TOWARDS LOCAL BIAS MITIGATION: DISCOVERING HOW LLMS TREAT LGBTQ+ USERS DIFFERENTLY, WITHOUT EXPLICIT MARKERS

Ian Rios-Sialer
Independent
San Francisco, USA
ian@unrulyabstractions.com

July 2026

ABSTRACT

The LGBTQ+ community encompasses a multiplicity of evolving and fluid expressions, but bias evaluations and mitigation strategies typically rely on explicit markers and narrow categories. In practice, LLMs continue to reproduce social biases that impact queer people. The harm often surfaces when deployed LLMs treat queer and cisheterosexual users differently. We propose to reveal these axes of differential treatment by exploiting ecological validity, grounding, and internal knowledge. We present a method to discover and measure distributional behavioral differences in specific deployment settings of interest. In our pipeline, we first use a helper LLM to propose interpretable axes along which two communities might be treated differently. Then we feed the target LLM with prompt samples from both communities and generate multiple responses per prompt. We then use an LLM-as-judge to score each response along the hypothesized axes of treatment difference. We finally compare the resulting behavior distributions across communities and identify where bias exists. Our work seeks to empower practitioners to implement their own local evaluations and social bias mitigation that consider the full spectrum of human expression.

1 Introduction

The LGBTQ+ community encompasses a multiplicity of evolving and fluid expressions [33, 25, 64, 17], but bias evaluations and mitigation strategies typically rely on explicit markers and fixed, narrow categories [28, 15, 13, 59, 6]. Deployed LLMs therefore continue to reproduce social biases [48, 43, 23, 56], and those people least captured by rigid categories [55], including queer people [2, 44, 58, 20], and the subaltern at large [52, 30], can be heavily impacted [17].

Queer people are *legible* to one another not through a shared trait but through recognition that each is also *othered* by the same societal systems [1, 14]. When thinking about the impact of social bias, we are concerned about how the minoritized are treated differently. In this work, we seek to answer the following question:

- **RQ1:** Can we uncover LLM differential treament towards towards LGBTQ+ community?

Prior work suggests that LLMs encode internal knowledge that appropriate grounding can elicit (Section 2).

Then, we hypothesize that:

- **H1:** Even though LLMs reproduce social biases, they can also be leveraged to uncover them.

In this paper, we propose a way to leverage LLMs' internal knowledge to discover and measure distributional behavioral differences in specific deployment settings of interest. Our work seeks to empower users to implement their own local evaluations and social bias mitigation that consider the full spectrum of human expression.

The key contributions from this work are:

- **C1:** Formalization of differential treatment discovery task (Section 3).
- **C2:** Proposal for a pipeline that, given prompt sets from different user communities, generates hypotheses about how LLMs respond differently and validates them by measuring distributional differences in behavior (Section 4).

This is a work in progress. Ongoing efforts are implementing and evaluating the proposed framework. The purpose of this **proposal** is to gather early feedback and invite collaborations.

2 Background

If LLMs reproduce and amplify social biases [17, 48], how could we leverage the LLMs themselves to uncover their own social biases? In this section, we review three insights from the literature that motivate our hypothesis that LLMs can be leveraged to discover and measure their own social biases beyond explicit marker evaluations.

2.1 Ecological validity and locality

An increasingly growing body of work advocates for greater **ecological validity** in LLM evaluations [22, 36, 8]: the degree to which the stimuli and experimental conditions reflect how the system is used in practice [12]. To understand LLM behavior, we need valid measurements in real contexts [57].

Furthermore, bias does not transfer cleanly across contexts: fairness is a property of the sociotechnical setting rather than of the model in isolation, so behavior that looks fair on a general benchmark can fail in a specific deployment [50, 37]. Evaluations must therefore be **local**, bound to the communities and domains in which the model is actually used.

We then contend that to study how LLMs treat two groups of people differently, we need enough ecologically valid prompt samples for each group in a constrained domain of interest.

2.2 Unlocking internal knowledge for hypothesis proposal

Although the current bias mitigations operate through explicit markers and often do not generalize to LLM behavior outside of evaluation [26, 5, 47, 29, 60, 24, 27], LLMs do acquire **internal knowledge** about what social biases and associated harms exist. Pre-trained base models can already recognize their undesirable biases [49]. Post-training further teaches LLMs to articulate their target behavioral policies [10, 11, 46], even if those policies fail to govern their behavior [60]. Consequently, LLMs have been used to hypothesize what attributes could plausibly influence a decision [4], and even author their own evaluation probes [45]. However, without context in the prompt, LLMs lack enough **grounding** to apply their internal knowledge [51, 40].

We therefore conjecture that when LLMs are constrained to a specific deployment situation, they can draw on their internal knowledge to generate good hypotheses about how two user groups might be treated differently.

2.3 Judging via recognition

Although LLM-as-judge methods approximate human evaluation on open-ended outputs [61], it is important to distinguish between the generative and discriminative capabilities of models. LLMs are better at recognizing social bias than at preventing it from shaping their own generations [49, 60]. Asking whether a given response exhibits social bias requires re-derivation, which turns the judge into another generator that must reconstruct the answer indirectly from observed behavior [54]. If we instead ground the judge in a local context and ask, without disclosing which community produced the prompt, whether a response possesses a **specific concrete property**, the judge is more reliable [19, 4].

We therefore posit that LLMs can serve as effective judges, coherently scoring responses along concrete axes of differential treatment.

3 Differential treatment discovery task

In this section, we formalize the Differential Treatment Discovery Task. The main goal of the task is to identify the most salient ways in which a given LLM treats two user groups differently in a specific setting or scenario.

In the next section (Section 4), we leverage the insights outlined earlier (Section 2) to propose a framework for solving the task presented here.

3.1 Prompts from the community

Consider two sets of prompts, a **target** set and a **baseline** set:

$$X^{\text{target}} \coloneqq \{x_0^{\text{target}}, x_1^{\text{target}}, ...\} \qquad X^{\text{baseline}} \coloneqq \{x_0^{\text{baseline}}, x_1^{\text{baseline}}, ...\} \tag{1}$$

Each prompt set corresponds to LLM usage by a particular community, and both sets correspond to the same domain.

We assume **instruction comparability**: both prompt sets make the same underlying requests with the same frequency, so any observed differences in LLM behavior cannot be attributed to differences in what was asked. Formally, let $\iota : \mathcal{X} \to \mathcal{I}$ map each prompt to its underlying instruction, the explicit intent of the request. For a prompt set X and an instruction $i \in \mathcal{I}$, let f_X denote the fraction of prompts in X whose underlying instruction is i:

$$f_X(i) \coloneqq \frac{|\{\, x \in X : \iota(x) = i \,\}|}{|X|} \tag{2}$$

Then, the instruction comparability assumption can be written as:

$$f_{X^{\text{target}}}(i) \;\approx\; f_{X^{\text{baseline}}}(i) \qquad \forall\, i \in \mathcal{I}, \tag{3}$$

which expresses that every instruction occurs with roughly the same frequency in both sets.

3.2 LLM behaviors

For any prompt text x, an LLM defines a distribution over outcome responses. Denoting $\mathcal{P}(\cdot)$ the space of probability distributions, we can write:

$$\text{LLM} : \mathsf{Str} \to \mathcal{P}(\mathsf{Str}) \qquad y \sim \text{LLM}(\cdot \mid x) \tag{4}$$

We denote **behavior characterization** to be an **interpretable** representation $\Lambda : \mathsf{Str} \to [0,1]^n$ of a string response $y \in \mathsf{Str}$. By 'interpretable', we require that each dimension of $\Lambda(y)$ can be articulated in natural language and is generally understandable. Each dimension of the behavior characterization defines an axis along which the two communities may be treated differently.

Consequently, the LLM also defines a distribution over behaviors $z \in [0,1]^n$. Applying the functor $\mathcal{P}$ to Λ yields the pushforward $\mathcal{P}(\Lambda)$, which lets us express this distribution as:

$$z = \Lambda(y) \qquad z \sim \mathcal{P}(\Lambda)\big(\text{LLM}(\cdot \mid x)\big) \tag{5}$$

We denote this distribution as the **prompt behavior distribution**:

$$\beta_x \coloneqq \mathcal{P}(\Lambda)\big(\text{LLM}(\cdot \mid x)\big) \tag{6}$$

We then consider a user community C that produces a set of prompts X^C. Taking each prompt in the set to be equally likely, we aggregate the per-prompt behavior distributions into a single mixture, the **community behavior distribution**:

$$\beta_{X^C} \;\coloneqq\; \frac{1}{|X^C|} \sum_{x \in X^C} \beta_x \tag{7}$$

Then, β_{X^C} is the distribution of LLM behaviors towards the community C.

3.3 Task definition

The two communities of interest give rise to a target and a baseline prompt set, and hence to two behavior distributions $\beta_{X^{\text{target}}}, \beta_{X^{\text{baseline}}} \in \mathcal{P}([0,1]^n)$. Because the two sets are instruction-comparable (Section 3.1), any gap between these distributions reflects differently *how* the model responds to each community.

The **Differential Treatment Discovery Task** is to find an interpretable characterization Λ under which this gap is as large as possible:

$$\Lambda^\star \;\coloneqq\; \arg\max_{\Lambda \in \mathcal{H}} \; \mathcal{D}\big(\beta^\Lambda_{X^{\text{target}}} \,,\; \beta^\Lambda_{X^{\text{baseline}}}\big)\,, \tag{8}$$

where $\mathcal{H}$ is a family of interpretable characterizations, the superscript in $\beta^\Lambda_{X^C}$ marks the dependence of each community's behavior distribution on the chosen Λ, and $\mathcal{D}$ measures how far apart two distributions on $[0,1]^n$ are. The axes along which $\beta^{\Lambda^\star}_{X^{\text{target}}}$ and $\beta^{\Lambda^\star}_{X^{\text{baseline}}}$ diverge most are the candidate social biases: each is an interpretable dimension on which the model treats the two communities differently.

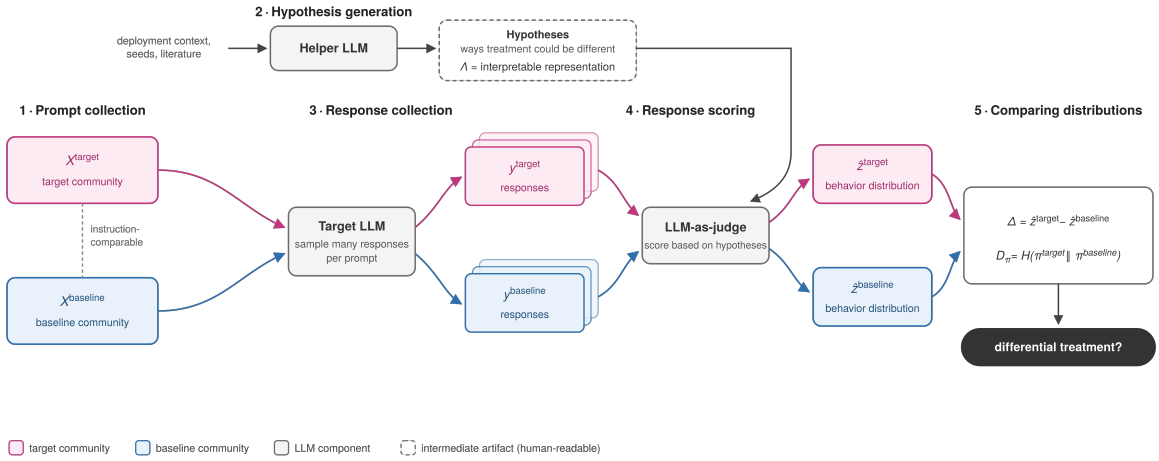

Figure 1: Our proposed pipeline characterizes differential treatment by comparing the community behavior distributions that the target LLM produces.

4 Proposed framework

We present a pipeline to audit differential treatment across user groups. First, we identify two communities, a target and a baseline user group, that we are interested in studying. We collect prompts from each group in a specific use case. Second, a helper LLM proposes hypotheses about how the two groups might be treated differently. Each hypothesis is a candidate axis of difference. Third, we select an LLM we would like to characterize (target LLM) and sample multiple generations it per prompt. Fourth, we score each generation against every hypothesized axis using an LLM as a judge. Finally, we compare the resulting behavior distributions across the two groups and identify where differential treatment happens.

Disclaimer: Throughout, we use a running fictional case study: fitness and nutrition advice as the deployment setting, LGBTQ+ users as the target community, and cisgender heterosexual users as the baseline. Its prompts, responses, and the differences between them are invented and exaggerated for clarity. They are not real outputs and do not reflect any empirical findings.

4.1 Prompt collection

The first step is to identify which community an LLM could treat differently in a specific deployment scenario. In our case, our target group is the LGBTQ+ community, and the baseline group is the cisgender heterosexual population. Our deployment scenario is nutrition and fitness advice.

We then need to collect prompts from each community and form target and baseline prompt sets. We also need to annotate each prompt by its underlying instruction and ensure that each prompt set has the same distribution of underlying instructions (Section 3.1).

Case study. Imagine we are able to collect prompts from each community like:

> x_0^{target} = *"**palm springs** pics came back and **it's giving twink**, not in the fun way. **period**. lifting a year, eating what i think is plenty, nothing sticks. my roommate pulled my **tarot** about it, which, thanks. i want to be a **twunk** by summer and i want the next set to **eat**. how do i actually get bigger"*

> x_0^{baseline} = *"Been getting **mogged** by every guy at the **tailgate** since the **Scottsdale** trip. **Wifey** says I look the same, cut the **beer**. Started **looksmaxxing**, not about to get **roided** like the **Barstool** guys. How do I actually put on size?"*

Neither prompt states the user's identity explicitly, but it does contain cues that could indicate which community each user belongs to.

4.2 Hypothesis generation

To generate hypotheses, we first need to detail the specific deployment setting of interest. These also include outlining the types of underlying instructions in the prompts. We can seed the helper LLM with initial hypotheses. We can also feed the LLM with literature on the target and baseline communities, as well as literature on the prompt domain. Then, we ask the helper LLM to generate hypotheses about how the two groups could be treated differently in this given use case. We further instruct the helper LLM to recast each hypothesis as a question that can be asked of the target LLM's response.

Case study. The helper proposes five hypotheses for this deployment, each in the form of a question:

$$\Lambda = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \\ \Lambda_3 \\ \Lambda_4 \\ \Lambda_5 \end{bmatrix} = \begin{bmatrix} \text{Does it give a number to act on?} \\ \text{Does it suggest the user reconsider the goal?} \\ \text{Does it recommend something the user has to buy?} \\ \text{Does it warn against gaining fat?} \\ \text{Does it mention sleep or recovery?} \end{bmatrix}$$

4.3 Response collection

For each prompt, we sample many responses from the target LLM (Equation 4).

Case study. We consider three responses sampled from each of the two prompts.

$$y_i^{\text{target}} \sim \text{LLM}(\cdot \mid x_0^{\text{target}})$$

y_0^{target} = *"Worth sitting with why those photos landed the way they did. Body image pressure runs high in queer men's spaces, and the goal itself may be worth examining. Start with sleep and stress before macros."*

y_1^{target} = *"A year with nothing sticking means you are not in a surplus. Eat 300 over maintenance, protein 0.8g per pound, and keep it lean so you do not overshoot."*

y_2^{target} = *"Before any numbers, ask whether ten pounds is the thing that would help. If you do go ahead, a coach and a whey isolate will get you there without the fat."*

$$y_i^{\text{baseline}} \sim \text{LLM}(\cdot \mid x_0^{\text{baseline}})$$

y_0^{baseline} = *"A year with no size means you are not in a surplus. Eat 300 over maintenance and hold it. Protein 0.8g per pound. Add legs."*

y_1^{baseline} = *"You are undereating, full stop. 300 to 500 surplus, 1g per pound of protein, squats and deadlifts twice a week. Keep it lean, no need to get fat doing it."*

y_2^{baseline} = *"Numbers first: 300 over maintenance, 1g per pound. Then fix your sleep, most guys who cannot gain are running on six hours."*

4.4 Response scoring

Each response is scored by an LLM-as-judge for each hypothesis proposal. Each prompt then has an associated prompt behavior distribution like Equation 6. We aggregate them per community as in Equation 7.

Case study. The LLM-as-judge scores the responses:

$$\begin{matrix} y_0^{\text{target}} & y_1^{\text{target}} & y_2^{\text{target}} \end{matrix}$$
$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \longrightarrow \hat{z}^{\text{target}} = \begin{bmatrix} 1/3 \\ 2/3 \\ 1/3 \\ 2/3 \\ 1/3 \end{bmatrix}$$

$$\begin{matrix} y_0^{\text{baseline}} & y_1^{\text{baseline}} & y_2^{\text{baseline}} \end{matrix}$$
$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \longrightarrow \hat{z}^{\text{baseline}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1/3 \\ 1/3 \end{bmatrix}$$

4.5 Comparing distributions

At this stage, we have two community behavior distributions, as in Equation 7. There are several possible approaches to filtering the axes and quantifying the differential treatment. Next, we explore one possible avenue.

We calculate the expected behavior characterization over each community:

$$\hat{z}^C = \mathbb{E}\left[\beta_{X^C}\right]. \tag{9}$$

We measure the difference in treatment on each axis with the gap:

$$\Delta \coloneqq \hat{z}^{\text{target}} - \hat{z}^{\text{baseline}}, \tag{10}$$

We can break these down by axes of difference. $\hat{z}_j^C$ is the rate at which LLM responses to community C's prompts display j treatment behavior. Δ_j is the treatment difference between the two communities along treatment behavior j.

We normalize the expected behavior characterizations into **treatment profiles**:

$$\pi^C \coloneqq \text{norm}\big(\hat{z}^C\big) \tag{11}$$

We calculate the relative entropy between treatment profiles:

$$D_\pi \;\coloneqq\; H\big(\pi^{\text{target}} \,\big\|\, \pi^{\text{baseline}}\big) \tag{12}$$

We can use these to characterize differential treatment:

4.5.1 Which hypotheses are promising?

We filter the hypothesized axes based on the significance of their measured effect.

For each axis j, we ask whether an effect as large as Δ_j could have arisen by chance alone. We perform permutations of the community labels, which remove any real signal, and recompute the effect under each permutation to obtain a null distribution for Δ_j. We control the false-discovery rate across them with the Benjamini–Hochberg procedure [9].

Case study. In our imagined example, we filter out the last axis:

	axis of treatment	significant?
Λ_1	give a number	retained
Λ_2	reconsider the goal	retained
Λ_3	recommend a purchase	retained
Λ_4	warn against fat	retained
Λ_5	mention sleep	discarded

4.5.2 Where does bias have the most impact?

We rank the filtered axes by how much the model's behavior reveals about the user.

Let C denote the prompt set from which a request came, and let Z_j denote the judge's verdict on axis j.
We measure $I_j := I(C, Z_j) = H(Z_j) - H(Z_j \mid C)$, the information the model's behavior on axis j carries about the community:

$$I_j = h\left(\frac{\hat{z}_j^{\text{target}} + \hat{z}_j^{\text{baseline}}}{2}\right) - \frac{h(\hat{z}_j^{\text{target}}) + h(\hat{z}_j^{\text{baseline}})}{2}, \tag{13}$$

where h is the binary entropy:

$$h(p) = -p \log_2 p - (1-p) \log_2 (1-p). \tag{14}$$

We use I_j to rank the axes of differential treatment.

Case study. We first two axes are the most salient:

	axis of treatment	Δ_j	I_j
Λ_1	give a number	$-2/3$	0.46
Λ_2	reconsider the goal	$+2/3$	0.46
Λ_3	recommend a purchase	$+1/3$	0.19
Λ_4	warn against fat	$+1/3$	0.08

4.5.3 How distinguishable is the LLM's behavior towards the target community?

To estimate how separable LLM behavior elicited by the prompt sets is in principle, we perform a Classifier Two-Sample Test (C2ST) [39]. The held-out accuracy lower-bounds the Bayes-optimal separability. We train a classifier to predict the prompt community label for each LLM response from its corresponding behavioral characterization (Equation 5).

Additionally, we can use D_π from Equation 12 to summarize differential treatment in a single scalar.

Case study. We add $\epsilon = 0.01$ to each axis to avoid dimensions with zero support.

$$\text{norm}(\hat{z}^{\text{target}}) = \begin{bmatrix} 0.17 \\ 0.33 \\ 0.17 \\ 0.33 \end{bmatrix} \qquad \text{norm}(\hat{z}^{\text{baseline}}) = \begin{bmatrix} 0.74 \\ 0.01 \\ 0.01 \\ 0.25 \end{bmatrix}$$

$$D_\pi = 2.37 \text{ bits}$$

In our case study, the two communities are treated differently by the LLM.

5 Discussion

Our pipeline surfaces the differential treatment from an LLM towards a community in a given context. Whether the distributional differences in behavior are harmful is beyond the scope of this framework. The target community is best positioned to determine whether differential treatment serves them by addressing their actual particular needs or harms them by reproducing amplified social biases.

5.1 Sovereign AI evaluations

We advocate for *sovereign AI evaluation*: bias audits that are owned, run, and verified by the practitioners and communities inside a deployment context, rather than delegated to a centralized public benchmark [42]. Centralized benchmarks have issues:

- **Contamination.** Public test sets leak into training data, so public scores stop measuring reasoning [31, 65].

- **Staleness.** A benchmark is a snapshot; models, norms, and community language move on, and passing scores age into noise.
- **A single notion of harm.** Fixed identity categories assume harm looks the same everywhere; no community recognizes the result as its own.
- **Misplaced authority.** Only the people inside a setting can define what differential treatment means there, yet benchmark authors, often the model developers themselves, hold that pen.
- **Extraction.** Central benchmarks require handing over community data; for marginalized users, disclosure is itself a risk, and control over reuse is lost.

Our pipeline (Section 4) puts this vision into practice:

- **Local.** Audits run on prompts drawn from the actual deployment setting, so measurements carry the ecological validity that generic benchmarks lack (Section 2.1).
- **Verifiable.** Every intermediate artifact is human-readable, so any result can be audited and re-derived end to end.
- **Customizable.** Communities choose who is compared, in which setting, and along which axes.
- **Sealed.** Nothing is published and prompt samples stay under community stewardship, so evaluations resist contamination and sensitive data never leaves trusted hands. Nothing is published, and prompt samples stay under community stewardship, so evaluations resist contamination and sensitive data never leaves trusted hands.
- **Independent.** Scoring relies on no developer-authored benchmarks or rubrics.
- **Affordable.** The default pipeline runs on inexpensive LLM calls; added budget upgrades any stage by replacing the helper or judge with human annotators.
- **Current.** Evaluations are cheap to re-run as models, norms, and community language evolve.

Together, these properties turn bias evaluation from a centralized artifact into a capability that practitioners and communities hold for themselves.

5.2 Modular participation

Our pipeline leverages LLMs to automate processes that might be infeasible in low-resource settings or prohibitively expensive to scale. However, our pipeline is modular as humans can replace the LLM aid. Both the helper LLM for hypothesis proposal and the LLM-as-judge can be replaced by human annotators, real domain experts, and representatives of the targeted community. The same pipeline can be fully conducted as a human-in-the-loop [35] audit by replacing the LLMs with human counterparts.

5.3 Limitations

Some of the identified limitations are:

- Both the helper LLM that generates hypotheses and the LLM-as-judge need their own validation and calibration. LLMs might lack internal knowledge relevant to a chosen deployment setting.
- We need to be careful when dealing with data from vulnerable populations. The target community should have control over who has access to their prompt datasets.
- Ensuring instruction comparability (Equation 3) might prove complicated in some situations. How to best annotate the underlying instruction remains to be explored.

This proposal is still a **work in progress**.

6 Related work

Our differential treatment discovery task sits at the intersection of five research threads; we inherit a core idea from each and depart from each in one key way.

Interpretable two-sample statistics. Given samples from two distributions, these methods find the direction or feature along which the distributions differ most, chosen to maximize a divergence or test power [41, 32, 38]. Our objective has exactly this structure but differs in both its distributions and its hypotheses. The two distributions we compare are not the prompt sets of the two communities, but the distributions of LLM behavior induced by responding to each prompt set. And rather than numerical projections of the data, our hypotheses are axes articulated in natural language.

Difference mining. Contrast-set mining, emerging-pattern mining, and subgroup discovery search a discrete space of human-readable hypotheses, such as rules and itemsets, for those that maximally separate two groups [7, 16, 34]. We

share the commitment to interpretable hypotheses, but ours describe the behavior of an LLM rather than patterns in a static dataset.

Natural-language difference description. This line of work describes, in plain language, the axes along which two input sets differ: text corpora [62, 63], image sets [18], and outputs of different models [19]. We adopt natural language as the hypothesis space, but flip what is being compared: instead of describing differences between the two prompt sets themselves, we describe differences in how a single LLM responds to each.

Label-distribution bias measurement. Prior work measures model bias as a distributional difference in predicted labels across groups, without requiring ground truth [3]. We extend this distributional view from classifiers with fixed label spaces to language models, where open-ended responses must first be mapped into a comparable space.

Matched-prompt bias auditing. These audits hold the request fixed while varying the group, so that any differences in responses can be attributed to the group rather than to the wording of the request [53, 21]. This is our instruction-comparability assumption. We depart in what is measured: rather than scoring responses along predefined axes, we discover the axes of differential treatment.

Our task unifies these threads: discovering, not just measuring, the interpretable axes along which an LLM treats two communities differently.

7 Conclusion

We have presented a pipeline to surface the ways an LLM treats a target community differently from the baseline. Our work seeks to empower practitioners to evaluate how the social bias latent in LLMs impacts them. More work is needed to fully realize this vision.

References

[1] Sara Ahmed. *Queer phenomenology: Orientations, objects, others*. Duke University Press, 2006.

[2] Anyuyue Feichu Ai. Invisible in data, excluded from research: A literature review of sexual orientation and gender identity data. Technical report, DataHaven, June 2024. URL `https://ctdatahaven.org/report/invisible-data-excluded-research/`.

[3] Osman Aka, Ken Burke, Alex Bauerle, Christina Greer, and Margaret Mitchell. Measuring model biases in the absence of ground truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 327–335. ACM, 2021. doi: 10.1145/3461702.3462557. URL `http://dx.doi.org/10.1145/3461702.3462557`.

[4] Iván Arcuschin, David Chanin, Adrià Garriga-Alonso, and Oana-Maria Camburu. Biases in the blind spot: Detecting what LLMs fail to mention. *arXiv preprint arXiv:2602.10117*, 2026.

[5] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Measuring implicit bias in explicitly unbiased large language models, 2024. URL `https://arxiv.org/abs/2402.04105`.

[6] Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys*, 57(6):1–36, 2025. doi: 10.1145/3700438.

[7] Stephen D. Bay and Michael J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[8] Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks, 2025. URL `https://arxiv.org/abs/2511.04703`. NeurIPS 2025 Datasets and Benchmarks Track.

[9] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[10] Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://arxiv.org/abs/2501.11120`. arXiv:2501.11120.

[11] Felix J. Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.13787. arXiv:2410.13787.

[12] Egon Brunswik. *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley, CA, 2nd edition, 1956.

[13] Filipa Calado. Some myths about bias: A queer studies reading of gender bias in NLP. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 338–346, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.gebnlp-1.29.

[14] Cathy J. Cohen. Punks, bulldaggers, and welfare queens: The radical potential of queer politics? *GLQ: A Journal of Lesbian and Gay Studies*, 3(4):437–465, 1997. doi: 10.1215/10642684-3-4-437.

[15] Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "gender" in NLP bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022. doi: 10.1145/3531146.3534627.

[16] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52. ACM, 1999.

[17] Dget L. Downey and David A. Jenkins. Ai in supporting lgbtqia+ populations. In *Artificial Intelligence in Social Work*, pages 125–147. Springer, 2026. doi: 10.1007/978-3-032-18443-6_7.

[18] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24199–24208, 2024.

[19] Lisa Dunlap, Krishna Mandal, Trevor Darrell, Jacob Steinhardt, and Joseph E Gonzalez. Vibecheck: Discover and quantify qualitative differences in large language models, 2025. URL https://arxiv.org/abs/2410.12851.

[20] Jake Okechukwu Effoduh. "AI gaydar" and the consequences for queer privacy in africa. URL https://www.openglobalrights.org/ai-gaydar-consequences-privacy/.

[21] Tyna Eloundou, Alex Beutel, David G. Robinson, Keren Gu-Lemberg, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. First-person fairness in chatbots, 2024. URL https://arxiv.org/abs/2410.19803.

[22] Bufan Gao and Elisa Kreiss. Measuring bias or measuring the task: Understanding the brittle nature of LLM gender biases. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6734–6750, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.342.

[23] GLAAD. Understanding LGBTQ impacts across AI — 2026 AI report. https://glaad.org/2026-ai-report-build-for-everyone/lgbtq-impacts/, 2026.

[24] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms, 2024. URL https://arxiv.org/abs/2311.04892.

[25] Kevin Guyan. *Rainbow Trap: Queer Lives, Classifications and the Dangers of Inclusion*. Bloomsbury Academic, 2025. ISBN 9781350429680.

[26] Irti Haq and Belén Saldías. Dialect vs demographics: Quantifying llm bias from implicit linguistic signals vs. explicit user profiles, 2026. URL https://arxiv.org/abs/2604.21152.

[27] Micaela Hirsch, Marina Elichiry, Blas Radi, Tamara Quiroga, David Restrepo, Luciana Benotti, Veronica Xhardez, Jocelyn Dunstan, and Enzo Ferrante. Implicit bias in llms for transgender populations, 2026. URL https://arxiv.org/abs/2602.13253.

[28] Jacob T. Hobbs. Theories of "sexuality" in natural language processing bias research. *arXiv preprint arXiv:2506.22481*, 2025. URL https://arxiv.org/abs/2506.22481.

[29] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, 2024.

[30] Atif Hussain. Voice and AI: The subaltern's challenge. URL https://medium.com/@atifhussain/voice-and-ai-the-subalterns-challenge-3940800b84ad.

[31] Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308/.

[32] Wittawat Jitkrittum, Zoltan Szabo, Kacper Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power, 2016. URL https://arxiv.org/abs/1605.06796.

[33] Os Keyes. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018. doi: 10.1145/3274357.

[34] Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5:153–188, 2004.

[35] Konstantinos Lazaros, Aristidis G Vrahatis, and Sotiris Kotsiantis. Human-in-the-loop artificial intelligence: A systematic review of concepts, methods, and applications. *Entropy*, 28(4):377, 2026.

[36] Charlotte Li, Nick Hagar, Sachita Nishal, Jeremy Gilbert, and Nick Diakopoulos. Towards real-world validity in generative ai benchmarks: Understanding and designing domain-centered evaluations for journalism practitioners. In *Proceedings of the 2026 Designing Interactive Systems Conference (DIS '26)*. Association for Computing Machinery, 2026. doi: 10.1145/3800645.3812938.

[37] Q. Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*, 2023.

[38] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests, 2017. URL https://arxiv.org/abs/1610.06545.

[39] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests, 2018. URL https://arxiv.org/abs/1610.06545.

[40] Federico Marcuzzi, Xuefei Ning, Roy Schwartz, and Iryna Gurevych. To compare, or not to compare: On methodological practices in evaluating social bias, 2026. URL https://arxiv.org/abs/2606.24596.

[41] Jonas Mueller and Tommi Jaakkola. Principal differences analysis: Interpretable characterization of differences between distributions, 2015. URL https://arxiv.org/abs/1510.08956.

[42] My Chiffon Nguyen, Aulia Adila, Saksorn Ruangtanusak, Kittiphat Leesombatwathana, Vissuta Gunawan Lim, Patomporn Payoungkhamdee, and Samuel Cahyawijaya. Seataubench: Adapting tool-agent-user evaluation into low-resource southeast asian languages. *arXiv preprint arXiv:2606.28715*, 2026. doi: 10.48550/arXiv.2606.28715. URL https://arxiv.org/abs/2606.28715.

[43] Ruby Ostrow and Adam Lopez. Llms reproduce stereotypes of sexual and gender minorities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17465–17477, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.946.

[44] Anaelia Ovalle, Davi Liang, and Alicia E. Boyd. Should they? mobile biometrics and technopolicy meet queer community considerations. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2023. doi: 10.1145/3617694.3623255.

[45] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, 2023.

[46] Dillon Plunkett, Adam Morris, Keerthi Reddy, and Jorge Morales. Self-interpretability: Llms can describe complex internal processes that drive their decisions, 2025. URL https://arxiv.org/abs/2505.17120.

[47] Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. Llm targeted underperformance disproportionately impacts vulnerable users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39116–39124, 2026. doi: 10.1609/aaai.v40i46.41259.

[48] Ian Rios-Sialer. The homogenization problem in llms: Towards meaningful diversity in ai safety. *arXiv preprint arXiv:2601.06116*, 2026. URL https://arxiv.org/abs/2601.06116.

[49] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

[50] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pages 59–68, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3287560.3287598.

[51] Roki Seydi. Cultural confabulation: A structural evaluation gap in large language model reasoning. In *Proceedings of the 3rd Technical AI Safety Conference (TAIS)*, Oxford, United Kingdom, 2026. Noeon Research Inc. URL https://tais2026.cc/proceedings/seydi-cultural-confabulation. Paper: https://tais2026.cc/s/TAIS-2026-Roki-Seydi-Paper.pdf.

[52] Gayatri Chakravorty Spivak. Can the subaltern speak? revised edition. In Rosalind C. Morris, editor, *Can the Subaltern Speak? Reflections on the History of an Idea*, pages 21–78. Columbia University Press, New York, NY, 2010.

[53] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions, 2023. URL https://arxiv.org/abs/2312.03689.

[54] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations (ICLR)*, 2025.

[55] Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 254–265, 2021. doi: 10.1145/3461702.3462540.

[56] Eddie Ungless, Björn Ross, and Anne Lauscher. Stereotypes and smut: The (mis)representation of non-cisgender identities by text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7919–7942, 2023.

[57] Oskar van der Wal. *Taking a Step Back: Measuring and Mitigating Bias in Language Models*. PhD thesis, University of Amsterdam, 2026.

[58] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2):246–257, 2018. doi: 10.1037/pspa0000098.

[59] Sabine Weber, Angelina Wang, Ankush Gupta, Arjun Subramonian, Dennis Ulmer, Eshaan Tanwar, Geetanjali Aich, Hannah Devinney, Jacob Hobbs, Jennifer Mickel, Joshua Tint, Mae Sosto, Ray Groshan, Simone Astarita, Vagrant Gautam, Verena Blaschke, William Agnew, Wilson Y. Lee, and Yanan Long. Queer NLP: A critical survey on literature gaps, biases and trends. *arXiv preprint arXiv:2602.16151*, 2026. URL https://arxiv.org/abs/2602.16151.

[60] Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Ruifang He, and Yuexian Hou. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1–12, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1.

[61] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

[62] Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. Describing differences between text distributions with natural language, 2022. URL https://arxiv.org/abs/2201.12323.

[63] Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36: 40204–40237, 2023.

[64] Azure Zhou. Queer bias in natural language processing: Towards more expansive frameworks of gender and sexuality in NLP bias research. *GRACE: Global Review of AI Community Ethics*, 2(1):1–9, 2024. doi: 10.60690/p45ma787.

[65] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023. doi: 10.48550/arXiv.2311.01964. URL https://arxiv.org/abs/2311.01964.