Locate-then-Debias: Understanding and Mitigating Model Bias in Hidden Representations Explanation of Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated strong capabilities for hidden representation interpretation through Patchscopes (Ghandeharioun et al., 2024), a framework that uses LLMs themselves to generate human-readable explanations by decoding from internal hidden representations. However, our work shows that LLMs tend to rely on inherent linguistic patterns, which can override contextual information encoded in the hidden representations during decoding. For example, even when a hidden representation encodes the contextual attribute "purple" for "broccoli", LLMs still generate "green" in their explanations, reflecting a strong prior association. This behavior reveals a systematic unfaithfulness in Patchscopes. To systematically study this issue, we first designed a dataset to evaluate the faithfulness of Patchscopes under biased cases, and our results show that there is an 18.84% faithfulness decrease on average. We then propose Locate-then-Debias through Logit Recalibration (LDLoR), which treats the output logits from an unpatched prompt as capturing model bias and contrasts them with logits obtained under patched contextual information. By recalibrating the logit distribution through this contrast, LDLoR suppresses model bias and amplifies contextual information during generation. Experiments across multiple LLMs demonstrate that LDLoR consistently outperforms vanilla Patchscopes, achieving up to 33% relative show rate improvement.

1 Introduction

Large language models (LLMs) work by transforming input tokens into high-dimensional vectors and repeatedly feeding them through multiple layers of feed-forward transformations. The outputs of each feed-forward transformation are hidden representations that fundamentally govern how models encode attributes, reason over context, and ultimately generate predictions (Orgad et al., 2025).

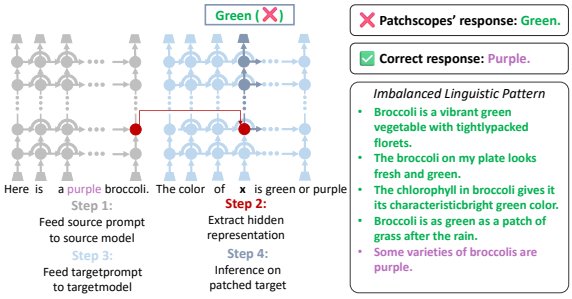

Figure 1: Overview of the Patchscopes framework and an example illustrating its unfaithfulness under an imbalanced linguistic pattern. The patched hidden representation encodes the contextual attribute "purple" for "broccoli", but under Patchscopes' framework, LLMs still generate "green" in their explanations.

Thus, understanding what information is stored and how it is used inside these hidden representations is essential for building trustworthy LLMs. Hidden representation explanation methods, such as linear probing (Alain and Bengio, 2018; Belinkov, 2022), train a linear classifier to analyze whether content can be extracted from hidden representations, and causal interventions (Vig et al., 2020; Meng et al., 2023) change keywords in the input prompts to analyze the changes in hidden representations. More recent work, Patchscopes (Ghandeharioun et al., 2024) explains hidden representations by using the LLMs themselves as interpreters. As illustrated in Figure 1, a source prompt provides contextual information describing broccoli as purple. Then, the hidden representation of broccoli, which corresponds to the purple attribute value, is extracted. This hidden representation is then patched into a target prompt by replacing the hidden representation at a placeholder token x during the target prompt's inference process. Conditioned on the patched broccoli's hidden representation, the model generates a natural-language response that serves as an explanation of the information encoded in the patched hidden representation.

Despite the promising nature of the Patchscopes framework, it cannot guarantee that the content generated for explanation faithfully reflects the information contained in the hidden representations (Lyu et al., 2024; Bhan et al., 2025; Hu et al., 2025). Our work observed that the generated explanation content is heavily affected by the model's inherent bias, which is induced by imbalanced linguistic patterns and prevents models from faithfully reflecting the contextual information in their hidden representations. For example, as Figure 1 shows, in most linguistic patterns, broccoli is associated with the green color attribute values, while only a few are associated with purple. When we want to explain the color-related hidden representations of broccoli encoded with this less-frequent purple attribute, Patchscopes ignores the purple information in this hidden representation and instead generates the "green" attribute. While Figure 1 provides an illustrative example of this phenomenon, we further validate this effect through a systematic controlled analysis in Section 3.3. We demonstrate that the faithfulness of the Patchscopes framework is distorted by this model bias, which leads to an 11.84% to 28.47% decrease in explanation faithfulness. This observation highlights a fundamental challenge in Patchscope: during generation, LLMs often rely on high-frequency associations learned from imbalanced linguistic patterns instead of the contextual information encoded in their hidden representations.

It is notable that current debiasing methods are difficult to integrate directly into the Patchscopes framework, making a simple "debiasing-then-patching" pipeline impossible. Supervised finetuning (SFT) requires large annotated datasets to capture both biased and unbiased instances (Lin et al., 2022), which can be very expensive. Also, SFT alters model parameters globally, thereby changing the original hidden representations and undermining Patchscopes' goal of faithfully explaining what the original model encodes. Prompt-based debiasing methods (Amplayo et al., 2023) are training-free, but they depend heavily on prompt designations. The results are unstable and vary with minor lexical or syntactic changes. Steering vector intervention-based methods (Li et al., 2025) attempt to remove or steer bias directions within latent space by identifying and manipulating subspaces of internal representations. However, these approaches assume that model bias can be isolated into well-defined and linearly separable directions, and the exact directions corresponding to model bias are unknown and hard to estimate reliably.

In order to overcome these limitations, we propose Locate-then-Debias through Logit Recalibration (LDLoR), a logit-recalibration decoding method that directly targets the source of distortion, which is motivated by the knowledge that logit-level probability distributions directly determine text generation and reflect the influence of model bias. By subtracting logits obtained from an unpatched contrastive prompt, internal evidence is contrastively separated from empirical priors in LDLoR. This approach amplifies contextual information encoded in the patched representation while suppressing undesired model bias. In summary, this paper makes three key contributions:

- We designed a Q&A dataset according to linguistic pattern-induced model bias. Our analysis demonstrates that model bias fundamentally limits the faithfulness of hidden-representation explanations in LLMs.
- We introduce a logit-recalibration decoding method LDLoR to locate and mitigate model bias at the inference stage.
- We provide extensive experiments across four modern LLMs, demonstrating that LDLoR substantially improves explanation faithfulness by up to 33% while remaining robust across hyperparameters and decoding temperatures.

2 Preliminaries

We follow the framework and notation of Patchscopes (Ghandeharioun et al., 2024). Let $S = \langle s_1, \ldots, s_n \rangle$ be a source prompt of length n, and M be a source LLM with L layers. Feeding S to M yields hidden representations $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^d$, where $i \in \{1, \ldots, n\}$ indexes token positions and $\ell \in \{1, \ldots, L\}$ indexes layers. A source representation to be inspected is specified by the tuple (S, i, M, ℓ), corresponding to $\mathbf{h}_i^{(\ell)}(S)$.

To decode information from this representation, Patchscopes defines a separate target execution. Let $T = \langle t_1, \ldots, t_m \rangle$ be a target prompt and M^* be a target LLM with L^* layers. We denote by $\mathbf{h}_{i^*}^{(\ell^*)}(T)$ the hidden representation at target position $i^* \in \{1, \ldots, m\}$ and layer $\ell^* \in \{1, \ldots, L^*\}$ during the forward pass of M^* on T.

Given an optional mapping function $f : \mathbb{R}^d \to \mathbb{R}^{d^*}$, the patching operation intervenes on the target computation by replacing

$$\mathbf{h}_{i^*}^{(\ell^*)}(T) \leftarrow f\left(\mathbf{h}_i^{(\ell)}(S)\right),$$

while keeping all other hidden representations unchanged. In this work, we restrict our attention to the case where the source and target models are the same. Consequently, the mapping function $f(\cdot) = \mathrm{id}(\cdot)$. The modified hidden representations are then propagated through the remaining layers of M^*, producing a patched output distribution over the next token:

$$p_{\text{patched}}(w_{t+1} \mid T) = \sigma\left(W_o \mathbf{h}_t^{(L^*)}(T) + \mathbf{b}_o\right),$$

where σ denotes the softmax function, W_o and $\mathbf{b}_o$ denote the output projection parameters of M^*.

By construction, this intervention halts further contextualization from the source sequence S; thus, any information reflected in the patched generation must be recoverable from the source representation $\mathbf{h}_i^{(\ell)}(S)$ via the post-patching computation.

3 Inherent Model Bias will Affect Explanation Faithfulness

Although the Patchscopes framework is widely used, whether its use is a faithful method remains a problem. Building on the insight that LLMs exhibit model bias, we need to determine how much faithfulness is affected by this. In this section, we design a dataset to analyze the causal relationship between linguistic patterns, model bias, and the faithfulness of hidden representation explanations.

3.1 Dataset

There are four tasks in this dataset: color, gender, culture, and age. For each datapoint, there will be one *noun* to be explained and two attribute values a_{pri} and a_{sec} for the model to select. Table 1 provides an overview of these four tasks. The source and target prompt examples of each task are provided in Appendix G.3. The dataset consists of two parts: a biased part and a non-biased one. For datapoints in the biased part, the model exhibits a clear preference toward a_{pri} in the absence of additional contextual information, whereas no such preference is observed for datapoints in the non-biased subset. We use $\mathcal{D}$ for each task in each part.

Biased Subset The biased subset contains datapoints from all four tasks. For the color task, we use the category names of the Large Vocabulary Instance Segmentation dataset (Gupta et al., 2019) for nouns as item names. Then, the color attribute values related to these nouns are extracted from the RedPajama subset (Weber et al., 2024), which is described as the best-effort reproduction of the Llama training dataset. Thus, it can reflect the linguistic patterns in natural language. The color with the highest co-occurrence frequency is set as a_{pri}, and the color with the second highest co-occurrence frequency is set as a_{sec}. The nouns are then selected such that the model exhibits a clear preference for a_{pri} in the biased dataset. The details of the selection strategy are provided in Appendix G.1. For gender, culture, and age tasks, we extract the nouns and two attribute values from the dataset provided by (Hernandez et al., 2024). We define the model-preferred attribute identified by Hernandez et al. (2024) as a_{pri}, and the other one as a_{sec}. More details are provided in Appendix G.2.

Task	Noun	Attribute value
Color	item name	Color1, Color2
Gender	degree	He, She
	characteristic	He, She
	person name	He, She
	occupation	He, She
Culture	person name	Country1, Country2
	person name	Religion1, Religion2
Age	occupation	Young, Old

Table 1: Details for each task in the dataset.

Non-biased Dataset The non-biased dataset includes the color task. The datapoints of the color task that are not selected for the biased dataset construct non-biased dataset.

3.2 Experiment Settings

Dataset and Models. We use the biased and non-biased subsets of the color task to analyze the causal relationship between linguistic patterns, model bias, and Patchscopes faithfulness. Due to the larger size of the non-biased dataset, we perform five repeated experiments with undersampling. Four models are used for experiments: Llama2-7b-chat (Meta, 2023), Llama3.2-1b-Instruct, Llama3-8b-Instruct (Meta, 2024), and Qwen3-4b-Instruct-2507 (Qwen, 2025).

Prompts. The source prompt (S_i) is formatted as: `Here is an {a`$_{i,sec}$`} {noun`$_i$`}`. The target prompt (T_i) is formatted as: `The color of x is`

{a$_{i,pri}$} or {a$_{i,sec}$}? During patching, the position of x in the target prompt defines the target index, while the hidden representations at the noun's token indices in the source prompt are extracted for interpretation. For simple implementation, we repeat the placeholder x to match the token length of the noun. To mitigate option position bias, we additionally use a counterpart target prompt with the attribute options swapped and report performance averaged over both prompts.

Evaluation Metric. We use Show Rate (SR ⇑) to evaluate the faithfulness of Patchscopes; SR measures the ratio of responses that LLM generates a_{sec}, the attribute value aligns with contextual information, as an explanation. The SR is

$$SR = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}_i \in \mathcal{D}} \mathbb{I}[M(T_i) = a_{i,sec}] \,.$$

3.3 Analysis

In this section, we aim to determine the relationship between linguistic patterns, model bias, and Patchscopes faithfulness.

First, we examine whether model bias arises from the imbalanced linguistic patterns, which are the co-occurrence frequencies between color names and item names here (Brill, 2024; Kiyomaru et al., 2024). For each noun-color pair, we compute the frequency difference Δf, while $\Delta f = f_{pri} - f_{sec}$ where f_x denotes the co-occurance frequency between a_x and $noun$, $x \in \{pri, sec\}$. Take Llama3.2-1b as an example, across five repeated runs, we fit a logistic regression with Δf standardized to one standard deviation. The logit coefficient, corresponding odds ratio (OR), and ROC-AUC are used to measure predictive performance. The mean value and 95% confidence interval of each metric are shown in Table 2. It shows that every $+1$ standard deviation increase in Δf increases the odds of exhibiting bias by a factor of 9.338 on average, and the ROC-AUC value of 0.619 also shows that the discriminative power suggests that an imbalanced linguistic pattern is a meaningful determinant of model bias. The causal analysis results between linguistic patterns and model bias on additional models are provided in the Appendix A.

Then we test the relationship between model bias and Patchscopes' (Ghandeharioun et al., 2024) faithfulness on biased and non-biased data. The results are shown in Figure 2. The performance gap between biased and non-biased datasets shows that model bias indeed undermines the faithfulness of Patchscopes.

Metric	Mean Value	95% CI
Logit coeff (+1 SD)	2.234	(1.487, 2.981)
OR (+1 SD)	9.338	(4.425, 19.706)
ROC-AUC	0.619	(0.580, 0.655)

Table 2: Repeated undersampling results.

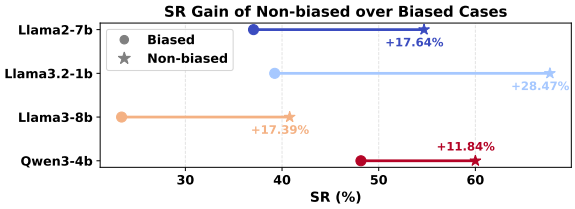

Figure 2: The result of attribute SR for vanilla Patchscopes on biased and non-biased datasets.

4 Methodology

We proposed LDLoR, a locate-then-debiasing method which first locates the bias-sensitive layer, then uses a decoding method through logit recalibration to debias at inference stage.

4.1 Locate the Bias-sensitive Layer

Lepori et al. (2025) finds that model bias is not uniformly distributed across layers but gradually emerges in intermediate layers. We refer to these intermediate layers, where biased priors are actively encoded and start to influence downstream computation, as bias-sensitive layers. After these layers, model predictions tend to collapse onto empirical priors, making later interventions ineffective.

For a model with n layers, exhaustively searching all source–target layer pairs incurs $O(n^2)$ complexity, which is expensive for LLM inference. Even restricting source and target to the same layer as Ghandeharioun et al. (2024) suggests, it still requires an $O(n)$ scan. We therefore design efficient quantitative metrics to identify bias-sensitive layers.

Logit Difference. To locate bias-sensitive layers, patching should be applied only after the attribute has been encoded in the model's internal representations. We therefore assess whether the model's representations already differentiate the context-aligned attribute from the biased one at each layer. Thus, we use logit-lens (Elhage et al., 2021) to

measure the logit difference (LD) between the secondary (context-aligned) and primary (biased) attributes at layer l. A larger positive difference indicates that the layer encodes evidence favoring the contextual attribute. Implementation details of this procedure are provided in Appendix B.1. We quantify this effect by averaging the LD over dataset $\mathcal{D}$.

$$LD_l = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}_i \in \mathcal{D}} \left[(h^{(l)} W_o^\top)_{a_{i,\text{sec}}} - (h^{(l)} W_o^\top)_{a_{i,\text{pri}}} \right]$$

Gradient Similarity Alignment. If patching is applied at layers close to the output, the model may have already committed to empirical priors, making bias difficult to override (Lepori et al., 2025). To locate bias-sensitive layers where patched information still affects model decisions, we measure the alignment between the output gradient and the attribute direction, which indicates whether the model relies on biased or contextual information.

A linear probe is trained to find the attribute direction w. More details about this are shown in Appendix B.2. Then we patch hidden representation $h^{(l)}(n)$ into target prompt T, the probability to generate correct color c is $p(c \mid T^l)$. Let the gradient of the patched hidden representation be g_i, then the gradient similarity alignment (GSA) score for layer l is defined as the dataset-averaged absolute cosine similarity between the gradient direction and the attribute direction:

$$GSA_l = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}_i \in \mathcal{D}} \left| \frac{g_i^\top w}{\|g_i\|_2 \, \|w\|_2} \right|, \quad g_i = \frac{\partial \log p(c|T^l)}{\partial h^{(l)}(n)}.$$

Layer Selection. After analyzing the individual effectiveness of LD and GSA, we integrate the two metrics to compute a final score for each layer. To justify their combination, we analyze the statistical relationship between the two metrics. Using Llama3.2-1b as an example, we compute the Spearman correlation coefficient (Spearman, 1961) between LD and GSA, obtaining $\rho = -0.147$ with $p < 0.001$, which indicates a weak correlation. To further characterize their relationship, we fit an isotonic regression (Miles, 1959), shown as the solid blue curve in Figure 3. The deviation from this idealized curve further suggests that LD and GSA are complementary, supporting their joint use in layer selection. The definitions of Spearman correlation and isotonic regression are provided in Appendix B.3.1 and B.3.2. The correlation analysis for additional models is reported in Appendix B.3.3. After normalizing LD_l and GSA_l

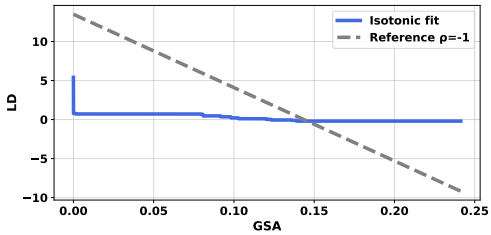

Figure 3: Isotonic Regression of GSA and LD on Llama3.2-1b

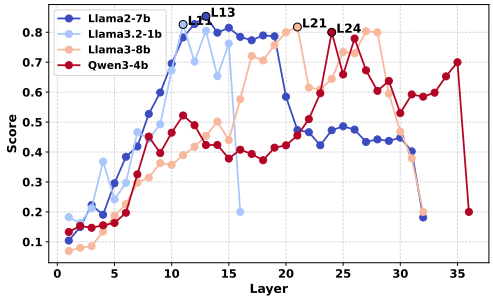

Figure 4: The selected best layer of each model for patching. Layer index starts from 1 with w=0.8.

to the $[0, 1]$ range, we combine them into a single layer-selection objective score and choose the optimal patching layer as

$$l^\star = \arg\max_{l \in \mathcal{L}} \Big(\beta \, LD_l + (1 - \beta) \, GSA_l \Big),$$

where β controls the relative importance of representation-level encoding and gradient-level utilization. We set $\beta = 0.8$ for all models, which yields the best empirical performance; the selection of β is determined via grid search, with results reported in Appendix E.1, and the layer scores and best layers for four models are shown in Figure 4.

4.2 Debias through Logit Recalibration

In the LDLoR's debiasing stage, we use a decoding method through logit recalibration. A contrastive prompt pair $\langle T, T^* \rangle$ is built, where T is the target prompt following the designation of Ghandeharioun et al. (2024), enabling an expressive decoding of a wide range of features. T^* is the contrastive prompt, which replaces the placeholder token in T with the original token in the source prompt that needs to be explained. Thus, the token in the contrastive prompt provided no contextual information, and the model can only explain the token with empirical knowledge and model bias. From T^*, we can obtain logits $l^* := l_\theta(y \mid T^*)$ that only contain model priors. To emphasize the contribution

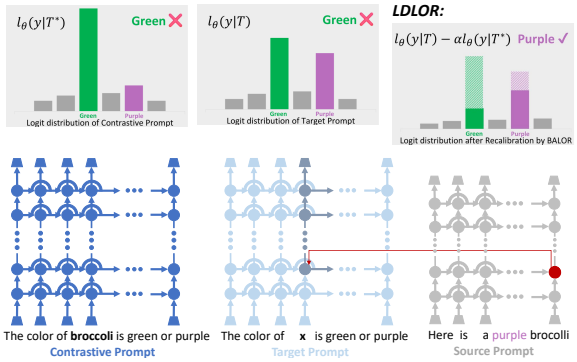

Figure 5: An overview of LDLoR's debiasing method. We use the same example with Figure 1, while green is a biased attribute value and purple is a correct attribute value. With the recalibration approach between the logit distributions of contrastive prompt and target prompt in LDLoR process, the LLM provides a correct response.

of patched information, we aim to mitigate the undesired model bias highlighted by the contrastive prompt and allow the model to generate a response based on the remaining information when provided with the injected hidden representations. Thus, we recalibrate the logit distribution:

$$p_{\text{ldlor}}(y \mid T) = \sigma\left[(1+\alpha)l_\theta(y \mid T) - \alpha l_\theta(y \mid T^*)\right],$$

σ is the softmax function, and α is a hyperparameter that denotes the amplification level of the recalibration method.

Figure 5 provides an intuitive illustration of how LDLoR operates through logit recalibration with an example. As shown in the left panel, when a contrastive prompt T^*, *"The color of broccoli is green or purple?"* is used without any contextual information, and the model exhibits a strong prior preference for the biased attribute value (*green*). This preference reflects the model's internal bias induced by imbalanced linguistic patterns. The middle panel shows that under the target prompt T, *"The color of X is green or purple?"*, the contextual information is encoded in the patched hidden representation. Although the context-aligned attribute value (*purple*) receives a higher logit than that in the contrastive prompt, its logit remains lower than that of the biased attribute value (*green*). As a result, the response is still unfaithful. LDLoR addresses this issue by isolating the bias component encoded in the unpatched prompt. As illustrated in the right panel, LDLoR subtracts a scaled bias logit $l_\theta(y \mid T^*)$ from the target logit $l_\theta(y \mid T)$, yielding a recalibrated logit distribution. Consequently, the biased preference is suppressed, and the context-aligned attribute value (*purple*) is amplified, as visualized by the dashed regions in the right panel. This logit-level adjustment effectively realigns the model's decision with the injected contextual information without modifying model parameters or internal representations. We then provide a theoretical characterization of LDLoR in the log-odds space, which is listed in Appendix C due to page limitations. It shows that if a candidate token is more supported by the patched token evidence than by the contrastive prior, LDLoR monotonically increases its relative preference as the scaling factor α grows.

While Figure 5 depicts a single decoding step, LDLoR is applied iteratively during autoregressive next-token prediction with the target and contrastive prompts decoding in parallel at each step. There are two modes of LDLoR, **S**hared and **D**ivided. In mode S, the recalibrated distribution $p_{\text{ldlor}}(y \mid T)$ is used to sample the next token for both the target prompt and the contrastive prompt, keeping their generated sequences synchronized. In mode D, $p_{\text{ldlor}}(y \mid T)$ is applied only to the target prompt for next-token prediction, while the contrastive prompt continues decoding with its original logits. This design allows LDLoR to flexibly control whether bias correction is enforced jointly across both prompts or applied solely to the target prompt during generation.

5 Experiments

This section introduces the experimental setup in Section 5.1, and then discusses LDLoR's overall performance in Section 5.2, and the ablation study in Section 5.3.

5.1 Settings

Datasets and Models. We use the biased part of the dataset across four tasks introduced in Section 3.1, fixing the sampling temperature to zero as the main result. In the ablation study, we use the biased part of the color task. In the sampling temperature ablation, we set Top-k=50 and Top-p=0.9. The models used in all experiments are those of Section 3.3.

Baselines. We compared LDLoR with six baselines: Vanilla, which means the original Patchscopes without any debiasing; Closed-Book (CB), Internal-Evidence (IE), and Data-Bound (DB), three prompt-guidance approaches to guide LLMs only use contextual information (Amplayo et al.,

2023); SFT (Rafailov et al., 2023) and FairSteer (Li et al., 2025). The implementation details of these baselines are provided in Appendix D.

Prompts under Zero-shot and Few-shot Settings. In both settings, the source prompt S is `Here is an` $\{a_{i,sec}\}$ $\{noun_i\}$. The target prompt T for zero-shot setting is `The {task} of x is` $\{a_{i,pri}\}$ `or` $\{a_{i,sec}\}$? The target prompt T for n-shot setting is `The {task} of` $noun_1$ `is` $\{a_{1,sec}\}$`... the {task} of` $noun_n$ `is` $\{a_{n,sec}\}$`, the {task} of x is`, where $1...n$ are are distinct from i. Contrastive prompts T^* for both settings are replaced `x` with $noun_i$. For the gender task, target prompts are modified to bypass gender alignment, details shown in Appendix G.2.

5.2 Main result analysis

The results of the baseline methods under zero-shot and few-shot settings on biased datasets are reported in Table 3 and Table 4. Across all four biased tasks, LDLoR achieves the best overall performance.

Overall, zero-shot settings tend to yield higher SR than few-shot settings, which is caused by task formulation. Zero-shot settings are framed as binary-choice questions with a constrained output space, whereas few-shot settings require free-form generation conditioned on in-context examples, which substantially expands the space of possible attribute values. This effect is more pronounced for tasks with a large attribute value space, such as color and culture, where open-ended generation introduces greater linguistic variability. In contrast, for tasks with relatively constrained attribute spaces, such as gender and age, the performance gap between zero-shot and few-shot settings is less evident.

The three prompt-based methods, CB, IE, and DB, can sometimes improve the SR compared to vanilla Patchscopes. In particular, IE achieves the highest SR on the gender task for Llama2-7b and Llama3-8b under few-shot settings. However, their performance is highly unstable, and they frequently underperform the vanilla baseline. Moreover, determining an optimal prompt is difficult, as different prompts perform better across different tasks, resulting in inconsistent outcomes. Similarly, SFT methods do not consistently outperform vanilla Patchscopes. Their performance degrades substantially when the biased dataset used for fine-tuning conflicts with the evaluation data, often leading to much lower SR than the vanilla baseline. FairSteer becomes more pronounced when the debiasing direction is inaccurate. In the gender task under zero-shot settings, FairSteer performs worse than the vanilla baseline across all four models.

In contrast, LDLoR performs debiasing at the decoding stage by modifying the logit distribution, thereby avoiding the limitations of prompt-based and fine-tuning-based methods. Under both zero-shot and few-shot settings, LDLoR consistently outperforms vanilla Patchscopes and achieves better performance than all six baselines in almost all cases. These results demonstrate that LDLoR attains both strong performance and high consistency, validating its effectiveness as a more faithful mechanism for hidden representation explanation.

We also observe that mode S is more stable than mode D. While mode S consistently outperforms all six baselines, mode D occasionally yields a lower SR. This instability arises because the target and contrastive prompts in mode D follow different next-token prediction strategies, causing the two input sequences to diverge progressively. Thus, the resulting logit distributions reflect not only differences in the hidden representations, but also discrepancies introduced by newly generated tokens. This divergence reduces the stability of mode D, although it can sometimes lead to higher SR.

5.3 Ablation Study

We also provide the ablation study on the effect of hyperparemeter α and sampling temperature. The detailed results are reported in Appendix E.2 and Appendix E.3.

6 Related Work

LLM Interpretability. Traditional interpretability methods, such as probing classifiers (Alain and Bengio, 2018; Belinkov, 2022), aim to measure whether a particular linguistic or semantic property can be linearly predicted from intermediate activations. More recent studies have explored interpretability by translating hidden representations into natural language. Belrose et al. (2025) use tuned lens to reconstruct the semantics of hidden vectors by projecting them back into interpretable token spaces. Building upon this direction, the Patchscopes framework (Ghandeharioun et al., 2024) introduces a modular approach to interpreting hidden representations by decoding them through a model's generative process.

Task	Model	Vanilla	CB	IE	DB	Fine-tuning	FairSteer	LDLoR (S)	LDLoR (D)
Color	Llama2-7b	37.05%	48.70%	49.09%	47.59%	40.36%	40.36%	**49.39%**	49.39%
	Llama3.2-1b	39.10%	43.17%	58.58%	62.94%	21.80%	38.81%	**71.66%**	48.84%
	Llama3-8b	23.40%	32.40%	24.43%	24.32%	43.48%	32.29%	**56.42%**	39.23%
	Qwen3-4b	48.20%	50.79%	46.61%	46.21%	51.70%	51.49%	73.45%	**74.80%**
Gender	Llama2-7b	24.06%	36.04%	41.70%	37.92%	41.02%	22.64%	**46.42%**	43.70%
	Llama3.2-1b	3.33%	46.67%	9.99%	26.64%	3.33%	0.00%	59.94%	**63.27%**
	Llama3-8b	0.00%	50.00%	50.00%	47.56%	32.93%	0.00%	54.39%	**51.71%**
	Qwen3-4b	34.78%	2.17%	8.70%	2.17%	48.58%	31.13%	**76.09%**	58.26%
Culture	Llama2-7b	36.84%	37.72%	43.49%	42.11%	38.60%	17.54%	**54.39%**	45.61%
	Llama3.2-1b	35.87%	29.35%	35.87%	39.13%	4.35%	45.65%	**57.61%**	40.43%
	Llama3-8b	9.68%	8.87%	10.48%	22.58%	18.55%	6.45%	**39.52%**	37.90%
	Qwen3-4b	45.69%	43.96%	43.73%	47.41%	54.03%	50.09%	**74.14%**	70.16%
Age	Llama2-7b	50.00%	46.34%	58.54%	60.98%	58.54%	43.90%	**69.51%**	67.07%
	Llama3.2-1b	50.00%	50.00%	47.22%	50.00%	47.78%	41.46%	50.00%	**91.67%**
	Llama3-8b	67.24%	50.00%	39.66%	50.00%	65.56%	44.83%	70.69%	**72.41%**
	Qwen3-4b	56.67%	56.10%	50.00%	35.37%	70.76%	56.10%	78.05%	**98.78%**

Table 3: Compare LDLoR with baselines under zero-shot settings. **Bold** and underline denote the highest and second-highest SR.

Task	Model	Vanilla	CB	IE	DB	Fine-tuning	FairSteer	LDLoR (S)	LDLoR (D)
Color	Llama2-7b	5.42%	7.83%	3.01%	1.51%	18.37%	4.82%	**52.10%**	48.19%
	Llama3.2-1b	14.24%	21.07%	8.14%	15.12%	3.78%	19.77%	**69.77%**	65.99%
	Llama3-8b	10.97%	6.63%	8.70%	8.60%	41.72%	14.50%	**47.62%**	43.47%
	Qwen3-4b	18.41%	13.66%	14.44%	15.50%	29.17%	22.09%	**68.90%**	67.54%
Gender	Llama2-7b	34.91%	69.34%	**82.08%**	80.19%	36.23%	36.80%	37.74%	40.57%
	Llama3.2-1b	6.67%	16.67%	10.00%	13.33%	13.33%	10.00%	26.67%	**36.67%**
	Llama3-8b	12.20%	28.05%	**34.15%**	17.07%	1.22%	21.95%	29.27%	28.05%
	Qwen3-4b	26.08%	17.39%	15.21%	21.74%	17.39%	21.74%	**41.30%**	39.13%
Culture	Llama2-7b	10.53%	14.03%	9.65%	14.91%	8.77%	8.77%	55.26%	**59.64%**
	Llama3.2-1b	0.00%	2.17%	9.78%	10.87%	8.70%	13.04%	**40.65%**	40.21%
	Llama3-8b	14.52%	4.03%	8.06%	4.03%	1.61%	12.90%	**37.09%**	35.32%
	Qwen3-4b	8.62%	4.31%	9.48%	8.62%	27.59%	12.07%	**62.07%**	57.76%
Age	Llama2-7b	56.10%	42.68%	48.78%	39.02%	59.76%	51.22%	**80.48%**	74.39%
	Llama3.2-1b	0.00%	19.44%	13.89%	2.78%	16.67%	55.56%	**63.89%**	58.33%
	Llama3-8b	17.24%	31.03%	27.58%	34.48%	10.34%	39.93%	**89.65%**	84.48%
	Qwen3-4b	39.02%	40.24%	51.21%	60.97%	41.46%	46.34%	**86.58%**	79.26%

Table 4: Compare LDLoR with baselines under few-shot settings. **Bold** and underline denote the highest and second-highest SR.

Debiasing LLMs. The LLMs exhibit modality preferences and inherent biases (Shrawgi et al., 2024; Gallegos et al., 2024a; Zhang et al., 2025), motivating the development of debiasing methods. Debiasing methods can be grouped into two main categories: (1) Fine-tuning-based methods, which mitigate bias in LLMs by retraining the model on a well-constructed dataset (Chen et al., 2023). The retraining methods such as contrastive learning (Zhou et al., 2022; Dong et al., 2023; Wu et al., 2025) and reinforcement learning (Allam, 2024; Qureshi et al., 2024) and are usually combined with a parameter-efficient-fine-tuning technique such as LoRA (Hu et al., 2022) and BitFit (Zaken et al., 2022). (2) Inference-time intervention methods intervene in the model at the inference stage. Prompt-based intervention methods use carefully designed prompts to lead LLMs toward unbiased outputs during generation (Dong et al., 2023; Oba et al., 2024; Gallegos et al., 2024b; Li et al., 2025). Other methods, such as Fairsteer (Li et al., 2025), learn a steering vector and then use it to adjust biased activation during inference.

7 Conclusion

In this work, we show that imbalanced linguistic patterns will lead to model bias in LLMs, and this kind of model bias makes the interpretation tool Patchscopes unfaithful. We built a dataset to investigate this, and we show that the model bias induced by imbalanced linguistic patterns overrides contextual information, which leads to unfaithfulness. Building on this insight, we propose LDLoR, a locate-then-debias strategy to mitigate this, enhancing the faithfulness of Patchscopes.

Limitations

The LDLoR builds on the Patchscopes framework, so its effectiveness is inherently influenced by the quality of the patched representations and may vary when these representations are noisy or only partially informative. Additionally, current experiments focus on relatively simple attribute settings, and further evaluation on more complex and compositional scenarios would help assess its broader applicability.

AI Assistants

We used ChatGPT to support writing and coding. Specifically, AI assistance was used for language polishing, grammar checking, improving clarity, and suggesting code snippets or debugging ideas. AI assistants were not used to generate experimental results, fabricate citations, replace human analysis, or make final research decisions. All AI-assisted content were reviewed, verified, and edited by the authors.

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Ahmed Allam. 2024. BiasDPO: Mitigating bias in language models through direct preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 42–50, Bangkok, Thailand. Association for Computational Linguistics.

Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das, and Shashi Narayan. 2023. Query refinement prompts for closed-book long-form QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7997–8012, Toronto, Canada. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob https://arxiv.org/help/api/indexSteinhardt. 2025. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

Milan Bhan, Jean-Noel Vittaut, Nicolas Chesneau, Sarath Chandar, and Marie-Jeanne Lesot. 2025. Did i faithfully say what i thought? bridging the gap between neural activity and self-explanations in large language models. *Preprint*, arXiv:2506.09277.

Ari Brill. 2024. Neural scaling laws rooted in the data distribution. *arXiv preprint arXiv:2412.07942*.

Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. 2023. Fast model debias with machine unlearning. *Preprint*, arXiv:2310.12560.

Xiangjue Dong, Ziwei Zhu, Zhuoer Wang, Maria Teleki, and James Caverlee. 2023. Co^{2}PT: Mitigating bias in pre-trained language models through counterfactual contrastive prompt tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5859–5871, Singapore. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

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

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024b. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *Preprint*, arXiv:2402.01981.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. *Preprint*, arXiv:2401.06102.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *Proceedings of the 2024 International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Jingyu Hu, Shu Yang, Xilin Gong, Hongming Wang, Weiru Liu, and Di Wang. 2025. Monica: Real-time monitoring and calibration of chain-of-thought sycophancy in large reasoning models. *Preprint*, arXiv:2511.06419.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and 1 others. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596, Tokyo, Japan. Association for Computational Linguistics.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Michael A. Lepori, Michael Curtis Mozer, and Asma Ghandeharioun. 2025. Racing thoughts: Explaining contextualization errors in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3020–3036, Albuquerque, New Mexico. Association for Computational Linguistics.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. 2025. Fairsteer: Inference time debiasing for llms with dynamic activation steering. *Preprint*, arXiv:2504.14492.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Meta. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

RE Miles. 1959. The complete amalgamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika*, 46(3/4):317–327.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2024. In-contextual gender bias suppression for large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1722–1742, St. Julian's, Malta. Association for Computational Linguistics.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Rameez Qureshi, Naïm Es-Sebbani, Luis Galárraga, Yvette Graham, Miguel Couceiro, and Zied Bouraoui. 2024. Refine-lm: Mitigating language model stereotypes via reinforcement learning. *arXiv preprint arXiv:2408.09489*.

Qwen. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857, St. Julian's, Malta. Association for Computational Linguistics.

Charles Spearman. 1961. The proof and measurement of association between two things.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *Preprint*, arXiv:2004.12265.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

Shushan Wu, Xilin Gong, Yongkai Chen, Yufang Liu, Wenxuan Zhong, and Ping Ma. 2025. Fisher contrastive learning: A robust solution to the feature suppression effect.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9.

Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. 2025. When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms. *Preprint*, arXiv:2511.02243.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130, Dublin, Ireland. Association for Computational Linguistics.

A Casual Analysis on Other Models

In this section, we provide the causal analysis results on all four models we used. Across all models, the logistic regression results show a consistent positive association between the frequency gap Δf(defined in Section 3.3) and the likelihood of color bias, As summarized in Table 5, one standard deviation increase in Δf leads to a substantial rise in the odds of exhibiting bias, with odds ratios (OR) ranging from 5.726 (Llama2-7B) to 9.338 (Llama3.2-1B). The corresponding logit coefficients are all positive and statistically significant, confirming that training frequency imbalance meaningfully predicts model bias. Furthermore, the ROC-AUC values between 0.579 and 0.619 indicate moderate but consistent discriminative power across models.

Model	Metric	Mean Value	95% CI
Llama2-7b	Logit coeff (+1 SD)	1.745	(0.866, 2.624)
	OR (+1 SD)	5.726	(2.376, 13.796)
	ROC-AUC	0.606	(0.549, 0.659)
Llama3.2-1b	Logit coeff (+1 SD)	2.234	(1.487, 2.981)
	OR (+1 SD)	9.338	(4.425, 19.706)
	ROC-AUC	0.619	(0.580, 0.655)
Llama3-8b	Logit coeff (+1 SD)	1.894	(1.295, 2.493)
	OR (+1 SD)	6.645	(3.651, 12.097)
	ROC-AUC	0.579	(0.546, 0.609)
Qwen3-4b	Logit coeff (+1 SD)	1.866	(1.312, 2.420)
	OR (+1 SD)	6.464	(3.715, 11.247)
	ROC-AUC	0.599	(0.569, 0.629)

Table 5: Repeated undersampling results for each model.

B Details about Locating the Bias-sensitive Layer

B.1 Logit Difference

To verify that a given layer indeed reflects the model's understanding of the input, we append an interrogative clause "What {attribute} is {noun}?" to the source prompt. This addition explicitly forces the model to process the attribute associated with the noun, thereby embedding the desired semantic information within the evolving hidden representations.

After injecting this clause, we use logit-lens to measure how the model internally differentiates the secondary attribute, which is shown less in linguistic pattern but provided in contextual information, from the primary attribute, which is the biased attribute from linguistic pattern at layer l. We examine the logit values difference of the secondary and primary attributes, where the logit value is computed by mapping the hidden representation at layer l to the vocabulary space using the unembedding matrix W_o. This reflects how strongly the layer-level hidden representation favors the secondary attribute over the primary one. Intuitively, if a layer meaningfully encodes the color information carried by the patched representation, the logit associated with the secondary color should be higher than that of the primary color.

B.2 Linear Probing

We train a multi-class linear probe that can classify the noun tokens into different colors on a general dataset. $W \in \mathbb{R}^{C\times d}$ is the probe's weight matrix for C color classes when d is the dimension of hidden representations. After training, each row vector W_c corresponds to a learned attribute direction of color c (Kim et al., 2018). Thus, this attribute direction w can be directly taken as $w = W_c \in \mathbb{R}^d$.

B.3 Correlation Analysis

B.3.1 Spearman Correlation

The Spearman correlation coefficient ρ is a non-parametric measure of rank correlation that evaluates the strength and direction of a monotonic relationship between two variables (Spearman, 1961). Given paired observations (x_i, y_i) for $i = 1, \ldots, n$, we first convert them to ranks $R(x_i)$ and $R(y_i)$, and then compute:

$$\rho = \frac{\sum_{i=1}^{n}(R(x_i) - \overline{R_x})(R(y_i) - \overline{R_y})}{\sqrt{\sum_{i=1}^{n}(R(x_i) - \overline{R_x})^2}\sqrt{\sum_{i=1}^{n}(R(y_i) - \overline{R_y})^2}},$$

where $\overline{R_x} = \frac{1}{n}\sum_{i=1}^{n} R(x_i)$ and $\overline{R_y} = \frac{1}{n}\sum_{i=1}^{n} R(y_i)$. Alternatively, when there are no tied ranks, it can be expressed as

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad d_i = R(x_i) - R(y_i).$$

The coefficient ranges from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with $\rho = 0$ indicating no monotonic association.

B.3.2 Isotonic Regression

Isotonic regression (Miles, 1959) is a non-parametric technique for fitting a monotonic (non-decreasing) function to a set of paired observations (x_i, y_i). It estimates values $\hat{f}_i$ by solving

$$\min_{f_1 \le f_2 \le \cdots \le f_n} \sum_{i=1}^{n} w_i (y_i - f_i)^2,$$

subject to the ordering constraints $f_i \le f_{i+1}$, where w_i are optional non-negative weights. This problem is efficiently solved using the Pool Adjacent Violators Algorithm (PAVA), which produces a piecewise-constant fit that enforces the desired monotonic relationship.

B.3.3 Correlation Analysis Results

The results in Table 6 demonstrate that these two metrics, LD and GSA, do not have a monotonic relationship since ρ is close to zero with a small p-value. We further characterize the correlation of LD and GSA with isotonic regression, and the fitted curves are shown in Figure 6. These results further demonstrate that these two metrics are independent.

Model	Spearman ρ	p-value
Llama2-7B	0.010	2.491×10^{-1}
Llama3.2-1B	-0.147	4.852×10^{-33}
Llama3-8B	-0.006	5.023×10^{-1}
Qwen3-4B	-0.040	9.281×10^{-6}

Table 6: Spearman correlation between Logit Difference and Gradient Similarity Alignment across models.

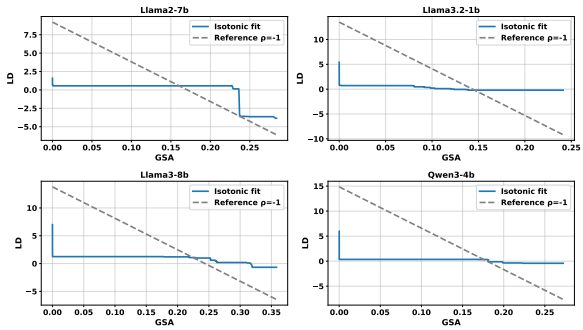

Figure 6: Isotonic regression results across models. The gray line is the reference curve while the blue line is the fitted curve.

C Theorem and Proof

Theorem C.1 shows that if a candidate token is more supported by the patched token evidence than by the contrastive prior, LDLoR monotonically increases its relative preference as the scaling factor α grows. This effect is linear in α in log-odds space, ensuring that sufficiently large α promotes the context-consistent token to the highest probability.

Theorem C.1 (Log-odds contrastive amplification of LDLoR)**.** *Let $l_\theta(y \mid T)$ and $l_\theta(y \mid T^*)$ be the (pre-softmax) logits over vocabulary $\mathcal{V}$. Define LDLoR as*

$$p_{ldlor}(y \mid T) = \sigma((1+\alpha)\, l_\theta(y \mid T) - \alpha\, l_\theta(y \mid T^*)),$$

where $\alpha \ge 0$. Then, for any two tokens $y_1, y_2 \in \mathcal{V}$,

$$\log \frac{p_{ldlor}(y_1 \mid T)}{p_{ldlor}(y_2 \mid T)} = \log \frac{p_\theta(y_1 \mid T)}{p_\theta(y_2 \mid T)} + \alpha \left(\log \frac{p_\theta(y_1 \mid T)}{p_\theta(y_2 \mid T)} - \log \frac{p_\theta(y_1 \mid T^*)}{p_\theta(y_2 \mid T^*)} \right),$$

where $p_\theta(\cdot \mid T) = \sigma(l_\theta(\cdot \mid T))$ and $p_\theta(\cdot \mid T^*) = \sigma(l_\theta(\cdot \mid T^*))$. Equivalently,

$$\log \frac{p_{ldlor}(y_1 \mid T)}{p_{ldlor}(y_2 \mid T)} = (1+\alpha)(l_\theta(y_1 \mid T) - l_\theta(y_2 \mid T)) - \alpha(l_\theta(y_1 \mid T^*) - l_\theta(y_2 \mid T^*)).$$

In particular, if y_1 is more supported by the patched target evidence than by the contrastive prior, $\log \frac{p_\theta(y_1|T)}{p_\theta(y_2|T)} > \log \frac{p_\theta(y_1|T^*)}{p_\theta(y_2|T^*)}$, *then the LDLoR preference for y_1 over y_2 increases monotonically with α.*

Proof.

$$p_{\text{ldlor}}(y \mid T) \propto \exp\left((1+\alpha) l_\theta(y \mid T) - \alpha l_\theta(y \mid T^*)\right).$$

Therefore, for any y_1, y_2,

$$\log \frac{p_{\text{ldlor}}(y_1 \mid T)}{p_{\text{ldlor}}(y_2 \mid T)} = (1+\alpha)(l_\theta(y_1 \mid T) - l_\theta(y_2 \mid T)) - \alpha(l_\theta(y_1 \mid T^*) - l_\theta(y_2 \mid T^*)),$$

which gives the second equality. For the first equality, note that softmax log-odds satisfy $\log \frac{\sigma(l)(y_1)}{\sigma(l)(y_2)} = l(y_1) - l(y_2)$, so substituting $p_\theta(\cdot \mid T) = \sigma(l_\theta(\cdot \mid T))$ and similarly for T^* yields the stated decomposition. Finally, the monotonicity follows because the right-hand side is affine in α with slope $\log \frac{p_\theta(y_1|T)}{p_\theta(y_2|T)} - \log \frac{p_\theta(y_1|T^*)}{p_\theta(y_2|T^*)}$, which is positive under the condition.

D Baselines

D.1 Prompting

Prompt-guidance intervention approach is the easiest to implement way for model debiasing. We design three distinct guidance prompts that serve as prefix instructions to guide the model in generating non-biased responses based on provided contextual information only (Amplayo et al., 2023).

Closed-Book: Do not use outside or empirical knowledge, directly answer. <target prompt >.
Internal-Evidence: Use only internal evidence from the input, exclude background knowledge. <target prompt >.
Data-Bound: Answer strictly from the given data, avoid any external reasoning. <target prompt >.

D.2 Fine-tuning

In order to prevent the effect of model bias, fine-tuning is a well-known debiasing method. We use Direct Preference Optimization (DPO) (Rafailov et al., 2023) combined with Low-Rank adaptation (LoRA) (Hu et al., 2022) here to perform lightweight preference fine-tuning on Llama2-7b, Llama3.2-1b, Llama3-8b, and Qwen3-4b. To reduce model bias, we adopt 1,507 pairs of biased and unbiased data from (Nangia et al., 2020), of which 95% are used for training and 5% are used for validation. This encourages the model to prefer more fairness and faithful responses.

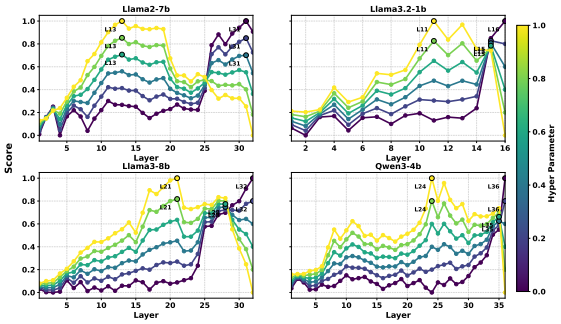

Figure 7: Layer-wise score progression across different models under varying hyperparameter β. Each curve represents a different β value, and the color of the curve indicates the value of β according to the color bar on the right.

D.3 FairSteer

FairSteer is a vector-steering intervention method. Following (Li et al., 2025), we first trained a steering vector to capture the overall direction from biased case to unbiased case. During the inference stage, we modify the midlayer activation before the patched layer along the steering vector's direction by adding the steering vector on the activation.

E Ablation Study

E.1 Hyperparameter β for Layer Locating.

The ablation study of hyper parameter β is shown in Figure 7. The best layer under each β is highlighted in the figure. Across all β values, there are only 2 layers as candidates for Llama2-7b, and 3 layers as candidates for Llama3.2-1b, Llama3-8b, and Qwen3-4b. As stated in (Lepori et al., 2025) and (Ghandeharioun et al., 2024), biases or misinterpretations often arise in the middle layers of transformer models. Thus, we set $w = 0.8$ for all models and select layer 13 for Llama2-7b, layer 11 for Llama3.2-1b, layer 21 for Llama3-8b and layer 24 for Qwen3-4b.

E.2 Hyperparemeter α for Debiasing.

As shown in Figure 8, LDLoR consistently yields positive SR gains over the vanilla Patchscopes baseline across all four models for a wide range of α values. The Shared mode achieves stable improvements throughout $\alpha \in [0.4, 4.0]$, achieving sub-

stantial peak gains of 12–33%, while maintaining positive performance as α increases. The Divided mode exhibits more variability and smaller gains. Although the relative gains are negative on $\alpha = 2.8$ for Llama2-7b and $\alpha \geq 0.8$, but still improve over the vanilla baseline for all other α settings on four models. While mode S is more stable, we recommend using this Shared mode in the application.

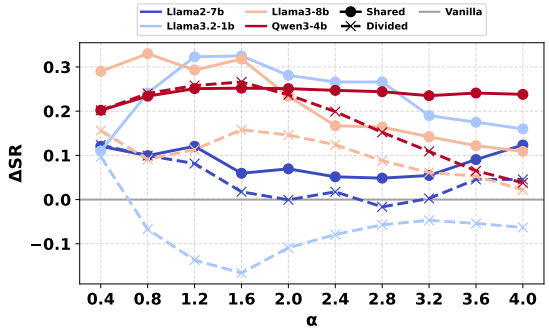

Figure 8: Relation SR gain by hyberparameter α of four models compared with vanilla Patchscopes.

E.3 Sampling Temperature.

We further evaluate LDLoR under varying sampling temperatures, which control inference randomness, ranging from 0.0 to 1.6 for all LLMs (Figure 9). Sampling temperature has little impact on the SR of mode S, whereas mode D exhibits a slight upward trend. This indicates that mode S remains stable across temperatures, while mode D benefits from higher temperatures, as divergence between prompt pairs means the optimal token may not always have the highest post-recalibration probability. For statistical analysis, we apply the Kruskal-Wallis test (Kruskal and Wallis, 1952) to the SR distributions. The definition and result of the Kruskal-Wallis test are provided in Appendix F.

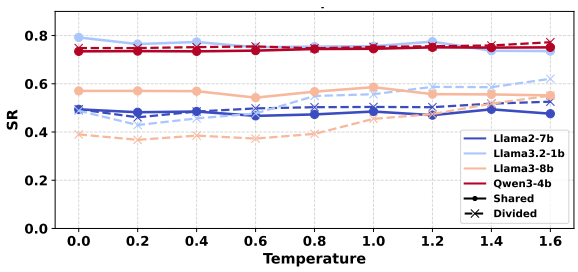

Figure 9: SR by sampling temperature of four models for LDLoR.

F Kruskal-Wallis test

F.1 Defination

The Kruskal-Wallis test (Kruskal and Wallis, 1952) is a non-parametric statistical test used to determine whether multiple independent groups originate from the same distribution. It generalizes the Wilcoxon rank-sum test (Wilcoxon, 1992) to the case of k groups.

Suppose we have k independent groups $G_1, G_2, \ldots, G_k$ with group sizes $n_1, n_2, \ldots, n_k$, and a total sample size $N = \sum_{j=1}^{k} n_j$. At the ranking step, all observations from all groups are pooled together and assigned ranks R_{ij}, where R_{ij} is the rank of the i-th observation in group j after sorting the entire dataset. The sum of ranks for each group is$T_j = \sum_{i=1}^{n_j} R_{ij}$.

Then the Kruskal–Wallis statistic H is computed as

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{T_j^2}{n_j} - 3(N+1).$$

When tied ranks exist, a tie-adjustment factor is applied to ensure validity. Under the null hypothesis

$$H_0 : \text{all } k \text{ groups come from the same distribution.}$$

The statistic H follows an asymptotic chi-square distribution $H \sim \chi^2_{k-1}$.

The decision rule is that given the computed statistic H and degrees of freedom $k-1$, the p-value is

$$p = 1 - F_{\chi^2_{k-1}}(H),$$

where F is the chi-square CDF. A small p-value (< 0.05) indicates that at least one group differs in distribution from the others.

F.2 Kruskal-Wallis test Result

The result is shown in Table 7 and for all LLMs, p-value > 0.05 indicates that the result is non-significant and thus, sampling temperature does not introduce measurable distributional shifts in the SR of LDLoR (S). For LDLoR (D), while the SR shows a slight increase, the distribution also shifts during the sampling temperature increase. This result also supports the analysis presented in Section 5.3, regarding the performance differences between modes S and D, as well as the trend of SR varying with sampling temperature.

Model	H(9)	p-value
Mode S		
Llama2-7b	0.6584	0.9953
Llama3.2-1b	6.2894	0.3916
Llama3-8b	3.1877	0.7850
Qwen3-4b	1.8333	0.9344
Mode D		
Llama2-7b	7.6408	0.3653
Llama3.2-1b	113.6126	1.63×10^{-21}
Llama3-8b	140.8080	3.44×10^{-27}
Qwen3-4b	3.4114	0.8445

Table 7: Kruskal–Wallis test results by model for LD-LoR steering.

Model	Biased	Non-biased
Llama2-7b	166	980
Llama3.2-1b	344	802
Llama3-8b	483	663
Qwen3-4b	516	630

Table 8: Model bias statistics. *Hard* column shows the number of cases where the model chose the primary color under the hard prompt, *Easy* under the easy prompt, and *Bias* under both.

Task	Concept	Number
Gender	degree	38
	characteristic	30
	name	19
	occupation	19
Birthplace (Culture)	name	31
Religion (Culture)	name	31
Age	occupation	45

Table 9: Number of examples per relation in the implicit biases subset. Birthplace and Religion bias on name are merged as Culture bias in the main paper.

G Dataset

G.1 Color Task

G.1.1 Make datapoints for Color Task

The color task contains 1,146 product nouns and uses their related colors as attributes. The nouns are selected from the LVIS (Gupta et al., 2019) dataset. We only use the category names, which are 1,202 nouns in total. The colors related to these nouns are then extracted from the Redpajama subset (Weber et al., 2024). Based on the co-occurrence frequency of nouns and colors in the RedPajama dataset, we matched each noun with two colors: the most frequently co-occurring color (primiary color) and the second most frequently one (secondary color). Some nouns basically do not have color (e.g Internet, person, chap), or we only get "color not found". We manually removed 56 such data points and leaving 1,146 nouns have valid primary and secondary color.

G.1.2 Select Datapoints in the Color Task

We divide the color datapoints by conducting an *option-swapping QA task* to find the subset where the model shows bias toward primary color. Since LLMs have bias in multipul-choice task (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024), which will make them favor Option A, we use prompt "The color of a {noun} is {color$_{pri}$} or {color$_{sec}$}?" as Easy to select color$_{pri}$ prompt and "The color of a {noun} is {color$_{sec}$} or {color$_{pri}$}?" as Hard to select color$_{pri}$ prompt, we can filter the noun-color pairs which LLMs have strong empirical knowledge on. If it constantly selects {color$_{pri}$} instead of the first option, this means that they have a strong empirical bias on the color they chose. Thus, the biased subset can be used to test the performance of patchscopes in biased cases. When we patch these nouns with secondary color information, and the target model successfully reveals this color in the context it generated, this means that the model overcomes empirical bias and effectively translates the hidden representation. Table 8 shows the case statistics of each model.

G.2 Gender, Culture and Age tasks

G.2.1 Dataset Description

These datapoints are a subset of the dataset provided by (Hernandez et al., 2024). The dataset proposed by (Hernandez et al., 2024) contains four categories: factual associations, commonsense knowledge, implicit biases, and linguistic knowledge, while we only use the implicit biases subset. In this subset, there are seven relation and for easy statistics, we merge them into three tasks, gender, culture and age. Table 9 shows what is contained in each task and also the number of examples per relation.

Category	Prompts
Occupation	**Source:** This is an exceptional {occupation} {gender$_{sec}$}. **Target:** Replace ? with "She" or "He". ? is an exceptional {x}.
Name	**Source:** {Gender$_{sec}$} has the name {name}. **Target:** Replace ? with "She" or "He". ? has the name {x}.
Degree	**Source:** This is a {gender$_{sec}$} outstanding student in department of {degree}. **Target:** Replace ? with "She" or "He". ? stands out in the department of { x }.
Characteristic	**Source:** This {gender$_{sec}$} is a {characteristic} person. **Target:** Replace ? with "She" or "He". ? is a {x} person.

Table 10: The source and target prompts for different categories of gender biased tasks. The designed prompt is to jailbreak the sexism alignment.

G.2.2 Source and Target Prompts for Gender Biased Dataset

For the four subcategories of gender biased tasks, the prompts we use are shown in Table 10. For simplify notation, in source prompt, we only denote the gender which model does not toward to as gender$_{sec}$. When implemented, the word can be substituted with the pronoun that corresponds to the gender of the subject. In the occupation and characteristic category, gender$_{sec}$ is man or woman. In the name category, gender$_{sec}$ is She or He. In the degree category, gender$_{sec}$ is male or female. Also, the target prompt T shown in the table has "She" as the first option and "He" as the second. In experiments, we also forward the counterpart T', where the options' positions are reversed.

G.3 Examples for each task

In this section, we provide examples of the source and target prompts for each task. The noun we want to explain is highlighted with underline.

- **Color:** Source Prompt: *Here is a purple broccoli.* Target Prompt: *The color of x is green or purple?*
- **Gender:** Source Prompt: *This man is a compassionate person.* Target Prompt: *Replace z with He or She. z is a x person.*
- **Culture:** Source Prompt: *Here is a Jewish Ajahn.* Target Prompt: *The religion of x is Buddhism or Judaism?*
- **Age:** Source Prompt: *Here is an exceptional young mentor.* Target Prompt: *The age of x is old or young?*