

This final experiment serves as the ultimate test of the steering method's practical reliability, using a large, diverse dataset of 53 unique, "in-the-wild" sentences for each gender context. These sentences, introduce realistic linguistic variation and complexity, moving far beyond the sterile templates of the previous experiment. Faced with this complexity, the intervention's effects became chaotic and unpredictable, confirming its fundamental unsuitability as a robust debiasing tool.

Across both male and female contexts, the most striking result was the high prevalence of two failure modes: the original bias persisting and, most notably, the model producing **conflicting and incoherent outputs**. Unlike in simpler contexts where the intervention would cleanly suppress the bias, here it frequently pushed the model into a confused state. This indicates that in a complex linguistic environment, the blunt force of the steering vector interferes with the model's generative process in unpredictable ways, resulting in unreliable and often nonsensical completions.

While parameter tuning could occasionally increase the number of successful `Bias Flipped` outcomes (peaking at 15 for males and 22 for females), these "sweet spots" were erratic and followed no clear pattern. There was no monotonic relationship between increasing the intervention's force and improving the results. Furthermore, the catastrophic collapse observed in Experiment 12 reappeared here with even greater intensity. At the most aggressive setting ($N = 10, \alpha = 50$), the outcome for both genders collapsed into `Bias Neutralized`, with the count skyrocketing to 26 for male sentences and 28 for female sentences. This proves that the brittle nature of steering is not an artifact of a sterile environment but a fundamental characteristic of the method. The intervention frequently fails, produces incoherent outputs, and its effectiveness collapses unpredictably at higher strengths, making it unsuitable for reliable, real-world application.

Parameters		Outcome Counts			
Features (N)	Multiplier (α)	Flipped	Persists	Neutralized	Conflicting
Debiasing Male-Context Sentences					
1	1	9	17	7	20
	10	9	17	6	21
	50	9	18	7	19
3	1	9	19	6	19
	10	8	21	5	19
	50	7	20	3	23
5	1	8	22	5	18
	10	9	20	4	20
	50	15	15	15	8
10	1	8	19	5	21
	10	15	10	6	22
	50	9	14	26	4
Debiasing Female-Context Sentences					
1	1	15	15	4	19
	10	15	15	4	19
	50	15	15	5	18
3	1	15	14	4	20
	10	16	14	4	19
	50	22	9	4	18
5	1	15	14	5	19
	10	19	12	5	17
	50	14	10	20	9
10	1	16	12	5	20
	10	18	13	8	14
	50	12	9	28	4

Table 6: Detailed Outcomes for Experiment 13: Realistic Nurse/Doctor Context (Absolute Difference)

5. Conclusions

This paper investigated the practical reliability of feature steering as a technique for mitigating gender bias in LLMs. Our findings demonstrate a critical disconnect between the theoretical potential of direct neural intervention and its real-world applicability. While our methodology confirmed that steering can, under highly controlled conditions, achieve a near-perfect reversal of stereotypical outputs, our primary contribution is the empirical evidence that this success is exceptionally brittle and context-dependent.

Across four distinct experimental contexts, we observed a consistent pattern of unreliable behavior. In naturalistic settings, the intervention hit a rigid performance ceiling that could not be surpassed by increasing its force (**Professional Bias**). In more focused contexts, we found that excessive force was often counterproductive, causing the intervention to backfire and the original bias to become more persistent (**Adjective Bias**). This instability culminated in the core finding of non-monotonic collapse: in a sterile environment designed for optimal performance, a perfectly tuned intervention could fail with a minor parameter change, flipping from a state of high corrective power to one of mere suppression (**Focused Context**). When tested in a realistic "in-the-wild" scenario, these failure modes converged, resulting in chaotic and frequently inco-

herent outputs (**Realistic Context**).

A key contribution of this work was the systematic comparison of four distinct strategies for identifying the bias circuit. Our statistical analysis revealed that for the vast majority of conditions, the choice of strategy had no significant impact on the outcome. The subtle differences between them only became apparent under the most extreme and unrealistic steering intensity. This implies that the observed unreliability is not simply a matter of finding the "correct" features; it is a fundamental limitation of the steering mechanism itself. More complex heuristics for circuit identification do not solve the underlying problem of brittleness.

The core implication is that simple activation steering, while a powerful diagnostic and proof-of-concept for interpretability research, is not a robust tool for ensuring AI fairness in deployment. Its unreliability highlights that merely identifying and pushing on bias-implicated features is an insufficient, blunt-force approach prone to unpredictable failure. Our findings provide a rigorous empirical foundation for the field's shift towards more sophisticated and robust control mechanisms, underscoring the deep challenges that remain on the path to creating truly fair and reliable AI systems.

6. Limitations and Future Work

Our study's conclusions are based on a single model, `gemma-2-2b`, and focus specifically on gender bias within the English language. The primary technical limitation is the current reliance on a model-specific transcoder required for the analysis, which is not yet available for other models such as `gemma-9b`. Furthermore, our focus on English avoids the complexities of grammatical gender; it would be highly insightful to examine how these circuits behave in morphologically richer languages like Spanish, where gender is encoded in articles and adjectives

Future work should proceed in three key directions. First, our comparative methodology should be applied to other models as the necessary open-source tools and transcoders become available. Second, research should move from steering polysemantic features to intervening directly on the monosemantic features within the Sparse Autoencoder's latent space, which may offer more precise control. Finally, the development of adaptive steering mechanisms, where the intervention's strength is dynamically adjusted based on context, could potentially mitigate the brittle failure modes we identified.

Acknowledgements

We wish to express our sincere gratitude to **NeuronPedia** for providing access to the computational models and software tools via their API. Their infrastructure was fundamental to the experimental phase of this research and the validation of our results.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852) funded by Junta de Andalucía.

Data and Code Availability

The complete experimentation code is publicly hosted and accessible on GitHub at: Project Code. All the generated resources as part of the experimentation are also in the repository inside the data folder.

Additionally, a live report viewer of the experimental results and the full dataset of used sentences and resulting tokens can be accessed here: Results and dataset

Ethics Statement

The research presented in this paper is motivated by the goal of mitigating harmful societal biases, specifically gender stereotypes, that are perpetuated by LLMs. Our work aims to contribute to the development of more equitable AI systems.

However, we acknowledge several ethical considerations. The core technique of activation steering, while used here for bias mitigation, is a general method for manipulating model behavior. It could potentially be misused to amplify biases, generate malicious content, or otherwise alter a model's output in undesirable ways. Our research, which explores the limitations of this technique, is intended to highlight the care that must be taken before such methods are deployed.

Furthermore, our experiments necessarily involve prompting a model to generate text that may contain stereotypical or biased language. This was done in a controlled research environment to analyze and measure the model's behavior, not to generate harmful content for public consumption.

Finally, a central conclusion of our work is the unreliability of simple feature steering. We explicitly caution against interpreting this method as a ready-to-deploy solution for debiasing. The brittleness we observe underscores the risk of implementing superficial fixes that may provide a false sense of security while failing to address the underlying bias in a robust manner.

7. References

Adam Zewe Adam Zewe | MIT News. Unpacking the bias of large language models.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. Circuit tracing: Revealing computational graphs in language models.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Jacy Reese Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, Alexander D'Amour, and Chenhao Tan. 2024. The impossibility of fair llms. *ArXiv*, abs/2406.03198.

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. 2025. Steering large language model activations in sparse spaces. *ArXiv*, abs/2503.00177.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

A. Bhattacharyya. 1946. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406.

Ahsan Bilal, David Ebert, and Beiyu Lin. 2025. Llms for explainable ai: A comprehensive survey.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features. *ArXiv*, abs/2411.02193.

Bhavik Chandna, Zubair Bashir, and Procheta Sen. 2025. Dissecting bias in llms: A mechanistic interpretability perspective. *ArXiv*, abs/2506.05166.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Zhoujie Ding, Ken Ziyu Liu, Pura Peetathawatchai, Berivan Isik, and Sanmi Koyejo. 2024. On fairness of low-rank adaptation of large models. *ArXiv*, abs/2405.17512.

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

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7).

Gemma Team. 2024. Gemma.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *ArXiv*, abs/1903.03862.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Amr Hegazy, Mostafa Elhoushi, and Amr Alanwar. 2025. Guiding giants: Lightweight controllers for weighted activation steering in llms. *ArXiv*, abs/2505.20309.

Aliyah R Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Y Odisho, Peter R Carroll, and Bin Yu. 2024. Efficient automated circuit discovery in transformers using contextual decomposition. *arXiv preprint arXiv:2407.00886*.

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

Adam Karvonen and Samuel Marks. 2025. Robustly improving llm fairness in realistic settings via interpretability. *ArXiv*, abs/2506.10922.

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

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2.

Johnny Lin. 2023. Neuronpedia: Interactive reference and tooling for analyzing neural networks. Software available from neuronpedia.org.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangde. 2024. Steering language model refusal with sparse autoencoders. *ArXiv*, abs/2411.11296.

Avash Palikhe, Zhenyu Yu, Zichong Wang, and Wenbin Zhang. 2025. Towards Transparent AI: A Survey on Explainable Large Language Models. *arXiv e-prints*, page arXiv:2506.21812.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *ArXiv*, abs/2312.06681.

Marco Scialanga, Thibault Laugel, Vincent Grari, and Marcin Detyniecki. 2025. SAKE: Steering Activations for Knowledge Editing. *arXiv e-prints*, page arXiv:2503.01751.

Emily Sheng, Kai-Wei Chang, P. Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *ArXiv*, abs/2105.04054.

Manan Suri, Nishit Anand, and Amisha Bhaskar. 2025. Mitigating memorization in llms using activation steering. *ArXiv*, abs/2503.06040.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C Daniel Freeman, Theodore R Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Published May 21, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David S. Udell, Juan J. Vazquez, Ulisse Mini, and Monte Stuart MacDiarmid. 2023. Steering language models with activation engineering.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. Taxonomy, opportunities, and challenges of representation engineering for large language models. *ArXiv*, abs/2502.19649.

Xin Xu, Wei Xu, Ningyu Zhang, and Julian McAuley. 2025. Biasedit: Debiasing stereotyped language models via model editing. *ArXiv*, abs/2503.08588.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Troy Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. 2023. Representation engineering: A top-down approach to ai transparency. *ArXiv*, abs/2310.01405.