Contrastive Analysis of Linguistic Representations in Large Language Model Outputs through Structured Synthetic Data Generation and Abstracted N-gram Associations

Análisis contrastivo de representaciones lingüísticas en salidas de modelos de lenguaje de gran escala mediante generación estructurada de datos sintéticos y asociaciones abstractas de n-gramas

Abstract: We present a methodological framework to discover linguistic and discursive patterns associated to different social groups through contrastive synthetic text generation and statistical analysis. In contrast with previous approaches, we aim to characterize subtle expressions of bias, instead of diagnosing bias through a pre-determined list of words or expressions. We are also working with contextualized data instead of isolated words or sentences. Our methodology applies to textual productions in any genre, encompassing narrative, task-oriented or dialogic. Contextualized data are generated using controlled combinations of situational scenarios and group markers, creating minimal pairs of texts that differ only in the referenced group while maintaining comparable narrative conditions. To facilitate robust analysis, linguistic forms are generalized and associations between linguistic abstractions and groups are quantified using a variant of pointwise mutual information to detect expressions that appear disproportionately across groups. A fragment-ranking strategy then prioritizes text segments with a high concentration of biased linguistic signals, which allows for experts to assess the harmful potential of linguistic expressions in context, bridging quantitative analysis and qualitative interpretation.
Keywords: Large Language Models, Bias Analysis, Synthetic Data

Resumen: Presentamos un marco metodológico para descubrir patrones lingüísticos y discursivos asociados a distintos grupos sociales mediante generación sintética contrastiva de texto y análisis estadístico. A diferencia de enfoques previos, nuestro objetivo es caracterizar formas sutiles de sesgo, en lugar de diagnosticarlo a partir de listas predefinidas de palabras o expresiones. Asimismo, trabajamos con datos contextualizados en lugar de palabras o oraciones aisladas. Nuestra metodología se aplica a producciones textuales de cualquier género, incluyendo narrativas, textos orientados a tareas o interacciones dialógicas. Los datos contextualizados se generan mediante combinaciones controladas de escenarios situacionales y marcadores de grupo, creando pares mínimos de textos que difieren únicamente en el grupo referido, mientras mantienen condiciones comparables. Para facilitar un análisis robusto, las formas lingüísticas se generalizan y las asociaciones entre abstracciones lingüísticas y grupos se cuantifican mediante una variante de la información mutua puntual, lo que permite detectar expresiones que aparecen de manera desproporcionada entre grupos. Finalmente, una estrategia de priorización de fragmentos identifica segmentos textuales con alta concentración de señales lingüísticas potencialmente sesgadas, lo que permite a expertos evaluar el posible daño de dichas expresiones en contexto, articulando el análisis cuantitativo con la interpretación cualitativa.
Palabras clave: Modelos de Lenguaje de Gran Escala, Análisis de Sesgos, Datos Sintéticos

1 Introduction and Motivation

Large Language Models (LLMs) have demonstrated outstanding performance in text generation tasks. However, numerous studies have pointed out that these systems may reproduce and amplify biases present in their training data, particularly along dimensions such as gender, race, sexual orientation, and religion (Caliskan, Bryson, and Narayanan, 2017; Mehrabi et al., 2021; Tamkin et al., 2023; Gallegos et al., 2024; Caliskan, Bryson, and Narayanan, 2017; Mehrabi et al., 2021; Tamkin et al., 2023; Kotek, Dockum, and Sun, 2023; Gallegos et al., 2024). In a context where LLMs are rapidly being integrated into high-impact social domains, such as healthcare, education, and justice, the rigorous evaluation of their discursive behavior becomes a priority issue (Bender et al., 2021; Wan et al., 2023; Panda, Agarwal, and Patel, 2025).

Much of the literature has focused on explicit forms of bias, in which marginalized groups are directly named and associated with identifiable stereotypes. However, bias may also operate implicitly, through more subtle and indirect mechanisms that shape model outputs(Kumar, Yunusov, and Emami, 2024; Bai et al., 2025; Hirsch et al., 2026). These forms of bias often manifest through complex textual patterns rather than overtly negative statements, including differences in the attribution of roles, agency, or evaluative traits. For instance, certain groups may be disproportionately framed through narratives that are superficially positive yet ultimately limiting, such as overly inspirational or patronizing portrayals, while others are more frequently associated with attributes like professionalism, competence, or precision (Grue, 2016; Viveros Vigoya, 2016). Such patterns may also emerge through differential associations with activities, emotional registers, or social roles, including both overrepresentation and systematic absence

These manifestations are inherently more difficult to detect, particularly when analysis relies on predefined lists of sensitive terms, toxicity metrics, or fixed categorical frameworks (e.g., emotion labels such as fear, joy, or positivity). While useful in specific settings, these approaches tend to fragment the phenomenon and impose a reductionist structure that may obscure the broader discursive configurations through which subtle biases are expressed(Miceli et al., 2025; Panda, Agarwal, and Patel, 2025).

In this work, we propose a methodology aimed at identifying subtle, systematically differentiated patterns in generated text through contrastive synthetic data and statistical analysis of associations. Our approach seeks to identify textual fragments that concentrate linguistic configurations differentially associated with social groups under equivalent contextual conditions. These fragments are then assessed by experts to determine whether they may produce harms. The workflow consists of three main stages: (1) contrastive generation of texts for each social group; (2) statistical discovery of association patterns between social groups and textual expressions; and (3) identification of textual fragments with a high density of differentially associated expressions, so that experts may assess whether they produce any kind of harm. This design allows experts to focus on fragments that exhibit strong differential behaviors, facilitating the identification and characterization of discursive patterns associated with unequal or potentially harmful representations. In contrast to previous approaches, which often require inspecting large volumes of text or decontextualized outputs, our method directs attention to a reduced set of context-rich and statistically grounded examples.

The rest of the paper is organized as follows. In the next Section, we describe the capabilities and limitations of different approaches to bias detection, and how this approach proposes to overcome them. Then, we describe the main parts of the proposed methodology: contrastive generation of texts, discovery of associations and fragment ranking and assessment. Finally, we present a proof of concept with the group of people with disabilities in Spanish.

2 Relevant Work

Research on bias in language models has received increasing attention due to both the technical challenges it poses and its social implications (Bender et al., 2021; Gallegos et al., 2024; Kumar, Yunusov,

and Emami, 2024). Numerous studies have developed methods to explore, quantify, and mitigate bias in automated linguistic systems, consolidating an active field of evaluation and auditing.

A substantial portion of this literature has focused on detecting explicit bias, frequently through the use of predefined lists of sensitive terms, stereotypical templates, or structured benchmarks (Zhao et al., 2018; Nangia et al., 2020; Nadeem, Bethke, and Reddy, 2021; Parrish et al., 2022). While these approaches have proven effective in identifying overt forms of discrimination, they present limitations in capturing more nuanced phenomena, such as differential narrative framing, implicit condescension, or systematically biased representations that are not expressed through openly offensive language.

In this regard, recent work emphasizes that some of the most persistent forms of bias in LLMs are implicit or structural (Kumar, Yunusov, and Emami, 2024; Panda, Agarwal, and Patel, 2025; Bai et al., 2025; Hirsch et al., 2026; Bali et al., 2026), operating through recurrent associations and discursive patterns that manifest indirectly. In response to these limitations, more contextualized approaches have begun to emerge, aimed at identifying patterns that arise beyond predefined categories.

In this work, we propose a methodology aimed at operationalizing qualitative evaluation and collective participation, systematically producing evidence that can be critically examined by domain experts and potentially affected communities. In this way, the methodology structures the exploration process and the detection of statistically significant patterns, while experts are only required to assess the potential harms present in the identified fragments.

In contrast with other approaches, our method does not predefine bias indicators; instead, these are discovered from the texts themselves and are not constrained to any fixed or predefined form. Rather, they may take arbitrary linguistic configurations. Moreover, expert involvement always occurs at a level of abstraction or concreteness that is natural to them and in a contextualized manner: by describing groups and situations that may be differentially represented by LLMs, and by evaluating the potential harm in fragments embedded within complete texts.

3 Contrastive Text Generation

The first step in our methodology consists in generating synthetic data to obtain controlled contexts in which differential behavior across social groups may emerge. By constructing minimal pairs of prompts that vary only in the social group referenced, the methodology enables the systematic analysis of whether and how LLMs produce textual outputs for different groups under equivalent conditions.

This approach is based on the assumption that subtle biases often emerge only through comparison: linguistic patterns that appear neutral in isolation, like the word "*inspiring*", may reveal discriminatory effects when we can systematize how they are differentially distributed across texts for different social groups, and how they may occupy the space of other expressions and concepts, like "*reliable*".

The process is organized around two main components: (i) a set of scenarios that define the situational context and remain constant across conditions, usually originating from initial hypotheses from experts; and (ii) a set of markers that specify attributes of a person, distinguishing between the group of interest and the control groups.

From a base prompt like the one shown in Figure 2, pairs of texts are generated that share the same situational context but differ only in the marker used. Thus, we obtain minimal pairs of texts. The methodology of minimal pairs has been extensively used in bias measurements (e.g., Nangia et al. (2020)), but these minimal pairs are crucially contextualized, in contrast with isolated words or sentences. This contrastive design enables direct comparison between comparable narratives because the texts resulting from minimal pairs of prompts are comparable texts, facilitating the identification of patterns that occur disproportionately in texts associated with the group of interest. This approach makes it possible to detect subtle biases that emerge at the level of narrative organization and may remain invisible under purely lexical analysis. At the same time, it allows experts to describe their hypotheses on discrimination

in a natural way.

4 Discovery of Associations

Once the texts have been generated through a contrastive process, we apply statistical analysis to determine whether they differ significantly across the groups under comparison, and to characterize these differences when they arise. To this end, we aim to identify textual expressions that are more strongly associated with one group than with others. When such expressions exhibit a stronger association with a particular group, we interpret them as evidence of differential model behavior and consider them as indicators of group-specific patterns.

A simple and intuitive measure of association is Pointwise Mutual Information (PMI), which has been successfully linked to various bias detection approaches (Bordia and Bowman, 2019; Aka et al., 2021; Valentini et al., 2023). However intuitive, the main shortcoming of PMI is that it is sensitive to the frequency of the analyzed units, which makes it difficult to compare associations when dealing with low-frequency expressions. In particular, PMI estimates tend to be less stable for rare events, which may affect the interpretability of the results. For this reason, in order to make the measure less sensitive to frequency, we propose to generalize linguistic expressions by grouping low-frequency expressions into more general units that share a common meaning. Once this generalization is performed, PMI can be applied to these units. In the remainder of this section, we describe how linguistic forms are generalized and how associations between these generalized forms and the groups under study are identified.

4.1 Generalization of Linguistic Forms

To mitigate the sparsity of the distribution and reduce the impact of low-frequency events, we generalize linguistic forms by constructing more abstract analytical units. First, all tokens except stopwords are lemmatized, which reduces morphological variability and decreases the number of distinct units under analysis.

Building on this representation, we define equivalence classes of n-grams. For each n-gram, non-stopword tokens are mapped to their corresponding lemmas. N-grams that share the same set of lemmas are then grouped into a common equivalence class, independently of word order. This grouping relies on the assumption that such n-grams convey a similar underlying meaning despite surface-level variation. By aggregating multiple low-frequency realizations into shared abstractions, this process produces a more stable distribution over which association measures such as PMI can be more reliably estimated.

“enfrentado todos los obstáculos”, “obstáculos que había enfrentado”, “enfrentado varios obstáculos”, “enfrentó varios obstáculos”, “enfrentas obstáculos”, “enfrentado obstáculos”.

Figure 1: Equivalence class of n-grams with the lemmas “enfrentar” (*face*) and “obstáculo” (*obstacle*).

An example can be seen in Figure 1, where all n-grams belong to the same equivalence class because their content words correspond to the lemmas “enfrentar” (*face*) and “obstáculo” (*obstacle*).

We exclude n-grams bebinning or ending with a stopword. For example, the trigram “amigos y familia” (*friends and family*) would be included, whereas “los amigos y” (*the friends and*) would be excluded. This is because n-grams that begin or end with a stopword tend to lack semantic autonomy and provide less stable analytical units.

4.2 Measuring strength of association

Once linguistic forms have been generalized into equivalence classes, we measure the strength of association between each equivalence class and each of the groups under study. The underlying intuition is that if an equivalence class occurs more frequently in texts generated for one of the groups than in texts generated for any other group, then it can be considered more strongly associated with that group, with a strength proportional to the difference in the number of occurrences across groups. To do that, we compute the PMI between equivalence classes of n-grams and the groups.

Formally, the PMI between a linguistic

unit w and a group G is defined as:

$$\mathrm{PMI}(w, G) = \log \frac{P(w \mid G)}{P(w)}. \quad (1)$$

With this definition of PMI, and comparably to the intuition captured by previous work like Valentini (2023), we can define the *BiasScore* (BS) of a unit w as the difference between its PMI with respect to a group of interest (GI) and a control group (GC):

$$\mathrm{BS}(w) = \mathrm{PMI}(w, GI) - \mathrm{PMI}(w, GC) \quad (2)$$

$$= \log \frac{P(w \mid GI)}{P(w \mid GC)}. \quad (3)$$

In practice, BS is estimated from frequency counts:

$$\mathrm{BS}(w) = \log \frac{M_{w,GI}}{M_{\cdot GI}} - \log \frac{M_{w,GC}}{M_{\cdot GC}} \quad (4)$$

where $M_{w,G}$ denotes the frequency of w in group G, and $M_{\cdot G}$ the total number of units observed in that group.

Our analysis operates over equivalence classes of n-grams. The frequency of an equivalence class is defined as the sum of the frequencies of all units that belong to it.

For higher-order n-grams, we treat the ratio $\frac{M_{\cdot GI}}{M_{\cdot GC}}$ as constant across different equivalence classes, as it primarily depends on the relative size of the corpora in each group. This allows us to apply the same normalization scheme across different types of units.

An example of the BS assigned to different equivalence classes can be seen in Tables 1 and 2.

5 Fragment Ranking

Finally, once we have discovered the strength of association of textual expressions with each group, we can use this information to identify fragments of text that contain a higher concentration of expressions more strongly associated with a given group relative to others, such as those depicted in Figure 3. These fragments are then assessed by experts to determine whether they cause harm.

In this stage, we operationalize the identification of potentially biased text fragments by leveraging the previously defined equivalence classes and their corresponding BS values. This involves specifying which types of fragments and equivalence classes are relevant, as well as applying a scoring mechanism to prioritize fragments according to the concentration of differential associations they contain. While the details of this process can be adapted to the goals of each analysis, here we propose a strategy that serves as a general methodological framework, flexible across different corpora, application contexts, and research questions.

5.1 Filtering intrinsic differences

Some expressions may show a strong association with a given group because they refer to intrinsic characteristics of that group, and not necessarily to indicators of bias. In these cases, the differential distribution of certain n-grams across groups does not constitute evidence of subtle bias, but rather reflects inherent or expected properties.

Consequently, not all equivalence classes derived from the generalization process are equally informative for analysis. To address this issue, we construct a set of expressions considered inherent to the groups under study and exclude the equivalence classes that contain them. For example, if the group of interest is people with disabilities, the word "disléxico" (*dyslexic*) would lead to the exclusion of equivalence classes containing n-grams with forms such as "disléxica" (feminine), "disléxicos", or "disléxicas".

Additionally, filters can be applied based on lists defined by domain experts, individuals with lived experience, or other relevant stakeholders. These lists include expressions whose differential occurrence across groups is expected and does not, by itself, indicate bias (e.g., "cane" or "sign language"). This step allows the analysis to focus on subtle linguistic patterns rather than on terminology explicitly linked to identity.

Furthermore, in order to mitigate the effects of low frequency and ensure the relevance of the analyzed units, we introduce minimum frequency thresholds and restrict the analysis to equivalence classes whose frequency exceeds a given value. Units with very low occurrences tend to produce unstable estimates and lack statistical relevance. Since units with a higher number of content-bearing tokens are naturally less frequent, thresholds are set differently

according to the number of non-stopword tokens in each unit.

5.2 Definition of Candidate Fragments

We define a set of fragments that meet previously established conditions. These conditions aim to delimit comparable textual units that are relevant for analysis. In particular, they may include structural constraints such as a fixed number of sentences, the presence of explicit markers associated with the group of interest, or other relevant criteria.

Based on this set of fragments, a scoring scheme is constructed to estimate the concentration of linguistic patterns that are differentially associated with the groups.

5.3 Fragment Scoring

In order to identify fragments that concentrate a higher density of bias, we propose a scoring scheme based on the BS values calculated over the previously defined equivalence classes. This score allows us to establish a priority ranking among fragments according to their degree of differential association with the groups under comparison.

Formally, given a fragment F, we consider the n-grams appearing in it and associate them with their respective equivalence classes. Under the assumption that longer units provide greater contextual specificity, we avoid double counting by discarding n-grams that are completely contained within longer n-grams present in the same fragment and at the same position.

The fragment score $Score(F)$ is defined as the sum of the BS values of the equivalence classes associated with the n-grams in the fragment that belong to equivalence classes considered in the analysis. Formally, let a denote an equivalence class associated with an n-gram in F then:

$$Score(F) = \sum BS(a)$$

where the sum runs over the occurrences of n-grams in F that are associated with equivalence classes that were not filtered. In particular, when multiple n-grams belonging to the same equivalence class appear in different positions of the fragment, their contributions are accumulated. In cases where $BS(a)$ is not finite, its value is replaced by the finite value with the largest absolute magnitude observed among the equivalence classes considered in the analysis (i.e., those that were not filtered).

It is important to note that frequency thresholds are adjusted according to the number of content-bearing tokens in each equivalence class. Classes with more words tend to be less frequent and, therefore, exhibit greater variability in BS; however, these longer units provide more specific and relevant information to identify differential discursive patterns.

Consequently, fragments are ranked according to the density of linguistic units differentially associated with the groups under comparison. Fragments ranked highest are more strongly associated with the group of interest, whereas fragments ranked lowest are more strongly associated with the control group. It is important to emphasize that this ranking mechanism allows expert evaluation to focus on segments most likely to contain relevant linguistic patterns.

6 Proof of Concept with Disability in Spanish

As a proof of concept, we have carried out an initial exploration on representations related to persons with disabilities. Unlike other bias dimensions that have been more extensively studied, disability has received comparatively less attention in LLM analysis, and associated biases often manifest in particularly subtle and narratively mediated ways (Phutane, Seelam, and Vashistha, 2025; Panda, Agarwal, and Patel, 2025). This case illustrates the potential of the method to detect discursive patterns that do not emerge through isolated lexical analyses.

To explore the behavior of the proposed methodology, we generated a set of stories in Spanish using the model `llama-3.3-70B`. The corpus was organized into contrastive pairs: in one set of stories the character is explicitly described as having a disability, while in the control set the same narrative context is maintained without that reference. In both cases, the scenario and narrative task remain constant, and only the marker associated with the character is modified. To do that, we used prompts such as those in Figure 2.

Escribe una historia sobre Ana, **una persona ciega** que tiene una reunión de trabajo.
Write a story about Ana, ***a blind person*** *who has a work meeting.*
Escribe una historia sobre Ana, **una persona con anteojos** que tiene una reunión de trabajo.
Write a story about Ana, ***a person wearing glasses*** *who has a work meeting.*

Figure 2: An example prompts to generate contrastive stories with constant scenarios (in this case, a work meeting) with a person with disability (in this case, a blind person) and a person with no disability (in this case, a person wearing glasses), marked in boldface in the example.

For fragment prioritization, we first defined candidate fragments consisting of three sentences extracted from the stories associated with the group of interest. We additionally required that the central sentence contain a term belonging to the disability group, in order to analyze the immediate discursive context in which this marker is introduced.

Regarding the minimum frequency thresholds for including equivalence classes in the analysis, we set 40 occurrences for unigrams, 20 for classes with two non-stopword tokens, 15 for those with three non-stopword tokens, and 10 for classes with four non-stopword tokens. These values reflect that longer units naturally tend to occur less frequently. To reduce statistical noise, we discarded equivalence classes whose absolute BS value was lower than 0.5, a threshold corresponding approximately the standard deviation observed in the distribution of values obtained from random corpus partitions.

To illustrate the type of discursive patterns identified by the proposed methodology, we present two example fragments in Figure 3. The first corresponds to a fragment associated with a high positive BS value, while the second corresponds to a fragment with a negative BS value.

We observe that the fragment with the highest Score presents a type of representation that may be problematic, particularly in the form of *inspirational porn* (Gadiraju et al., 2023; Li et al., 2024). The equivalence classes contributing most to the Score include expressions emphasizing personal overcoming and inspiration ("cualquier persona puede superar" *anyone can overcome*, "convirtió en una inspiración" *turned into an inspiration*, "alcanzar sus metas" *reach their goals*), suggesting the presence of this type of discriminatory representation. In contrast, the fragment with a negative Score contains expressions centered on everyday experiences ("compartían risas" *shared laughs*, "disfrutar del paisaje" *enjoy the scenery*), which do not emphasize individual effort or personal abilities.

As shown in Tables 1 and 2, the equivalence classes associated with each group help identify specific linguistic associations that differentiate these narratives. These classes serve as valuable indicators, both independently and as contributors to the detection of text fragments that may contain biased or potentially harmful content.

It is important to note that some equivalence classes may group n-grams that, although sharing the same lemmas, combine them in ways that produce different meanings. For example, the class that includes expressions such as "adecuado y la determinación" (*adequate and the determination*), "determinación adecuada" (*adequate determination*), and "determinación y el adecuado2 (*determination and the adequate*) combines the same lemmas ("adecuado" and "determinación"), but the word order slightly changes the focus or intent of the expression. Even so, we consider that using equivalence classes is preferable to analyzing individual n-grams, as it allows n-grams sharing a common underlying idea to be grouped into a single unit of analysis, reducing dispersion and facilitating the detection of linguistic patterns.

7 Discussion

In this work, we present a methodology for the analysis of discursive bias in language models, combining contrastive generation, statistical analysis, and the retrieval of contextualized fragments. In particular, the approach shifts the focus from isolated lexical units to broader discursive configurations, facilitating the identification of patterns that emerge at the narrative level.

Fragment with high positive BS

Ana demostró que, con determinación y el adecuado apoyo, cualquier persona puede superar los obstáculos y alcanzar sus metas, independientemente de las barreras que pueda encontrar. Su historia se convirtió en una inspiración para muchos, mostrando que la discapacidad no es un límite para el éxito, sino una parte de la diversidad humana que enriquece nuestras comunidades. Después de graduarse, Ana continuó su camino hacia el éxito, trabajando en proyectos innovadores y contribuyendo a hacer del mundo un lugar más inclusivo y accesible para todos.

Fragment with high negative BS

Se detuvo un momento para disfrutar del paisaje y luego continuó su camino por el sendero principal. La silla de ruedas se deslizaba suavemente sobre la superficie plana, permitiéndole avanzar con facilidad. A medida que avanzaba, Juan se encontró con otros visitantes del parque: familias con niños que jugaban en los espacios de juego, parejas que paseaban de la mano y amigos que compartían risas y conversaciones.

Figure 3: Example of fragments of texts with high positive BiasScore (strongly associated to the interest group, in this case, people with disabilities) and high negative BiasScore (strongly associated with the control group, in this case, people without any explicit mention of disabilities).

A central aspect of the proposed approach is the use of equivalence classes as analytical units. These classes reduce morphological variability and organize information in a way that supports both quantitative analysis and qualitative interpretation. They also help make explicit the linguistic signals underlying the detected patterns, facilitating the interpretation of results and their subsequent evaluation by domain experts.

It is important to note that the construction of these equivalence classes presents certain limitations. In their current form, some classes may be fragmented or group expressions whose semantic relationship is not fully homogeneous, which may affect both the precision of the analysis and its interpretability. Nevertheless, these units enable the identification of regularities that are unlikely to emerge when considering n-grams in isolation. In this regard, future work could explore strategies to improve grouping criteria and strengthen the internal consistency of these classes.

Overall, these elements open up avenues for future research aimed at refining analytical units, evaluating the robustness of the approach across domains, and advancing validation frameworks that systematically incorporate expert participation.

This methodology is specifically aimed to operationalize qualitative evaluation and collective participation, systematically producing evidence that can be critically examined by domain experts and potentially affected communities. Thus it is not experts that need to systematize, but they just need to exercise their judgement. It is the methodology that systematizes the exploration, the detection of statistically significant patterns, and experts only need to determine the possible harms in story fragments.

In contrast with other approaches, this approach does not pre-determine which are indicators of bias, but they are discovered from texts, they do not have a pre-determined form but can be arbitrary in form, and the intervention of experts is always at a level of abstraction or concretion that feels natural to them, and contextualized: in describing groups and situations that may be portrayed differently by llms for different groups, producing potential harms, and in judging the harm produced by fragments that are contextualized within full texts. In contrast with fully contextualized harms assessments, that usually require immersion in a given context or task, this methodology is less costly, more generalizable, and allows for prospective analysis before deployment.

References

Aka, O., K. Burke, A. Bauerle, C. Greer, and M. Mitchell. 2021. 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, New York,

Equivalence Class	BS
"adecuado y la determinación", "adecuada y la determinación", "determinación y el adecuado", "adecuadas y la determinación", "determinación adecuada"	4.42
"barreras que podían", "barreras que podría", "barreras que puedan", "barrera que pudiera", "barreras que pueda", "barreras que podrían", "barrera puede", "barrera podía", "barreras podrían", "barreras podían", "barreras pueden"	4.42
"inclusivo y accesible", "accesible y inclusivo", "inclusiva y accesible", "accesible e inclusiva", "accesible y inclusiva"	4.42
"adecuada y el apoyo", "adecuadas y el apoyo", "adecuado y el apoyo", "apoyo adecuado", "adecuado apoyo", "apoyo adecuados"	4.42
"cualquier persona puede superar"	3.38
"convirtió en una inspiración", "convertiría en una inspiración", "convirtió en inspiración"	1.57
"obstáculos y alcanzar", "obstáculo y alcanzar", "obstáculo para alcanzar", "obstáculos y alcanzando", "obstáculos para alcanzar"	1.13
"encontrar que podía", "pueden encontrar", "podía encontrar", "podría encontrar", "pudiera encontrar", "poder encontrar", "pudo encontrar", "podido encontrar", "podemos encontrar", "puede encontrar", "pueda encontrar", "puedo encontrar", "podrían encontrar", "podían encontrar", "puedan encontrar"	1.05
"convirtió en una historia", "historia se convirtió", "historia se convertiría"	1.03
"muchas metas que alcanzaría", "alcanzar sus metas", "alcanzar metas tan", "metas que alcanzaría", "alcanzar su meta", "alcanzó su meta", "alcanzó las metas", "alcanzar metas", "alcanzando metas", "alcanzaran esas metas"	1.00
"sino como una parte", "sino también una parte", "sino una parte", "sino como parte", "sino parte"	0.96
"lugar en el mundo", "mundo un lugar", "lugares y mundos"	0.89
"hacía que el mundo", "hace que el mundo", "hacer que el mundo", "mundo la hizo", "hacer del mundo", "haciendo del mundo", "mundo y hacer", "hacer el mundo", "mundo lo hizo", "mundo haciendo"	0.74

Table 1: Equivalence classes contributing to the high positive BS fragment in Figure 3 (expressions more strongly associated to the group of people with disabilities).

Equivalence Class	BS
"compartir tanto las risas", "compartir una risa", "risas que compartiría", "compartieron una risa", "compartir risas", "compartían risas", "compartiendo risas", "risas compartieron", "compartieron risas", "compartió risas"	-0.97
"conversación y la risa", "risas y las conversaciones", "risa y las conversaciones", "risas y conversaciones", "risas y conversación"	-0.65
"disfrutar de ese momento", "disfrutar de su momento", "disfrutaron de ese momento", "disfrutar de un momento", "disfrutando del momento", "disfrutaba del momento", "momento de disfrutar", "momento para disfrutar", "disfrutar los momentos", "disfrutar del momento", "disfrutar de momentos", "disfrutó del momento", "disfrutaron del momento", "disfrutar momentos"	-0.57
"disfrutar del paisaje", "disfrutando del paisaje", "disfrutaba del paisaje", "disfrutó del paisaje"	-0.57

Table 2: Equivalence classes contributing to the negative BS fragment in Figure 3 (expressions more strongly associated to the group of people without any explicit mention of disabilities).

NY, USA. Association for Computing Machinery.

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

Bali, S., F. Farsi, M. Hosseini, A. Khorramrouz, and E. Asgari. 2026. Detecting subtle biases: An ethical lens on underexplored areas in AI language models biases. In V. Demberg, K. Inui, and L. Marquez, editors, *Proceedings of the 19th Conference of*

the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7352–7379, Rabat, Morocco, March. Association for Computational Linguistics.

Bender, E., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? pages 610–623, 03.

Bordia, S. and S. R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In S. Kar, F. Nadeem, L. Burdick, G. Durrett, and N.-R. Han, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Caliskan, A., J. J. Bryson, and A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. volume 356, page 183–186. American Association for the Advancement of Science (AAAS), April.

Gadiraju, V., S. Kane, S. Dev, A. Taylor, D. Wang, R. Denton, and R. Brewer. 2023. "i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 205–216, New York, NY, USA. Association for Computing Machinery.

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

Grue, J. 2016. The problem with inspiration porn: A tentative definition and a provisional critique. *Disability & Society*, 31(6):838–849.

Hirsch, M., M. Elichiry, B. Radi, T. Quiroga, D. Restrepo, L. Benotti, V. Xhardez, J. Dunstan, and E. Ferrante. 2026. Implicit bias in llms for transgender populations.

Kotek, H., R. Dockum, and D. Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24. ACM, November.

Kumar, A., S. Yunusov, and A. Emami. 2024. Subtle biases need subtler measures: Dual metrics for evaluating representative and affinity bias in large language models.

Li, R., A. Kamaraj, J. Ma, and S. Ebling. 2024. Decoding ableism in large language models: An intersectional approach. In D. Dementieva, O. Ignat, Z. Jin, R. Mihalcea, G. Piatti, J. Tetreault, S. Wilson, and J. Zhao, editors, *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 232–249, Miami, Florida, USA, November. Association for Computational Linguistics.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. A survey on bias and fairness in machine learning. 54(6).

Miceli, M., A.-A. Dinika, K. Kauffman, C. Salim Wagner, L. Sachenbacher, A. Hanna, and T. Gebru. 2025. Methodological considerations for centering workers' epistemic authority in ai research. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2):1698–1710, Oct.

Nadeem, M., A. Bethke, and S. Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August. Association for Computational Linguistics.

Nangia, N., C. Vania, R. Bhalerao, and S. R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–

1967. Association for Computational Linguistics, November.

Panda, S., A. Agarwal, and H. L. Patel. 2025. Accesseval: Benchmarking disability bias in large language models.

Parrish, A., A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. R. Bowman. 2022. Bbq: A hand-built bias benchmark for question answering.

Phutane, M., A. Seelam, and A. Vashistha. 2025. "cold, calculated, and condescending": How ai identifies and explains ableism compared to disabled people. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 1927–1941. ACM, June.

Tamkin, A., A. Askell, L. Lovitt, E. Durmus, N. Joseph, S. Kravec, K. Nguyen, J. Kaplan, and D. Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions.

Valentini, F., G. Rosati, D. Blasi, D. Fernandez Slezak, and E. Altszyler. 2023. On the interpretability and significance of bias metrics in texts: a PMI-based approach. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 509–520, Toronto, Canada, July. Association for Computational Linguistics.

Viveros Vigoya, M. 2016. La interseccionalidad: una aproximación situada a la dominación. *Debate Feminista*, 52:1–17.

Wan, Y., G. Pu, J. Sun, A. Garimella, K.-W. Chang, and N. Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore, December. Association for Computational Linguistics.

Zhao, J., Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. 2018. Learning gender-neutral word embeddings. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November. Association for Computational Linguistics.