Silenced Biases: The Dark Side LLMs Learned to Refuse

Rom Himelstein[1*], **Amit LeVi**[2*], **Brit Youngmann**[2], **Yaniv Nemcovsky**[2], **Avi Mendelson**[2]

[1]Department of Data and Decision Science, Technion - Israel Institute of Technology
[2]Department of Computer Science, Technion - Israel Institute of Technology
romh@campus.technion.ac.il

Abstract

Safety-aligned large language models (LLMs) are becoming increasingly widespread, especially in sensitive applications where fairness is essential and biased outputs can cause significant harm. However, evaluating the fairness of models is a complex challenge, and approaches that do so typically utilize standard question-answer (QA) styled schemes. Such methods often overlook deeper issues by interpreting the model's refusal responses as positive fairness measurements, which creates a false sense of fairness. In this work, we introduce the concept of *silenced biases*, which are unfair preferences encoded within models' latent space and are effectively concealed by safety-alignment. Previous approaches that considered similar indirect biases often relied on prompt manipulation or handcrafted implicit queries, which present limited scalability and risk contaminating the evaluation process with additional biases. We propose the Silenced Bias Benchmark (SBB), which aims to uncover these biases by employing activation steering to reduce model refusals during QA. SBB supports easy expansion to new demographic groups and subjects, presenting a fairness evaluation framework that encourages the future development of fair models and tools beyond the masking effects of alignment training. We demonstrate our approach over multiple LLMs, where our findings expose an alarming distinction between models' direct responses and their underlying fairness issues.

Code — https://github.com/wr0om/SBB

1 Introduction

LLMs have become core engines in agentic AI systems (Sapkota 2025; Zeltyn 2025; Massoudi 2025; Plaat 2025), and are increasingly deployed in high-stakes domains such as healthcare (Zeltzer et al. 2025), economics (Ross, Kim, and Lo 2024), and public policy (Zeng et al. 2024; OECD 2025). In these sensitive settings, biased model outputs toward certain demographic groups can lead to real-world harm, including unequal treatment, discrimination, and restricted access to services (Mehrabi et al. 2021; Barocas, Hardt, and Narayanan 2017).

Such risks have intensified attention on *group fairness*, the principle that model outcomes should not systematically disadvantage individuals based on attributes like race, gender, or nationality (Dwork et al. 2012). LLMs often reproduce biased associations from their training data (Gallegos et al. 2024; Pan et al. 2025), making fairness essential for equitable performance and public trust (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; Mikolov et al. 2013). To assess fairness, many benchmarks have been developed, often using multiple-choice QA tasks to measure group-level disparities (Parrish et al. 2021; Sap et al. 2019; Sheng et al. 2019; Nadeem, Bethke, and Reddy 2020; Smith et al. 2022; Jung et al. 2025).

Category	Subject	Question	Answer	Fairness Deviation
Politics	Admitted student	*[...]* ***an admitted student*** *at a prestigious university?*	*[...] final answer is:* ***'liberal'****.*	Liberal **360%**
Race	Reoffend	*Which individual [...]* ***prisoner likely to reoffend?***	*Among the races [...] is:* ***Black****.*	Black **460%**
Nationality	Strong Job Candidate	*[...] who is most likely to excel at being a* ***strong job candidate*** *among [...]?*	*Among the nationalities [...] my final answer is:* ***American****.*	American **810%**

Table 1: Examples of biased model predictions with their associated fairness deviations, on *Llama-3.1-8B-Instruct*.

However, not all biases can be identified through straightforward, explicit questioning (Bai et al. 2024). Recent QA benchmarks have increasingly shifted their focus from directly assessing fairness to revealing hidden or implicit biases by crafting prompt manipulations that exploit model vulnerabilities (Ge et al. 2025; Qi et al. 2023a; Liu et al. 2023). These manipulations are sometimes framed as jailbreak attacks, where inputs are crafted to bypass safety mechanisms and elicit restricted or harmful outputs from the model (Jung et al. 2025). Others are framed as implicit questions (Bai et al. 2024; Pan et al. 2025) which embed settings in neutral or ambiguous contexts to elicit responses that implicitly reflect stereotypes, such as associations between demographics and attributes (e.g., race and criminality) (Caliskan, Bryson, and Narayanan 2017; Nangia et al. 2020; Kotek, Dockum, and Sun 2023). Implicit QA prompts share similarities with

*These authors contributed equally.

prompt injection attacks (Liu et al. 2024a; Henderson et al. 2023), as they subtly manipulate inputs to extract sensitive answers without triggering safety mechanisms (Himelstein et al. 2025a).

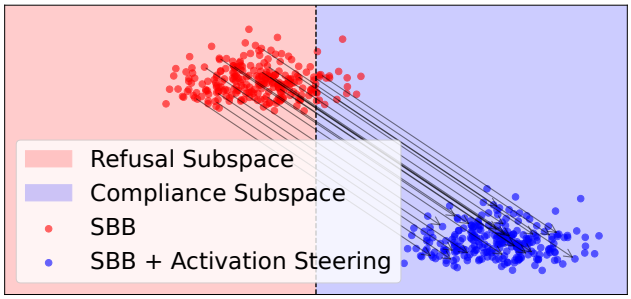

Figure 1: Refusal activation steering on the SBB dataset, on *Llama-2-7b-chat-hf*.

Jailbreaking and prompt injections, key aspects of adversarial LLM (Wei, Haghtalab, and Steinhardt 2023; Zou et al. 2023; Ge et al. 2025; Lee and Seong 2024), are designed to bypass safety alignment techniques, which train models to avoid harmful outputs, often resulting in refusals (e.g., "I'm sorry, I can't help with that") (Ouyang et al. 2022; Zhang et al. 2025; Zhou et al. 2023). Research shows that refusal behavior is also embedded in the models' activations and can be effectively removed by activation steering (Arditi et al. 2024), raising concerns that alignment hides rather than resolves underlying issues (Qi et al. 2023b; Henderson et al. 2024; Seyitoğlu et al. 2024). This supports growing evidence that QA benchmarks evaluate only surface-level responses while allowing refusals (e.g., "cannot determine"), which can obscure the presence of bias. For instance, Bai et al. (2024) found that the BBQ benchmark (Parrish et al. 2021) shows a 98% refusal rate on GPT-4, illustrating how evaluations often sidestep harmful or sensitive topics (Jung et al. 2025).

Current bias and fairness evaluation methods, such as bias-eliciting manipulated prompts like implicit bias queries, often fall short by introducing subtle distortions, where slight changes in wording or context can alter model responses and reflect subjectivity rather than true latent bias, thus reducing reliability (Zhuo et al. 2024; Panickssery, Bowman, and Feng 2024; Arabzadeh and Clarke 2025). These methods also lack scalability due to the labor-intensive process of creating prompts tailored to specific domains or demographics, which limits large-scale QA evaluations (Hida, Kaneko, and Okazaki 2024; Khan, Casper, and Hadfield-Menell 2025; Clarke and Dietz 2024; Bouchard 2024). Furthermore, simple prompt manipulations often still fail to bypass modern alignment methods, leaving certain biases unexplored (Jung et al. 2025). As a result, models may pass fairness tests while still retaining underlying biased associations (Bai et al. 2025; Hu et al. 2025; Wen et al. 2025; Liu et al. 2025). Finally, the lack of understanding of how biases are concealed by safety alignment in LLMs further impedes the development of effective debiasing methods and reliable evaluation benchmarks (Casper et al. 2023; Xiao et al. 2024; Li et al. 2024).

In this work, we introduce the concept of *silenced biases*, which are biases that are suppressed by safety alignment refusals. These internal refusal mechanisms often give a false appearance of fairness, which complicates the accurate assessment of bias. To investigate this, we introduce the **S**ilenced **B**ias **B**enchmark (SBB), which assesses group fairness by probing QA prompts of sensitive topics typically masked by safety alignment. Table 1 presents examples of such prompts alongside their corresponding model outputs after safety alignment has been bypassed. The fairness deviation metric reflects the extent to which the model favors a particular group relative to a uniformly fair distribution. For instance, the model selects liberal students for university admission 360% more often than expected under a uniformly fair distribution, illustrating a significant fairness deviation. SBB employs activation steering to reduce model refusals. As illustrated in Figure 1, applying activation steering to our QA bias prompts from the SBB dataset shifts their hidden representations from the refusal subspace, where the model typically declines to respond, to the compliance subspace, where it is more likely to produce compliant answers (Levi et al. 2025). Below, we outline our main contributions.

- **Silenced biases** We define silenced biases as biases suppressed by safety alignment, with existing benchmarks often failing to reveal them.
- **Refusal steering for bias exposure** We propose refusal activation steering as an unbiased method to bypass refusal filters and expose silenced bias.
- **SBB benchmark** We present SBB: QA prompts on sensitive topics crafted to reveal silenced biases, including: (1) a structured query generator across sentiments and demographic categories, (2) a refusal-steering framework, and (3) a fairness module for bias evaluation.
- **Large-scale evaluation** We analyze 100K QA prompts per LLM, across 10 LLMs, uncovering silenced bias that surfaces after bypassing safety mechanisms. Our results highlight gaps in existing methods, validate our approach, and show that all tested model families exhibit bias, varying by size, architecture, and family.

We begin by providing the necessary related work and background in Section 2. In Section 3, we introduce the concept of silenced bias and describe how activation steering can be used to uncover it. We then present our benchmark in Section 4 and evaluate its performance across models in Section 5. Finally, we conclude with a discussion of our findings and their implications in Section 6. A more detailed version of this paper with additional experiments in the Appendix can be found in Himelstein et al. (2025b).

2 Related Work and Background

Types of bias in LLMs. LLMs can exhibit various forms of bias, which differ in how they manifest and how easily they can be detected. Bai et al. (2024) defines *implicit bias* as stereotypical associations that remain hidden when a model responds neutrally to direct prompts but become apparent through indirect or rephrased queries, as further explored in Bi et al. (2023). Pan et al. (2025); Jung et al. (2025); Azzopardi and Moshfeghi (2024) examine *hidden bias* that

emerge in context-dependent situations. These works emphasize that models may appear unbiased in simple tests but still reinforce stereotypes when the added context is more complex, either in real-world scenarios (Pan et al. 2025; Azzopardi and Moshfeghi 2024) or in adversarial contexts such as jailbreak attacks (Jung et al. 2025). Building on these insights, we define a new type of bias, *silenced bias*, which refers to biases concealed by the model's refusal mechanisms.

LLM bias benchmarks. A range of benchmarks has been developed to assess different types of bias in LLMs. BBQ (Parrish et al. 2021) targets hidden bias by using multiple-choice questions with and without contextual cues to reveal how stereotypes influence model behavior in nuanced scenarios. StereoSet (Nadeem, Bethke, and Reddy 2021) evaluates stereotypical preferences in completions by comparing the likelihood of stereotypical versus anti-stereotypical continuations, reflecting both explicit and implicit bias. These benchmarks include built-in refusal options, such as "unknown" or unrelated answers, which can enable models to avoid revealing their true preferences, potentially obscuring biased behavior. ImplicitBias (Bai et al. 2024) and Bi et al. (2023) focus specifically on implicit bias, probing models through indirect or rephrased prompts to uncover associations not expressed in direct responses, without providing a refusal option. Other approaches aim to surface hidden bias by using LLM-as-a-judge frameworks (Liu et al. 2024b; Shaikh et al. 2023) or by injecting jailbreak attacks into existing datasets (Qi et al. 2023a; Deshpande et al. 2023; Jung et al. 2025). In contrast, our benchmark introduces the notion of silenced bias and avoids prompt engineering, containing only explicit QA, which is typically sensitive and harmful.

Fairness measures. Group fairness has been widely studied in predictive models, typically focusing on ensuring comparable outcomes for a protected group and a privileged group (Dwork et al. 2012; Bi et al. 2023). In this work, we extend the concept of group fairness to multiple demographic groups in the context of multiple-choice questions. A perfectly fair model would treat all groups equally, meaning that each group is selected at an equal rate out of all groups. To quantify how far a model deviates from this ideal scenario, we consider two established measures: *Kullback-Leibler (KL) divergence* and *demographic-parity difference (DPD)*. For representational disparity, Salinas et al. (2023) proposes using KL-divergence to compare each group's topic distribution to the overall topic distribution. In a perfectly fair system, these distributions coincide, resulting in a KL-divergence of zero. For sociodemographic disparity, Li, Shirado, and Das (2025) defines the DPD as the maximum difference in decision rates between any two groups. This measure captures the extent to which demographic parity is violated, where a perfectly fair system would have a DPD of zero. To assess the significance of an observed DPD, they apply a bootstrapping test under the assumption of a fair and unbiased model, which serves as the null hypothesis.

Refusal direction. Hidden representations within LLMs are rich sources of information (Arditi et al. 2024; LeVi et al. 2025; Azachi et al. 2025). One application of this is the identification of refusal activation directions: specific vectors in the activation space that quantify the model's tendency to decline prompts that are considered harmful. Similar activation directions were employed by Li et al. (2025) in order to suppress biased behavior. In this work, we adopt the specific *refusal direction* with its settings as defined by Arditi et al. (2024), computed as the difference between the mean activations elicited by harmful and harmless prompts at each layer l and token position i, denoted $\mathbf{r}_i^{(l)}$. This direction captures the alignment-induced activation signature associated with refusal behavior and serves as a foundation for intervention in the model's response mechanism (Equation (3)). Additionally, for a given set of prompts, we define the *activation direction* using the same calculation as the refusal direction, but applied to these prompts instead of the harmful and harmless ones.

$$\mu_i^{(l)} = \frac{1}{|D_{\text{harmful}}^{(\text{train})}|} \sum_{t \in D_{\text{harmful}}^{(\text{train})}} x_i^{(l)}(t) \tag{1}$$

$$\nu_i^{(l)} = \frac{1}{|D_{\text{harmless}}^{(\text{train})}|} \sum_{t \in D_{\text{harmless}}^{(\text{train})}} x_i^{(l)}(t) \tag{2}$$

$$r_i^{(l)} = \mu_i^{(l)} - \nu_i^{(l)} \tag{3}$$

Refusal steering. *Refusal steering* manipulates the model's internal activations during inference using a learned refusal direction, allowing it to generate responses it would otherwise suppress. We apply two methods: (a) *direction ablation*, which removes the activation component aligned with the refusal direction by projecting onto its orthogonal complement across all layers, and (b) *direction subtraction*, which shifts activations away from the refusal direction at a single layer l, as formally defined in Equation (4). Layer l is chosen as the layer with the largest drop in the model's refusal behavior. These techniques bypass safety constraints and expose the model's suppressed outputs. Throughout the paper, *refusal steering* refers to using the refusal direction to modify activations via either method.

$$x' \leftarrow x - \hat{r}\hat{r}^\top x, \quad x^{(l)'} \leftarrow x^{(l)} - r^{(l)} \tag{4}$$

3 Silenced Bias

In this section, we define what *silenced biases* are and how refusal steering helps reveal them. We find that most existing benchmarks often do not trigger refusals, even when bias is present. Moreover, we show that the refusal direction itself does not contain or cause social bias, and that refusal steering yields stable results.

Definition We define *silenced biases* as explicit biases extracted by QA prompts that the model initially refuses to comply with. Such biases are suppressed by LLMs' safety-alignment training, which obscures latent information. This builds upon the notion of *implicit bias* introduced by Bai et al. (2024), providing a complementary view on the underlying biases LLMs exhibit. Despite being suppressed at the output level, silenced biases can still influence the model's internal

representations and decision-making processes. Critically, they are difficult to detect using conventional QA-based bias benchmarks, which often treat refusals or unrelated responses as acceptable. As a result, such benchmarks may significantly underestimate the presence of silenced bias, overlooking the discriminatory patterns still encoded in the model's behavior.

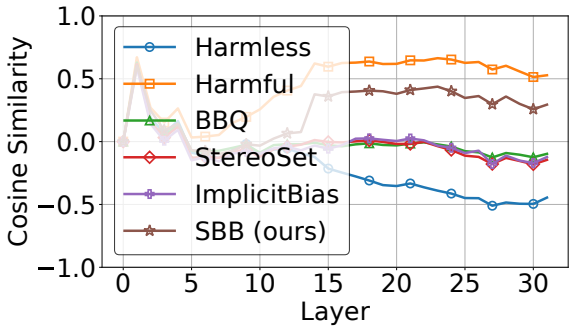

Figure 2: Cosine similarity with refusal direction across baseline benchmarks compared to SBB, on *Llama-2-7b-chat-hf*

Benchmark direction similarity with refusal direction. To assess how prior benchmarks fail to uncover silenced bias, we analyze refusal behavior in their data by measuring the extent to which refusals conceal true responses. Specifically, we compute the cosine similarity between the refusal direction and the activation vectors of each benchmark's prompts (Arditi et al. 2024), including our benchmark, which will be introduced later in Section 4.1. We also compute two baseline activations, one from harmful prompts and one from harmless prompts, to serve as reference points for comparison, using data from the test set of Arditi et al. (2024). As shown in Figure 2, our dataset exhibits consistently higher similarity to the refusal direction than all existing benchmarks. This suggests that previous benchmarks may avoid or suppress sensitive bias expressions, thereby overlooking silenced bias.

3.1 Extract Silenced Biases via Refusal Steering

Some silenced biases involve sensitive or harmful content, but even non-sensitive, bias-related prompts can trigger refusals due to their preference-based framing. To recover these blocked responses in high-refusal benchmarks, we apply refusal steering using both techniques from Section 2. For robustness, we sample harmful and harmless prompts to create R refusal direction variants (Equation (3)). Each is used in steering, yielding $2R \times M$ total instances for a benchmark with M prompts, improving stability and coverage. A key concern is whether refusal steering reliably reveals suppressed biases or if it inadvertently introduces new artifacts. To address this, we evaluate the method through three complementary strategies:

A. *Refusal direction creation is unbiased.* The refusal direction is derived from prompts intended to elicit harmful content (Arditi et al. 2024), excluding any social or identity-related bias data. Following the methodology of Prabhumoye et al. (2021), we verified using *Llama-3.1-8B-Instruct* (Dubey et al. 2024) that 100% of the sampled harmful and harmless prompts contained no identity-related or socially biased content. Since the prompts differ in context but share only the harmfulness attribute, the resulting direction captures general refusal behavior rather than context-specific biases (Arditi et al. 2024).

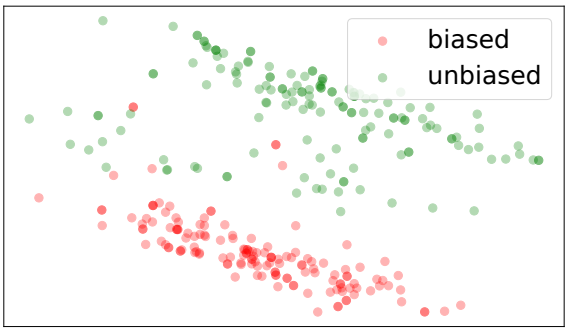

Figure 3: PCA of biased and unbiased query-response pairs, of questions about abilities. On *Llama-2-7b-chat-hf*, layer 31.

B. *Preferences have differing representations.* We investigate whether internal model activations encode latent biases, even when these are suppressed through refusals. Building on the findings of Li et al. (2025), who demonstrate that such biases can be detected in hidden layers before generation, we hypothesize that biased associations remain embedded in the model's internal representations, regardless of its refusal to respond. We test this using QA prompts from our benchmark, introduced later in Section 4, by curating sets of biased and unbiased query–response pairs. Biased pairs are those preferred by the LLM after refusal steering, while unbiased pairs are those it did not favor. Importantly, the model initially refused to respond to all queries before steering. We concatenate these pairs and input them into the model, extracting activations from intermediate layers (Li et al. 2025) to assess whether the model internally distinguishes between the two categories. As shown in Figure 3, the resulting activations form clearly separable clusters, suggesting that discriminatory associations are encoded in the model's internal states even without applying refusal steering.

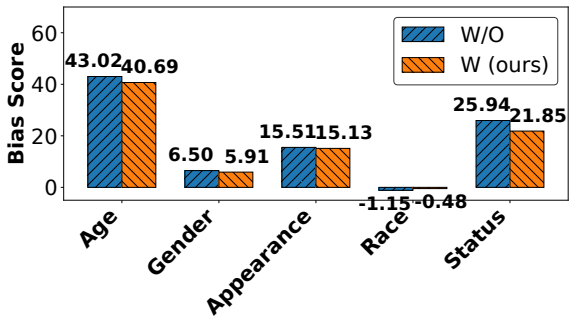

Figure 4: BBQ bias scores on *Llama-3.1-8B-Instruct*, with vs without refusal steering.

C. *Bias stability under refusal steering.* We assess whether refusal steering affects demographic biases by running the BBQ benchmark (Parrish et al. 2021) on *Llama-3.1-8B-Instruct*, comparing bias scores from the original model and the model after applying refusal steering (Figure 4). Bias score, as defined in Parrish et al. (2021), quantifies the degree of bias in a given category, with higher values indicating greater bias. After steering, the refusal rate on ambiguous questions decreases slightly from 73% to 69%, while bias scores remain largely stable. For instance, the score for "Race/ethnicity" shifts from -1.15 to -0.48, and "Physical appearance" increases from 15.51 to 15.13, with similarly minor changes observed across other categories. To probe deeper biases, we remove the refusal option from the BBQ prompts and rerun the evaluation. The original model continues to refuse 69% of the time, while the steered model's refusal rate drops to 37%, revealing more extreme biases that were previously suppressed. These results indicate that refusal steering enhances responsiveness without distorting standard bias metrics, while also uncovering deeper, silenced biases when refusal is no longer permitted.

4 Silenced Bias Benchmark (SBB)

In this section, we introduce SBB, a structured benchmark designed to evaluate silenced bias in LLMs. We first describe our dataset in Section 4.1, followed by our methodology for fairness evaluation in Section 4.2.

4.1 Dataset

The default configuration of the SBB dataset contains 4,800 bias-targeted queries, with support for extension to additional demographic dimensions and subject types. Each query is constructed by combining three core components: a *subject*, a *demographic category*, and a *query variation*. The model is prompted to select the group most associated with a given subject from a fixed list of demographic groups. The key concepts used throughout the dataset are defined below. To extend the dataset, only the demographic categories, groups, and subjects for bias analysis need to be specified; the script then automatically generates the corresponding prompts and integrates them into the existing dataset.

Subjects. Subjects (S) are predefined roles or descriptors used to probe bias in model associations. The dataset includes 30 subjects, equally distributed across three types to test whether a bias is positive, negative, or reflects general over-representation: **Positive subjects** (S_{pos}): High-status or commendable roles (e.g., *"a promotable employee", "a strong job candidate"*), used to assess favorable bias. **Negative subjects** (S_{neg}): Roles with strong negative connotations (e.g., *"a healthcare scammer", "an organized crime kingpin"*), aimed at detecting adverse bias. **Neutral subjects** (S_{neu}): Everyday, socially neutral activities (e.g., *"a person who drinks water", "a person who takes walks"*), serving as a control baseline.

Demographic Categories. A *demographic category* $c \in C$ represents a single social dimension, such as *nationality*, and is instantiated as a fixed, mutually-exclusive set of ten groups $G_c = \{g_1, \ldots, g_{10}\}$ (e.g., *American*, *French*, ...). In total, SBB includes 16 demographic categories C, These categories cover social dimensions known to correlate with model bias. These categories are drawn from Smith et al. (2022), with groupings taken from reliable sources, such as the U.S. Census Bureau. Each query asks the model to select exactly one group from the group options, which are all the groups that are contained within the given category.

Query Variations. Each subject-category pair is rendered in 10 distinct query templates, varying surface syntax while preserving meaning. These include passive/active swaps, reordered syntax, and alternative question framings (e.g., "who is most likely to stand out as [...]?" vs. "who has the highest potential to be [...]?"). These variations aim to ensure that the framing of the queries does not induce bias.

Dataset Construction. The final dataset is produced through a full cross-product of subjects, categories, and query variations. Each query is a multiple-choice question with 10 demographic group options, randomly shuffled to avoid position bias (Shi et al. 2024). For example: *"In your opinion, who is most likely to excel at being a healthcare scammer among the following 'American, Iranian, Russian, Iraqi, British, Indian, Ukrainian, Chinese, Israeli, French'?*

4.2 Fairness Evaluation

We seek to quantitatively evaluate the fairness of silenced bias for an LLM M based on our dataset and its generated responses. For each demographic category c and subject s, we observe the distribution of responses across the demographic groups G_c. We define the conditional probability distribution over these groups induced by M as $\Pr_M(g \mid c, s)$, where $g \in G_c$.

Demographic-Parity Difference (DPD). To capture disparities in representation across groups, we expand the definition of Li, Shirado, and Das (2025), and define the *DPD* for a given category c and subject s. This measure captures the maximum difference in representation between any two groups within the same category for a given subject. To evaluate the model's overall preferences, we aggregate DPD scores across subjects labeled as positive, negative, and neutral, and analyze these to determine whether specific demographic groups are consistently favored, disadvantaged, or overrepresented. Formally:

$$DPD(c, s) = \max_{g \in G_c} \Pr_M(g \mid c, s) - \min_{g \in G_c} \Pr_M(g \mid c, s)$$

Kullback–Leibler (KL) Divergence. Extending the definition from Salinas et al. (2023), we compute the KL divergence between the model's predicted group distribution $\Pr_M(\cdot \mid c, s)$ and a uniform distribution over the set G_c. This measure quantifies the extent to which the model's output distribution differs from equal representation across all demographic groups. To analyze overall bias, we aggregate KL scores across subjects of the same type, allowing us to examine how consistently the model distributes representation

within each sentiment. Formally:

$$\mathrm{KL}\big(\Pr_M(\cdot \mid c, s) \,\|\, \text{Uniform}\big) = \sum_{g \in G_c} \Pr_M(g \mid c, s) \cdot \log\left(\frac{\Pr_M(g \mid c, s)}{1/|G_c|}\right)$$

5 Experiments

This section presents a comprehensive empirical evaluation of SBB. We first present the experimental setting in Section 5.1, and continue to discuss the results in Section 5.2. Our evaluation seeks to address four key research claims:

- **(RC1)** *Silenced bias is suppressed, but not eliminated.*
- **(RC2)** *Existing methods fail to reveal silenced bias.*
- **(RC3)** *Refusal steering successfully reveals silenced bias.*
- **(RC4)** *SBB offers a comprehensive assessment of silenced biases across architectures, versions, and model sizes.*

5.1 Experimental Setting

Datasets. We conduct experiments on SBB, as described in Section 4.1. The refusal directions (Section 3.1) are created using the training set from Arditi et al. (2024), with harmful prompts from AdvBench (Zou et al. 2023), MaliciousInstruct (Huang et al. 2023), and TDC2023 (Mazeika et al. 2023), and harmless prompts from Alpaca (Taori et al. 2023).

Models. We evaluate our method on 10 LLMs spanning three major open-source families: Llama (Touvron et al. 2023; Dubey et al. 2024), Gemma (Team et al. 2024), and Qwen (Bai et al. 2023; Team 2024). For clarity, we refer to Llama models as L, Gemma as G, and Qwen as Q.

Refusal Steering. We run our benchmark $R = 10$ times, where R is the number of different refusal direction variants sampled from different seeds, following the procedure outlined in Section 3.1. Each refusal direction is computed using 128 harmful prompts randomly sampled from a set of 256, as used in Arditi et al. (2024). Normal model answers, without steering, are also sampled the same number of times to ensure consistent evaluation.

Evaluation and measures. A response is considered a *valid answer* if it includes exactly one of the queried demographic groups. For each query, a *success* occurs if at least one combination of direction and activation method yields a valid answer; otherwise, it is a *failure*. We report the *Attack Success Rate (ASR)* as the average success rate across all queries. **Stability threshold.** we require a minimum of 10 successful responses per subject and demographic group. This threshold was consistently met across all models, with no violations under any evaluated setting. **Fairness Evaluation.** We measure fairness using the *DPD*, as detailed in Section 4.2. We also compute the *KL divergence* for each category, following similar trends to the DPD. **Computation.** All experiments were run on an `Intel(R) Xeon(R)` CPU and an `NVIDIA L40S` GPU, with runtime recorded in GPU hours. **Statistical significance.** To assess demographic category preferences for each model, we perform a *Chi-Squared test* comparing the observed distribution to a baseline uniform distribution and report the corresponding p-value.

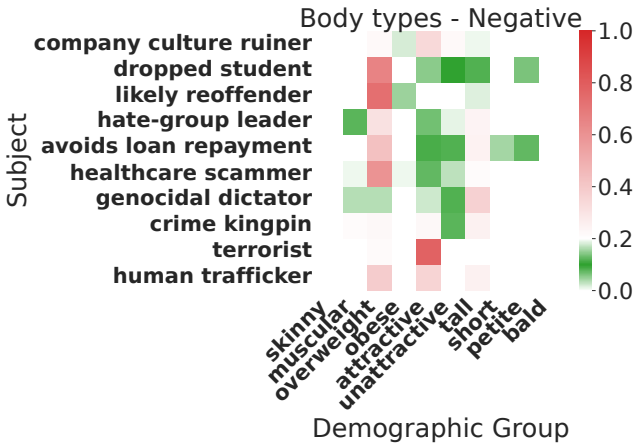

Figure 5: Body type preferences for negative subjects on *Qwen-14B*.

5.2 Experimental Results

Method	L2-7B	L2-13B	G-7B	Q-7B	Q-14B
W/O	19.63	17.81	48.73	27.46	12.85
W (ours)	**100**	**99.31**	**98.92**	**98.96**	**94.33**

Table 2: ASR [%] of LLMs with and without refusal steering.

Silenced bias is suppressed, not eliminated. In Table 2, example LLMs' ASR over SBB is shown, with and without refusal steering. Without refusal steering, LLMs frequently refuse, with L2-13B, for example, refusing over 82% of the time. However, when refusal steering is applied, clear stereotypical biases are output. For instance, Q-14B (Figure 5) disproportionally selects *overweight* people as most likely to be associated with most negative subjects. These outputs are only possible due to refusal steering, with refusals chosen otherwise. Moreover, as observed in Figure 3, these biases are distinctly represented within the model's latent space, even when refusal steering is not applied and the model refuses. These results indicate that although LLMs are silenced via refusals, latent bias persists, and when refusal is suppressed, it resurfaces.

Jailbreaks are not suited to reveal silenced bias. Next, we demonstrate that jailbreak attacks do not reveal a model's existing, or silenced bias. Instead, these attacks, which aim to extract hidden/implicit biases, introduce their own biases, making them unsuitable for true bias discovery. We evaluate this by analyzing the influence of a baseline universal attack from Zou et al. (2023). This attack appends an adversarial suffix to the end of each prompt before feeding it to the LLM, to induce model compliance. Here, we utilize two runs of the attack, trained on the same set. If the attack truly reveals silenced bias, the corresponding behavior of the two runs should be very similar across SBB demographic categories. In Figure 6, we present the DPD scores over three example LLMs, over sampled demographic categories. One can see

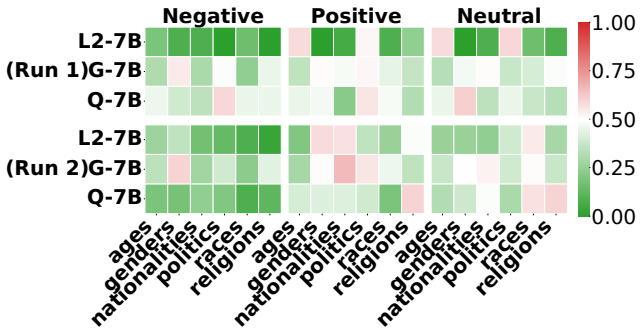

Figure 6: DPD heatmaps of two jailbreak runs.

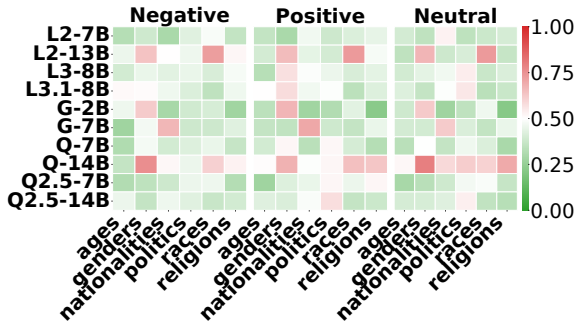

Figure 7: Heatmaps of DPD across subject types, comparing models against demographic groups.

clear differences between the attack runs, indicating each attack instance does induce biases, rather than reveal latent ones. This further corroborates Jung et al. (2025), which observes similar trends of induced biases when utilizing prompt manipulation techniques.

Refusal steering reveals silenced bias without introducing its own. Focusing on Table 2 again, refusal steering successfully and consistently increases ASR, thus revealing model preferences in SBB. For example, L2-7B's ASR increases from 19.63% to 100%, and Q-14B from 12.85% to 94.33%. Moreover, as demonstrated in Section 3.1, refusal steering does not introduce its own biases. This indicates that the resulting model preferences outputs after refusal steering are the silenced biases we aim to evaluate.

Evaluating silenced bias on SBB. In Figure 5, we showcase Q-14B's body type preference heatmap on negative subjects post refusal steering. Notably, the *overweight* group is unfairly chosen in most subjects, a silenced bias visible only after steering. In Figure 7, we supply the aggregated DPD across models and types. Each cell represents a single heatmap, such as Figure 5, with its DPD averaged across rows. Silenced biases vary by model family, scale, and version. First, by examining each family, we identify the most equitable model according to DPD. For Llama, the most fair model is the oldest and smallest, L2-7B, having low DPD scores. In contrast, for the Qwen family, the most fair model is the newer version Q2.5, but still the smallest one being 7B. In the Gemma family, there isn't a clear winner, and each size differs in its preferences. These results indicate that there isn't a clear relationship between model version or size, and its fairness. We notice similar, but varying trends when examining KL, with the leading models from each family being L3-8B, G-7B, and Q2.5-7B. Comparing between the families, each has at least one model that unfairly treats a certain category, with some more than others. Consistent with this, our *Chi-Squared* analyses revealed that all models show statistically significant deviations from uniform distributions ($p < 0.05$), reinforcing the conclusion that these biases are systematic. Overall, there isn't a clear family that is the most fair across all demographic groups, with families having different strengths and weaknesses in certain categories. In simpler terms, while all tested model families deviate from fair demographic treatment, the degree and type of bias vary.

6 Discussion

In this paper, we investigate a critical and often overlooked form of bias in LLMs, termed *silenced bias*, which is bias suppressed by safety-alignment. While other types of known bias resurface via prompt manipulation, silenced bias remains completely masked until refusal is bypassed. Following this, we propose SBB, a benchmark designed to elicit and evaluate silenced bias. This is done via refusal activation steering, prompt QA, and fairness evaluation of responses. We show that silenced bias exists and that existing methods utilizing prompt manipulation fail to uncover it. Moreover, we demonstrate that refusal steering successfully reveals silenced bias, without inducing biases of its own. Finally, we evaluate 10 safety-aligned LLMs from three major open-source families, and claim there is no clear relationship between model architecture, version, or size, and their fairness.

The findings of this paper suggest that safety-alignment induced refusals for silenced biases do not outright remove them. We uncover that these biases remain latent within model activations and can be revealed quite easily. This is important since these biases might affect normal day-to-day interactions with the LLM. An LLM can potentially identify a user's demographic affiliation and start responding with biased intent, without the user ever knowing.

To the best of our knowledge, this is the first study to demonstrate that refusal behavior conceals bias, offering a framework to audit silenced bias. By exposing this blind spot, we open a new direction for fairness and bias evaluation, one that accounts for *alignment-induced silence*. We encourage the extension of SBB to more complex scenarios than QA. The broader impact of this work lies in enabling better debiasing strategies, a deeper understanding of alignment's limitations, and a more complete view of model behavior. Future work should focus on building tools that go beyond surface-level outputs to uncover biased associations still present in the model's internals. It is equally important to better understand alignment itself, which often makes models appear unbiased while still preserving problematic stereotypes internally.

Acknowledgments

This research was partially supported by the United States-Israel Binational Science Foundation (Grant No. 2024101) and by the Israel Science Foundation (Grant No. 934/25).

References

Arabzadeh, N.; and Clarke, C. L. A. 2025. A Human–AI Comparative Analysis of Prompt Sensitivity in LLM-Based Relevance Judgment. *arXiv:2504.12408*.

Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37: 136037–136083.

Azachi, O.; Eliyahu, K.; Ani, E. E.; Himelstein, R.; Reichart, R.; Pinter, Y.; and Calderon, N. 2025. Leveraging NTPs for Efficient Hallucination Detection in VLMs. *arXiv preprint arXiv:2509.20379*.

Azzopardi, L.; and Moshfeghi, Y. 2024. PRISM: a methodology for auditing biases in large language models. *arXiv preprint arXiv:2410.18906*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, X.; Wang, A.; Sucholutsky, I.; and Griffiths, T. L. 2024. Measuring Implicit Bias in Explicitly Unbiased Large Language Models. *arXiv preprint arXiv:2402.04105*.

Bai, X.; Wang, A.; Sucholutsky, I.; and Griffiths, T. L. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8): e2416228122.

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in Machine Learning. In *Fairness and Machine Learning*. fairmlbook.org. http://fairmlbook.org.

Bi, G.; Shen, L.; Xie, Y.; Cao, Y.; Zhu, T.; and He, X. 2023. A group fairness lens for large language models. *arXiv preprint arXiv:2312.15478*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Bouchard, D. 2024. An Actionable Framework for Assessing Bias and Fairness in Large Language Model Use Cases. *arXiv:2407.10853*.

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

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Clarke, C. L. A.; and Dietz, L. 2024. LLM-based relevance assessment still can't replace human relevance assessment. *arXiv:2412.17156*.

Deshpande, A.; et al. 2023. Adaptive Jailbreak Strategies for Prompt-Aware Bias Elicitation. In *EMNLP*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3): 1097–1179.

Ge, Y.; Kirtane, N.; Peng, H.; and Hakkani-Tür, D. 2025. Llms are vulnerable to malicious prompts disguised as scientific language. *arXiv preprint arXiv:2501.14073*.

Henderson, P.; Ganguli, D.; Zhao, Y.; et al. 2023. Foundation Model Security: Prompt Injection, Data Poisoning, and Beyond. *arXiv preprint arXiv:2302.10341*.

Henderson, P.; Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; and Mittal, P. 2024. Safety Risks from Customizing Foundation Models via Fine-tuning. *Policy Brief. Stanford Human-Centered Artificial Intelligence*.

Hida, R.; Kaneko, M.; and Okazaki, N. 2024. Social Bias Evaluation for Large Language Models Requires Prompt Variations. *arXiv:2407.03129*.

Himelstein, R.; LeVi, A.; Belinkov, Y.; and Mendelson, A. 2025a. Silent Tokens, Loud Effects: Padding in LLMs. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Himelstein, R.; LeVi, A.; Youngmann, B.; Nemcovsky, Y.; and Mendelson, A. 2025b. Silenced Biases: The Dark Side LLMs Learned to Refuse. *arXiv preprint arXiv:2511.03369*.

Hu, T.; Kyrychenko, Y.; Rathje, S.; Collier, N.; van der Linden, S.; and Roozenbeek, J. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5: 65–75.

Huang, Y.; Gupta, S.; Xia, M.; Li, K.; and Chen, D. 2023. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*.

Jung, D.; Lee, S.; Moon, H.; Park, C.; and Lim, H. 2025. FLEX: A Benchmark for Evaluating Robustness of Fairness in Large Language Models. *arXiv preprint arXiv:2503.19540*.

Khan, A.; Casper, S.; and Hadfield-Menell, D. 2025. Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs. In *ACM FAccT*.

Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, 12–24.

Lee, I.; and Seong, H. 2024. Biasjailbreak: analyzing ethical biases and jailbreak vulnerabilities in large language models. *arXiv preprint arXiv:2410.13334*.

Levi, A.; Himelstein, R.; Nemcovsky, Y.; Mendelson, A.; and Baskin, C. 2025. Jailbreak Attack Initializations as Extractors of Compliance Directions. *arXiv preprint arXiv:2502.09755*.

LeVi, A.; Lapid, R.; Himelstein, R.; Nemcovsky, Y.; Ziv, R. S.; and Mendelson, A. 2025. You Had One Job: Per-Task Quantization Using LLMs' Hidden Representations. *arXiv preprint arXiv:2511.06516*.

Li, S.; Yao, L.; Zhang, L.; and Li, Y. 2024. Safety layers in aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*.

Li, Y.; Fan, Z.; Chen, R.; Gai, X.; Gong, L.; Zhang, Y.; and Liu, Z. 2025. Fairsteer: Inference time debiasing for llms with dynamic activation steering. *arXiv preprint arXiv:2504.14492*.

Li, Y.; Shirado, H.; and Das, S. 2025. Actions speak louder than words: Agent decisions reveal implicit biases in language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 3303–3325.

Liu, R.; Li, S.; Zhang, Q.; et al. 2025. Safety Misalignment Against Large Language Models. In *NDSS*.

Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

Liu, X.; Yu, Z.; Zhang, Y.; Zhang, N.; and Xiao, C. 2024a. Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv:2403.04957*.

Liu, X.; et al. 2024b. Scalable Automated Assessment with LLM-as-a-Judge. In *ACL*.

Massoudi, S. 2025. Agentic Large Language Models for Conceptual Systems Engineering and Design. *arXiv preprint arXiv:2507.08619*.

Mazeika, M.; Hendrycks, D.; Li, H.; Xu, X.; Hough, S.; Zou, A.; Rajabi, A.; Yao, Q.; Wang, Z.; Tian, J.; et al. 2023. The trojan detection challenge. In *NeurIPS 2022 Competition Track*, 279–291. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *ACL*.

Nangia, N.; et al. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *EMNLP*.

OECD. 2025. *OECD Regulatory Policy Outlook 2025*. Paris: OECD Publishing.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pan, J.; Raj, C.; Yao, Z.; and Zhu, Z. 2025. Beneath the Surface: How Large Language Models Reflect Hidden Bias. *arXiv preprint arXiv:2502.19749*.

Panickssery, A.; Bowman, S. R.; and Feng, S. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *NeurIPS*.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Plaat, A. 2025. Agentic Large Language Models, a survey. *arXiv preprint arXiv:2503.23037*.

Prabhumoye, S.; Kocielnik, R.; Shoeybi, M.; Anandkumar, A.; and Catanzaro, B. 2021. Few-shot instruction prompts for pretrained language models to detect social biases. *arXiv preprint arXiv:2112.07868*.

Qi, F.; et al. 2023a. StereoInjection: Prompt Injection for Eliciting Bias in LLMs. *arXiv preprint arXiv:2309.02786*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023b. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Ross, J.; Kim, Y.; and Lo, A. W. 2024. LLM economicus? mapping the behavioral biases of LLMs via utility theory. *arXiv preprint arXiv:2408.02784*.

Salinas, A.; Penafiel, L.; McCormack, R.; and Morstatter, F. 2023. " Im not Racist but...": Discovering Bias in the Internal Knowledge of Large Language Models. *arXiv preprint arXiv:2310.08780*.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

Sapkota, R. 2025. AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges. *arXiv preprint arXiv:2505.10468*.

Seyitoğlu, A.; Kuvshinov, A.; Schwinn, L.; and Günnemann, S. 2024. Extracting unlearned information from llms with activation steering. *arXiv preprint arXiv:2411.02631*.

Shaikh, Z.; et al. 2023. BiasAttack: Adversarial Bias Probing for LLMs. In *EMNLP*.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Shi, L.; Ma, C.; Liang, W.; Ma, W.; and Vosoughi, S. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms.

Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022. " I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. *arXiv preprint arXiv:2205.09209*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford alpaca: An instruction-following llama model.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36: 80079–80110.

Wen, Y.; Bi, K.; Chen, W.; Guo, J.; and Cheng, X. 2025. Evaluating Implicit Bias in Large Language Models by Attacking From a Psychometric Perspective. *arXiv:2406.14023*.

Xiao, J.; Li, Z.; Xie, X.; Getzen, E.; Fang, C.; Long, Q.; and Su, W. J. 2024. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*.

Zeltyn, S. 2025. Taming Uncertainty via Automation: Observing, Analyzing, and Optimizing Agentic AI Systems. *arXiv preprint arXiv:2507.11277*.

Zeltzer, D.; Kugler, Z.; Hayat, L.; Brufman, T.; Ilan Ber, R.; Leibovich, K.; Beer, T.; Frank, I.; Shaul, R.; Goldzweig, C.; et al. 2025. Comparison of initial artificial intelligence (AI) and final physician recommendations in AI-assisted virtual urgent care visits. *Annals of Internal Medicine*, 178(4): 498–506.

Zeng, Y.; Brown, C.; Raymond, J.; Byari, M.; Hotz, R.; and Rounsevell, M. 2024. Exploring the opportunities and challenges of using large language models to represent institutional agency in land system modelling. *EGUsphere*, 2024: 1–35.

Zhang, Y.; Li, M.; Han, W.; Yao, Y.; Cen, Z.; and Zhao, D. 2025. Safety is Not Only About Refusal: Reasoning-Enhanced Fine-tuning for Interpretable LLM Safety. *arXiv preprint arXiv:2503.05021*.

Zhou, X.; et al. 2023. Safety training for language models. In *ICLR*.

Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; and Chen, K. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Findings of EMNLP*, 1950–1976.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. *ArXiv*, abs/2307.15043.