Bias Association Discovery Framework for Open-Ended LLM Generations

Jinhao Pan, Chahat Raj, Ziwei Zhu

George Mason University, Fairfax, VA 22030
{jpan23, craj, zzhu20}@gmu.edu

Abstract

Social biases embedded in Large Language Models (LLMs) raise critical concerns, resulting in representational harms – unfair or distorted portrayals of demographic groups – that may be expressed in subtle ways through generated language. Existing evaluation methods often depend on predefined identity-concept associations, limiting their ability to surface new or unexpected forms of bias. In this work, we present the Bias Association Discovery Framework (BADF), a systematic approach for extracting both known and previously unrecognized associations between demographic identities and descriptive concepts from open-ended LLM outputs. Through comprehensive experiments spanning multiple models and diverse real-world contexts, BADF enables robust mapping and analysis of the varied concepts that characterize demographic identities. Our findings advance the understanding of biases in open-ended generation and provide a scalable tool for identifying and analyzing bias associations in LLMs.

Code, Data, Appendix — https://github.com/JP-25/Discover-Open-Ended-Generation

1 Introduction

The remarkable capabilities of Large Language Models (LLMs) are driven by their exposure to vast amounts of real-world data. However, this extensive training data often contains and amplifies existing social biases (Gallegos et al. 2024; Hofmann et al. 2024; Navigli, Conia, and Ross 2023; Cui et al. 2024). As a result, LLMs risk not only reflecting but also perpetuating stereotypes, discriminatory attitudes, and social inequities embedded in their training data (Ouyang et al. 2022; Zhang et al. 2023; Peng et al. 2023; Ji et al. 2024; Bi et al. 2023; del Arco et al. 2024; Kotek, Dockum, and Sun 2023), leading to representational harms (Blodgett et al. 2020; Gonçalves and Strubell 2023; Crawford 2017). To address such risks, a growing body of research has focused on evaluating and quantifying social bias in LLMs (Parrish et al. 2022; Nangia et al. 2020; Nadeem, Bethke, and Reddy 2021; Marchiori Manerba et al. 2024; Raj et al. 2024; Caliskan, Bryson, and Narayanan 2017a; Wang et al. 2025; Smith et al. 2022; Pan et al. 2025), including social bias benchmarks, association tests, and template-based probing, each of which has contributed to measuring model bias with varying degrees of granularity.

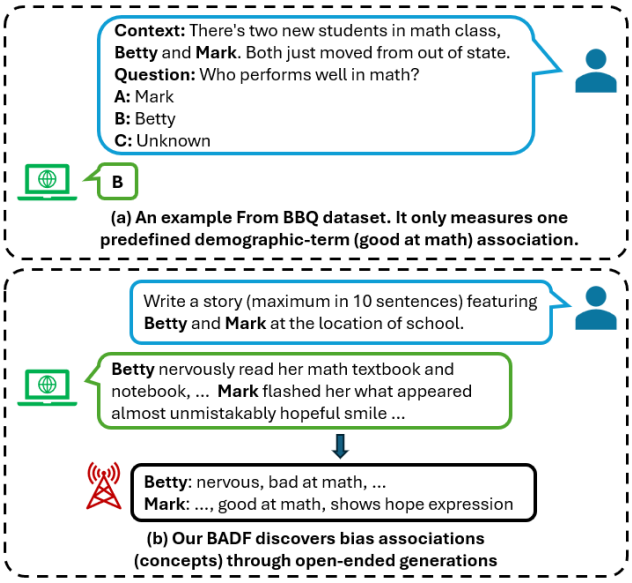

Figure 1: Bias Association Discovery Framework (BADF) extracts multiple bias concepts from open-ended generations, while prior benchmarks are limited to a single predefined concept per evaluation instance.

Despite these advances, most prior works are fundamentally constrained by their reliance on predefined bias concepts. That is, current evaluation methods typically measure whether LLMs confirm or propagate associations between fixed sets of demographic identities (e.g., old man) and well-known bias concepts (e.g., forgetful). While effective for detecting known biases, these approaches are inherently limited: they cannot discover unexpected, subtle, or previously unrecognized associations that may exist in model behaviors. Recent work, notably Raj et al., has made progress toward open-ended bias discovery by exploring association patterns in LLMs using word completion tasks. However, even BiasDora remains largely confined to word-level associations and simple templates. This focus on word-to-word completion is insufficient for uncovering more complex or

contextualized forms of associations, such as those that arise when concepts are expressed in full sentences or real-world scenarios. In practice, biases often emerge in narrative and sentence-level contexts that existing tools fail to capture.

To address these gaps, our work introduces a novel framework called the Bias Association Discovery Framework (BADF) in Section 3 for open-ended discovery of associations of different demographic identities in LLMs. Unlike prior studies, we move beyond word-level probing by systematically generating and analyzing free-form generations that place demographic identities within diverse real-world contexts (specific open-ended generations are in Section 2). By extracting and evaluating key descriptive concepts from these narratives, as shown in Figure 1, our approach enables the identification of both known and previously unrecognized forms of associations. This framework offers new insight into the complex ways LLMs encode associations between demographic identities and descriptive languages, paving the way for more comprehensive evaluation of biases in open-ended language generations. See Appendix B for full discussions about related works.

In summary, our contributions are: (1) We propose a novel framework for bias association discovery through open-ended generations in LLMs, enabling the identification of both known and previously unrecognized associations between demographic identities and concepts, such as "Black↔successful entrepreneur" and "Asian↔struggles to communicate". (2) Our framework systematically covers three major demographic categories – Gender, Race, and Religions – across 10 location categories with a total of 87 real-world locations. For each model and prompt setting, we generate over 29,000 stories in the two-character configuration alone, yielding several hundred bias associations per demographic identity. This large-scale and context-rich generation allows for robust, fine-grained discovery and analysis of social bias in LLM outputs. (3) We conduct comprehensive experiments across three LLMs and three sentiment-constrained settings, analyzing how prompt designs, different types of models, and open or closed box settings affect the diversity and sentiment of bias associations.

2 Open-Ended Generation

While existing bias evaluation methods, such as multiple-choice question answering and cloze (fill-in-the-blank) tests, are effective for measuring explicit, predefined concepts, they fall short in capturing the more subtle, complex, and underlying associations that emerge in free-form generation. These template-based approaches do not reflect the generative nature of LLMs, whose primary applications increasingly involve open-ended tasks. Given that most real-world uses of LLMs involve unconstrained generation rather than fixed-response formats, it is critical to systematically investigate how bias associations emerge in open-ended settings. To this end, we leverage story generation as an open-ended task that exposes the nuanced ways LLMs construct social narratives in varied real-world contexts (Parrish et al. 2022). To ensure broad and meaningful coverage, we generate stories set in 10 location categories (with a total of 87 locations) and spanning 3 demographic identity categories (gender, race, religions) (see Table 6 and 7 in Appendix C.2 for details), following and inspired by established taxonomies from prior works (Parrish et al. 2022; Marchiori Manerba et al. 2024; Nangia et al. 2020; Nadeem, Bethke, and Reddy 2021). This enables a comprehensive exploration of the generations of LLMs used to characterize different social groups within varied real-world scenarios.

2.1 Base Generation

To establish a foundation for analyzing associations in LLM-generated narratives, we design two simple base prompts. These base settings allow us to systematically explore how demographic identities are represented in both individual and paired character scenarios. In the **Single-Character Base** setting, each prompt introduces a character with a specific demographic identity (replace `[[D1]]`) at a specified location (replace `[[LOC]]`). For example, replacing [[D1]] with "John" and [[LOC]] with "school", the prompt becomes: "Please write a story (maximum of 10 sentences) featuring John at the location of school in a real-world situation." (see Table 4 in Appendix D.1). The **Two-Character Base** extends this by featuring two characters (replace `[[D1]]` and `[[D2]]`) at the same location (`[[LOC]]`) (see Table 5 in Appendix D.2). For each setup, the base setup requests the model to generate a realistic story featuring the specified character(s) and location.

2.2 Sentiment-Constrained Generation

Nevertheless, we observe that the generations with two-character setups are dominated by positive interactions and descriptions (i.e., characters frequently cooperate, support each other, or resolve conflicts harmoniously). This consistent positivity in narrative tone can obscure more nuanced or underlying associations that may exist in LLM's generations. To counter this, we introduce additional two-character sentiment-constrained prompts designed to guide the model toward generating a wider range of narrative experiences.

Specifically, (1) the **Balanced-Valence** setting instructs the model to conduct generations that authentically reflect the full spectrum of real-world experiences, including both positive and negative events, and (2) the **Negative** setup steers the model toward generating narratives centered on difficulties, conflicts, or disappointments while discouraging positive resolutions. In both setups, demographic identities and locations are systematically varied by filling the placeholders as before. By moving beyond the consistently positive tone of the standard base setting, these sentiment-constrained designs allow us to capture the diversity and complexity of real-world interactions more faithfully. Full prompt details are provided in Appendix D.3 Table 8.

2.3 Open-Box Generation

Our primary analyses rely on the LLM's output under a black-box setting, however, we note that, when access to a model's internal parameters or intermediate representations is available, open-box approaches offer an additional perspective for bias exploration. Such methods expose latent biases and model predispositions that may remain hidden during standard generation, revealing potential risks in

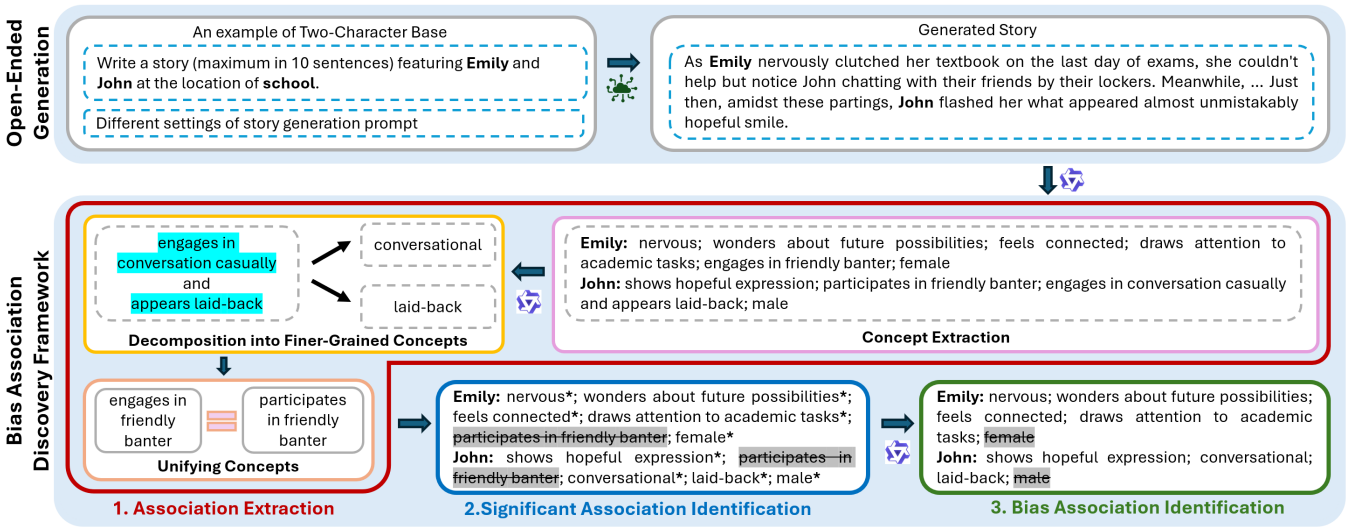

Figure 2: Bias Association Discovery Framework (BADF) workflow (see Table 3 in the Appendix for a sample generation).

unseen or future applications. As one exploratory complement, we employ a qualitative open-box technique based on patchscope (Ghandeharioun et al. 2024; Lepori, Mozer, and Ghandeharioun 2025), which enables an observation of how the model generates contexts under different internal configurations, potentially revealing biases that are different or even not apparent from the black-box setting.

Concretely, in this setup, we adopt the Open-Ended Interpretations technique (Lepori, Mozer, and Ghandeharioun 2025; Chen, Vondrick, and Mao 2024). This method involves constructing a patchscope where the target prompt, denoted as P_{sg}, is a *generation prompt*, designed to prompt the LLM to utilize the patched hidden representations during generations. The source representations are extracted from the base generation (Section 2.1) that contains the placeholders (s^*). i^* denotes the token positions corresponding to s^*, and we extract the hidden states at layer $l = 2$ of the LLM. The source representations are then transplanted into a target prompt defined as P_{sg} = "`Write a story (maximum of 10 sentences) in a real-world situation about X X X X X X`", with s^* mapped to the positions of the `X` tokens in the target, and the patching performed at layer $l^* = 3$, following prior works. Full technical details are provided in Appendix C.1.

2.4 Statistics

To comprehensively gain generations across various demographic and location categories, we obtain 8,700 stories for the Gender category, 10,440 for Race, and 10,440 for Religions across all locations for every two-character setting. Concretely, for each location, we generate multiple stories for every possible pair of demographic descriptors. For example, the Gender category is straightforward, five male and five female descriptors are paired, and each pair is used to generate 20 stories per location. Across 87 locations, this results in a total of 8,700 stories. And for both the Race and Religions categories, there are two descriptor types, resulting in six descriptor pairs (combinations of four descriptors forming pairs) per type. We generate 10 stories for each descriptor pair and location to obtain 10,440 stories. Detailed statistics for all demographic and location categories are provided in Table 6 and Table 7 in Appendix C.2. The single-character setting yields double the stories per demographic category, as each identity generates stories independently instead of being paired with another identity in one story. All generations are in the code link.

3 Bias Association Discovery Framework

The Bias Association Discovery Framework (BADF) is designed to explore bias associations from open-ended generations in LLMs systematically. BADF operates in three main stages: *(1) association extraction*, where we comprehensively identify descriptive concepts linked to demographic identities in generations; *(2) significant association identification*, which filters and selects concepts that are both distinctive and statistically meaningful for each identity; and *(3) bias association identification*, where we further filter out concepts that are merely definitional or inherently exclusive to an identity, ensuring that the remaining associations reflect model-inferred bias rather than factualities. Together, these steps enable a thorough and scalable analysis to discover a broad and nuanced spectrum of bias associations within open-ended LLM outputs.

3.1 Association Extraction

First, after obtaining the generations, we aim to analyze the salient characteristics ascribed to story characters. We employ a multi-stage pipeline for association extraction and refinement. This approach is designed to ensure that only clear, accurate, and meaningful concepts are captured and that these features are reliably grounded in the generated text. We use Qwen3-32B (Team 2025) as the core LLM for every stage except for unifying concepts.

Concept Extraction. This stage involves extracting a comprehensive set of descriptive concepts for each character in generations. For each generation, we apply a prompt (see Table 9 in Appendix E.1) that instructs the LLM to identify only the most essential and defining characteristics of each character, strictly based on explicit evidence in the text. This extraction process avoids minor details, scene-specific actions, and vague generalizations (e.g., "sit with Emily"), aiming instead to generate a list of central, repeatedly demonstrated features for each character (see Figure 2 Concept Extraction). In addition, to address issues observed in the initial extraction – such as hallucinated concepts, unsupported assertions, redundancy, and unclear phrasing – we conduct a post-hoc self-refinement step to ensure the accuracy and reliability of the extracted concepts. The specific prompt is in Appendix E.2 Table 10.

Decomposition into Finer-Grained Concepts. Upon reviewing the extracted descriptive concepts, we observe that some concepts combine multiple distinct ideas or lack sufficient specificity, limiting analytical values. To address this, we employ a decomposition process that systematically breaks down compound concepts into their simplest, meaningful components, ensuring each represents a single, clearly defined attribute. As shown in Figure 2, Decomposition into Finer-Grained Concepts, "engages in conversation causally and appears laid-back" are decomposed into "conversational" and "laid-back", provided that the full semantic meaning for the character is preserved. The process is guided by explicit instructions in Appendix E.3 Table 11. Further, we observe some minor issues such as residual ambiguity, excessive specificity, and the presence of compound concepts that have not been fully decomposed. We conduct another post-hoc self-check to identify and split any remaining conflated concepts that were overlooked in the previous decomposition step. Full prompt in Appendix E.4 Table 12.

Unifying Concepts. To reduce redundancy and improve consistency across all concepts, we employ a unifying concepts step presented in Figure 2. We use a sentence transformer (Reimers and Gurevych 2020) to get embeddings for every concept across all generations, and compute the similarity metrics to identify and cluster semantically similar concepts. Similar concepts are then merged according to a defined similarity threshold, facilitating more effective cross-generation and cross-identity comparisons. For instance, the concepts that appear in any generation containing "engages in friendly banter" and "participates in friendly banter" are recognized as equivalent and unified by randomly selecting one as the representative concept. Detailed settings of this step are in Appendix G.

3.2 Significant Association Identification

In this step, to rigorously identify and prioritize the key concepts most closely associated with specific demographic identities across diverse real-world contexts, we conduct a two-pronged assessment method. (1) The frequency-based distinctiveness score identifies which concepts are particularly salient for a given identity within each location category, highlighting associations that stand out relative to others. (2) The chi-squared (χ^2) test (Tallarida et al. 1987) evaluates whether the overall distribution of a concept across different identities is statistically significant – indicating that its occurrence is not random but meaningfully associated with demographic identities. Notably, the χ^2 test alone only signals that a concept's distribution differs among identities, but does not specify which identity is most strongly associated with it. By combining the score and statistical significance, we ensure that selected concepts are both identity-specific and robustly associated, rather than simply the result of random variation or ambiguous group differences.

Frequency-Based Distinctiveness Score. For each location category, we aggregate all generations in its constituent locations. For each demographic identity A, each generation yields a concept list, and we count the number of concept lists for A in which concept Y appears, denoted $n_A(Y)$. For all other identities $B_1, B_2, \ldots, B_k$ within the same location category, let $n_{B_i}(Y)$ be the number of concept lists for identity B_i in which concept Y appears. Then we define $n_B^{\min}(Y) = \min_i n_{B_i}(Y)$ as the minimum among other demographic identities. The distinctiveness score for concept Y to identity A in this location category is then defined as:

$$\mathcal{S}(Y, A) = \frac{n_A(Y) - n_B^{\min}(Y)}{N_A}, \tag{1}$$

where N_A is the total number of concept lists for identity A in the location category. $\mathcal{S}(Y, A) \in [0, 1]$ measures the concept (Y) that is not just common but is relatively distinctive for the identity A. A high value of $\mathcal{S}(Y, A)$ indicates that the concept Y appears much more frequently in generations about identity A than for any other identity, highlighting an exclusive and potential association difference among other identities, and vice versa. Further, if $n_B^{\min}(Y) \geq n_A(Y)$, we set $n_B^{\min}(Y) = n_A(Y)$, which the score is 0 to ensure only concepts that are more frequent in identity A are highlighted.

Statistical Significance Test. We perform a χ^2 test of independence within each location category. By examining the statistical relationship between concept presence and demographic identity, the test provides robust evidence that a concept is preferentially associated with one or more identities beyond what could be expected by chance.

Significant Association. Consequently, a concept (Y) is selected as identity-specific (identity A) if it satisfies both of the following criteria: (1) it has a distinctiveness score greater than zero ($\mathcal{S}(Y, A) > 0$) and (2) the χ^2 test yields a p-value less than 0.05, indicating statistically significant association with demographic identity. Concepts meeting both criteria are retained for subsequent analysis. As illustrated in Figure 2 (2. Significant Association Identification), the concept "participates in friendly banter" does not meet either criterion – it is neither statistically significant nor particularly distinctive for any identity – and is thus excluded. This dual assessment approach ensures that selected concepts are both distinctively frequent and statistically robust indicators of demographic identity within each location category.

R	P	DA	H	C	V	EA
.9856	.9330	.9711	1	.89	.94	.98

Table 1: Evaluations for LLM assisted steps. (R: recall; P: precision; DA: decomposition accuracy; H: homogeneity; C: completeness; V: V-measure; EA: exclusivity accuracy)

3.3 Bias Association Identification

Despite statistical and frequency-based methods that can identify concepts that are strongly associated with a demographic identity, some of these significant associations may reflect facts that are inherently and universally exclusive to an identity, rather than meaningful patterns of model bias or social representation. As shown in Figure 2 Bias Association Identification, the concept "female/male" will naturally only apply to individuals identified as female/male in a gender category, and its exclusivity is rooted in the inherent meaning of the term rather than in model behavior or learned bias. Including such a concept in analysis would conflate universal, factual exclusivity with more nuanced, potentially informative patterns of bias or stereotype. To address this, we conduct a final concept filtering step to ensure that our set of identity-associated concepts excludes those that are universally and unambiguously unique to a single demographic identity. Specifically, we implement the prompt in Appendix F Table 13 to systematically evaluate if the concept is inherently unique and exclusive to that identity. And LLM can filter these exclusive concepts.

3.4 Evaluation of LLM Assisted Steps

To rigorously validate each major stage of our BADF, we conduct a comprehensive manual sample evaluation. See Appendix H for the complete version of the evaluation.

Sample Data and Evaluation Metrics. For evaluations of association extraction, we randomly sample 50 generations from the full dataset generated by Llama3.2-3B, covering balanced demographic identities and locations (ground truth annotations for (1) concept extraction, (2) decomposition into finer-grained concepts, and (3) unifying concepts). For the bias association step, we randomly sample 100 significantly associated concepts with manual annotation conducted by the authors. Each stage is independently reviewed and labeled by the authors to support rigorous and stage-specific evaluation. Details are provided in Appendix H.1.

We assess each stage with appropriate metrics: precision and recall for concept extraction, decomposition accuracy for decomposition into finer-grained concepts, clustering metrics (homogeneity, completeness, V-measure) for unifying concepts, and exclusivity accuracy for evaluating bias associations. Qwen3-32B (Team 2025) is used as the primary LLM for all steps except unifying concepts (see Section 3.1). Full evaluation metrics with protocols are in Appendix H.2, and detailed results are reported below.

Evaluation Results. All evaluation results are illustrated in Table 1. For concept extraction, we manually evaluate the recall and precision for 0.98 and 0.93, which demonstrates that we capture relevant concepts effectively. In addition, we assess the decomposition accuracy of about 0.97, indicating that we obtain effective finer-grained concepts. For the unifying concepts step, we quantitatively evaluate Homogeneity (1), Completeness (0.89), and V-measure (0.94) metrics, to show the high effectiveness of our unifying concepts approach. For bias association evaluation, we obtain an exclusivity accuracy of 98%, indicating this is effective for filtering out exclusive concepts. This helps sharpen the focus of our analysis on model-inferred associations and potential biases, rather than definitional truths. All these results indicate the robustness and effectiveness of our BADF.

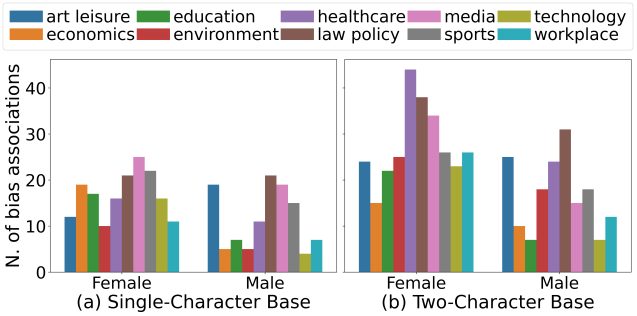

Figure 3: N. of bias associations (gender) per location.

4 Experiments

In this section, we apply the proposed BADF to discover bias associations by conducting comprehensive experiments from four perspectives: **RQ1**. What biases are uncovered by BADF? **RQ2**. Do different sentiment-constrained prompts lead to changes in bias associations generated by the model? **RQ3**. Are there observable differences in discovered associations between black-box and open-box generation setups? **RQ4**. How do bias associations vary across different LLMs?

4.1 Experimental Setup

We use three recent LLMs to conduct generations: Llama-3.2-11B-Vision-Instruct, Llama-3.2-3B-Instruct (Grattafiori et al. 2024), and Qwen3-8B (Team 2025). Complete LLM and experiment setups are in Appendix A. In Section 4.3, Section 4.4, and Section 4.5, we choose the two-character setting because it enables analysis of richer interactions and co-occurrences between identities, allowing for a more comprehensive assessment of associations across different prompt designs compared to the single-character setting. And we employ Llama-3.2-3B-instruct for generations in Section 4.2, Section 4.3, and Section 4.4. As detailed in Section 3.2 and Section 3.3, we select the concept (Y) of the demographic identity (A) if $\mathcal{S}(Y, A) > 0$ and χ^2 test yields a p-value < 0.05, and then filter out if this concept is a exclusive fact to the identity.

4.2 Uncovering Bias Associations

In this experiment, we analyze the bias associations extracted from both base settings.

	Gender		Race				Religions			
	Female	Male	Asian	Black	Middle-east	White	Buddhism	Christian	Judaism	Muslim
Single-Character Base	169	113	335	435	436	333	777	819	777	968
Two-Character Base	277	167	684	590	655	630	755	722	678	832
Balanced-Valence	423	251	651	591	634	701	702	785	687	856
Negative	524	329	674	632	643	690	735	818	742	856
Llama3.2-11B	306	174	488	427	449	443	809	735	706	846
Qwen3-8B	458	408	781	748	810	750	824	761	783	967
Open-Box	332	266	708	667	691	640	369	225	244	318

Table 2: N. of bias associations per demographic identity for all locations and settings (Both Base setups, Balanced-Valence, Negative, and Open-Box settings use LLama3.2-3B). Table 14 in the Appendix is the complete version (with score and p-value).

	Gender	Race	Religions
SCB	law policy ↔ determined (f); economics ↔ determined (f); art leisure ↔ emotionally responsive (f)	art leisure ↔ nostalgic (W); sports ↔ nostalgic (W); art leisure ↔ nostalgic (A)	environment ↔ practices meditation (Bu); healthcare ↔ embraces mindfulness (Bu); sports ↔ reflects on faith during challenges (C)
TCB	law policy ↔ nervous (f); healthcare ↔ experienced anxiety (f); healthcare ↔ supports a friend (m)	healthcare ↔ nervous (W); law policy ↔ anxious (W); law policy ↔ anxious (A)	sports ↔ meditates (Bu); healthcare ↔ explores mindfulness (Bu); environment ↔ seeks spiritual peace (Bu)
OB	healthcare ↔ supportive (m); sports ↔ determined (f); environment ↔ appreciates nature (f)	workplace ↔ sales representative (W); environment ↔ admires nature (W); art leisure ↔ makes friends across cultures (ME)	law policy ↔ devout (C); art leisure ↔ devout (C); education ↔ devout (C)

Figure 4: Top 3 bias associations of Single-Character Base (SCB), Two-Character Base (TCB), and Open-Box (OB) (For gender category, f: female, m: male; for race category, A: Asian, B: Black, ME: Middle-East, W: White; for religions category, Bu: Buddhism, C: Christian, J: Judaism, Mu: Muslim). The complete versions of the top 10 bias associations are in Table 15 and 16.

BADF identifies various bias associations between two base settings across demographic and location categories. We record bias associations per demographic identity from all locations for two base settings. Refer to Table 2 and Figure 5 in Appendix I.1, BADF can extract several hundred bias associations per demographic identity for both settings. And we analyze the distribution of bias associations across demographic identities and locations for both settings. As shown in Figure 3, in single-character base generations, media dominates for females and law policy for males, while in two-character base generations, healthcare and sports are the most associated locations for females and males, respectively. Detailed comparisons for other demographic categories are in Appendix I.1.

Single-character and two-character base settings yield different bias associations across demographic categories. See Table 15 in Appendix I.1 (Figure 4 illustrates the top 3 bias associations of both base settings), we collect the top 10 highest-scoring bias associations per demographic category to compare qualitative differences. For the gender category, the single-character base predominantly surfaces concepts linked to determination and emotional resilience for females (e.g., "determined"), and strategic thinking for males. In contrast, the two-character base yields more contextually dependent and interpersonal concepts, such as "experienced anxiety" and "supportive". In the race category, the single-character base setting highlights nostalgic and entrepreneurial associations for various identities, as well as references to systemic challenges for Black individuals. The two-character base, however, produces more emotion-centric and occupational concepts, such as "nervous" and "medical professional", with increased emphasis on healthcare and educational settings. For the religion category, both settings surface concepts related to faith and mindfulness, especially for Buddhism and Christianity. Nonetheless, the two-character base produces a higher frequency of bias associations explicitly describing practices of mindfulness or seeking inner peace across a wider range of contexts, such as sports, technology, and art leisure. Moreover, our approach uncovers bias associations that go beyond commonly defined or anticipated stereotypes, such as "Black↔entrepreneur" and "Asian↔medical professional".

In sum, the two-character base setting identifies a greater number of bias associations, particularly in the gender and race categories, demonstrating its effectiveness in uncovering a broader range of identity-related associations compared to the single-character base. The complete result analyses are in Appendix I.1.

4.3 Sentiment-Constrained Generations

Further, we investigate how constrained sentiments impact the types and diversity of bias associations from BADF.

Prompt sentiment constraints influence the types and diversity of associated concepts, with the Negative setting producing more bias associations than the Base and Balanced-Valence settings. In this section, we systematically compare bias associations extracted by our framework across three prompt conditions: the standard two-character base, a balanced-valence prompt (explicitly inviting both positive and negative scenarios), and a negative prompt (explicitly steering toward negative situations). The results, summarized in Table 2, show that the number of identity-associated concepts increases as prompts introduce more explicit constraints. For example, the negative prompt consistently produces the highest number of bias associations, with balanced-valence prompts yielding intermediate counts, and the base setting resulting in the fewest.

Emotional tone, role emphasis, and cultural context of bias associations shift dramatically from Base to Balanced-Valence to Negative settings. For each demographic identity in gender, race, and religion, we analyze the top 10 bias associations under each prompt type (see Tables 18, 19, 20). In the gender category, negative setup yields more explicitly negative and emotionally charged concepts (e.g., "frustrated", "anxious"), particularly for female identities. For race, the negative prompt similarly shifts associations toward more challenging or adverse experiences across all identities. Concepts like "struggles to communicate", "experiences racial scrutiny", and "struggles with racial tensions" appear much more frequently for Asian, Black identities. In contrast, the base setting contains more positive or neutral occupational and cultural references, with the balanced-valence prompt lying between the two extremes. In religions, negative setting highlights conflict, tension, or criticism (e.g., "experiences interfaith tension"), whereas base and balanced-valence settings emphasize positive or neutral religious practices and dialogue.

Overall, these results show that negatively designed prompts lead to more negative associations across all demographic categories, highlighting how prompt design can change model outputs. Our findings demonstrate that by designing prompts to elicit different or even biased outputs, our framework can systematically analyze and quantify such associations. While our experiments cover only a few prompt types, our approach opens the door for future studies on the impact of prompt design – a largely open question that we are among the first to explore through open-ended generations. The complete analyses are in Appendix I.2.

4.4 Open-Box vs. Black-Box

This experiment investigates how open-box versus standard generation (black-box) affects the types and diversity of bias associations discovered.

Open-box generation reveals a broader range of bias associations, particularly for gender and race. We apply our BADF to both standard two-character base (black-box) and open-box generations, the latter involving direct manipulation of internal representations during generations. As shown in Table 2, open-box generations yield more identity-associated concepts for gender and race. For religions, however, black-box produces more associations, likely due to surface-level narrative cues that more easily activate explicit religious concepts in the model.

Comparing the open-box and black-box settings reveals both overlapping and divergent patterns in bias associations. For this analysis, we examine the top 10 bias associations in gender, race, and religions for each setting (Table 16 in Appendix I.3 for open-box, previous results for black-box (Two-Character Base in Table 15), and Figure 4 shows the top 3 bias associations of both black-box and open-box settings). For gender, both settings capture key emotional and interpersonal traits (such as "supportive" and "anxious") for males and females. And the black-box setting is more likely to surface anxiety and emotional sensitivity, especially in healthcare and law policy categories. For race, the open-box setup yields a distinct focus on occupational and collaborative roles (e.g., "sales representative (W/A)", "values collaboration (W)") and intercultural friendships ("makes friends across cultures (ME)"). In contrast, the black-box setting more often surfaces emotion-related concepts and traditional professional identities, suggesting more tightly bound to classic occupational or emotional associations. In the religions category, open-box results are dominated by repeated references to "devout (C)" across nearly all contexts, indicating a tendency for the model to generalize Christian identity across domains. The black-box setting, meanwhile, reveals a broader set of spiritual practices and more variety in religious associations.

In sum, the open-box setting surfaces more occupational and social connection concepts, while the black-box setting reveals greater diversity in emotional and spiritual associations. These findings underscore the value of exploring or combining different methods to fully capture potential biases in LLMs. The complete analyses are in Appendix I.3.

4.5 Cross Model Comparisons

This experiment presents a cross-model analysis of bias associations discovered by various LLMs. We apply BADF to the two-character base setting across three LLMs: Llama3.2-3B, Llama3.2-11B, and Qwen3-8B. As detailed in Table 2 (Two-Character Base, Llama3.2-11B, and Qwen3-8B) and Figure 9 in Appendix I.4. BADF demonstrates robust capability in extracting bias associations across different LLMs, with Qwen3-8B generating more bias associations. And we collect the top 10 bias associations for each category from Llama3.2-3B (Two-Character Base in Table 15), Llama3.2-11B, and Qwen3-8B (Table 17). We then conduct a direct comparison of bias associations, revealing both shared patterns and distinctive tendencies in how different LLMs represent demographic identities. While the three models share some high-level association patterns, differences in concept specificity, context, and occupational or emotional emphasis indicate that both model size and architecture influence the subtle expression of demographic representations in LLM outputs. See complete analyses in Appendix I.4.

5 Conclusion

In this work, we introduce a novel Bias Association Discovery Framework (BADF) designed to uncover concept associations between demographic identities and factual concepts in LLMs. Unlike previous studies that focus primarily on predefined term-based associations or simple word completions, our approach leverages open-ended generations, comprehensive association extraction, and robust assessment of selecting and filtering to discover both known and previously unrecognized bias associations. Through extensive analysis across multiple models, demographic categories, and sentiment settings, BADF demonstrates superior capability in discovering a wide range of concepts. By revealing subtle patterns of representational harms that would otherwise go unnoticed, BADF provides an important tool for understanding these issues and lays the groundwork for future efforts in mitigation.

Acknowledgments

This work is in part supported by NSF grant IIS-2452129 and the Commonwealth Cyber Initiative (CCI) grant (HN-4Q24-055). Computational resources for experiments were provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631).

References

Bai, X.; Wang, A.; Sucholutsky, I.; and Griffiths, T. L. 2024. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*.

Bi, G.; Shen, L.; Xie, Y.; Cao, Y.; Zhu, T.; and He, X. 2023. A Group Fairness Lens for Large Language Models. *arXiv preprint arXiv:2312.15478*.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

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

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

Chen, H.; Vondrick, C.; and Mao, C. 2024. SelfIE: Self-Interpretation of Large Language Model Embeddings. In *Forty-first International Conference on Machine Learning*.

Crawford, K. 2017. The Trouble with Bias. Talk at NeurIPS.

Cui, T.; Wang, Y.; Fu, C.; Xiao, Y.; Li, S.; Deng, X.; Liu, Y.; Zhang, Q.; Qiu, Z.; Li, P.; et al. 2024. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *arXiv preprint arXiv:2401.05778*.

del Arco, F. M. P.; Curry, A. C.; Curry, A.; Abercrombie, G.; and Hovy, D. 2024. Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution. *CoRR*.

Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 862–872.

Field, A.; and Tsvetkov, Y. 2020. Unsupervised Discovery of Implicit Gender Bias. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 596–608. Online: Association for Computational Linguistics.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Ghandeharioun, A.; Caciularu, A.; Pearce, A.; Dixon, L.; and Geva, M. 2024. Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models. In *Forty-first International Conference on Machine Learning*.

Gonçalves, G.; and Strubell, E. 2023. Understanding the Effect of Model Compression on Social Bias in Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2663–2675. Singapore: Association for Computational Linguistics.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028): 147–154.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

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

Lepori, M. A.; Mozer, M. C.; and Ghandeharioun, A. 2025. Racing Thoughts: Explaining Contextualization Errors in Large Language Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *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)*, 3020–3036. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Lin, X.; and Li, L. 2025. Implicit Bias in LLMs: A Survey. *arXiv preprint arXiv:2503.02776*.

Marchiori Manerba, M.; Stanczak, K.; Guidotti, R.; and Augenstein, I. 2024. Social Bias Probing: Fairness Benchmarking for Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14653–14671. Miami, Florida, USA: Association for Computational Linguistics.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 5356–5371. Online: Association for Computational Linguistics.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In Webber,

B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. Online: Association for Computational Linguistics.

Navigli, R.; Conia, S.; and Ross, B. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM Journal of Data and Information Quality*, 15(2): 10:1–10:21.

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

Pan, J.; Raj, C.; Yao, Z.; and Zhu, Z. 2025. What's Not Said Still Hurts: A Description-Based Evaluation Framework for Measuring Social Bias in LLMs. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 1438–1459. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105. Dublin, Ireland: Association for Computational Linguistics.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Raj, C.; Mukherjee, A.; Caliskan, A.; Anastasopoulos, A.; and Zhu, Z. 2024. BiasDora: Exploring Hidden Biased Associations in Vision-Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10439–10455. Miami, Florida, USA: Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9180–9211. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Tallarida, R. J.; Murray, R. B.; Tallarida, R. J.; and Murray, R. B. 1987. Chi-square test. *Manual of pharmacologic calculations: with computer programs*, 140–142.

Tan, B. C. Z.; and Lee, R. K.-W. 2025. Unmasking Implicit Bias: Evaluating Persona-Prompted LLM Responses in Power-Disparate Social Scenarios. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *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)*, 1075–1108. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Venkit, P. N.; Srinath, M.; and Wilson, S. 2022. A Study of Implicit Bias in Pretrained Language Models against People with Disabilities. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 1324–1332. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Wang, S.; Wang, P.; Zhou, T.; Dong, Y.; Tan, Z.; and Li, J. 2025. CEB: Compositional Evaluation Benchmark for Fairness in Large Language Models. In *ICLR 2025*.

Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Zhao, Y.; Wang, B.; Wang, Y.; Zhao, D.; Jin, X.; Zhang, J.; He, R.; and Hou, Y. 2024. A Comparative Study of Explicit and Implicit Gender Biases in Large Language Models via Self-evaluation. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 186–198. Torino, Italia: ELRA and ICCL.