A Word Association Network Methodology for Evaluating Implicit Biases in LLMs Compared to Humans

KATHERINE ABRAMSKI*, Dept. of Computer Science, University of Pisa, Italy
GIULIO ROSSETTI, Institute of Information Science and Technologies, National Research Council, Italy
MASSIMO STELLA, Dept. of Psychology and Cognitive Science, University of Trento, Italy

Background: As large language models (LLMs) become increasingly integrated into our lives, their inherent social biases remain a pressing concern. Bias detection and evaluation methods have made great progress towards mitigating their harmful effects, but most existing methodologies exhibit important limitations: they commonly focus on explicit rather than implicit biases, they are often overly context specific, and crucially, they do not allow for comparisons with human benchmarks.
Objectives: The present work addresses several of these limitations. We present a novel word association network methodology for evaluating implicit biases in LLMs compared to humans. Our prompt-based approach taps into the implicit relational structures encoded in LLMs, measuring biases by simulating semantic priming within LLM-generated word association networks. Unlike most prompt-based evaluation methods, our comparative framework provides both quantitative and qualitative assessments of biases that can be directly compared to humans and other LLMs, enabling relative evaluations and offering new insights into the alignment of LLMs with human cognition.
Methods: The methodology involves building comparable word association networks using aligned datasets generated by English-speaking humans and LLMs. Within these networks, biases are evaluated by measuring the strength of association between words using a method that simulates semantic priming. Three different metrics are presented for evaluating biases in terms of: stereotypes, valence, and emotions. We apply our framework to both humans and several widely used LLMs to investigate a variety of implicit social biases.
Results: Our results reveal both convergences and divergences between LLM and human biases related to gender, religion, ethnicity, sexual orientation, and political party. These results provide new perspectives on the potential risks of using LLMs.
Conclusions: Our framework offers a novel approach to evaluating implicit biases in LLMs compared to humans, addressing several limitations of existing approaches. Our work contributes to a systematic, scalable, and generalizable approach that advances the goal of trustworthy and socially responsible language technologies.

JAIR Track: AI & Society

JAIR Associate Editor: Paolo Torroni

JAIR Reference Format:
Katherine Abramski, Giulio Rossetti, and Massimo Stella. 2026. A Word Association Network Methodology for Evaluating Implicit Biases in LLMs Compared to Humans. *Journal of Artificial Intelligence Research* 87, Article 28 (November 2026), 36 pages. DOI: 10.1613/jair.1.22120

*Corresponding Author.

Authors' Contact Information: Katherine Abramski, ORCID: 0009-0001-4210-5498, katherine.abramski@phd.unipi.it, Dept. of Computer Science, University of Pisa, Pisa, Italy; Giulio Rossetti, ORCID: 0000-0003-3373-1240, giulio.rossetti@isti.cnr.it, Institute of Information Science and Technologies, National Research Council, Pisa, Italy; Massimo Stella, ORCID: 0000-0003-1810-9699, massimo.stella-1@unitn.it, Dept. of Psychology and Cognitive Science, University of Trento, Rovereto, Trento, Italy.



This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2026 Copyright held by the owner/author(s).
DOI: 10.1613/jair.1.22120

1 Introduction

Large Language Models (LLMs) are rapidly transforming our lives by revolutionizing the way we produce and consume digital content (Padilla et al. 2025; Rossetti et al. 2026). Their accessibility and versatility have led to widespread use across diverse domains and NLP tasks—from writing emails and debugging code to making high-stakes decisions and participating in online discussions that influence opinion dynamics (Cau et al. 2025). Although LLMs hold great potential to enhance human productivity and understanding, they also raise serious societal concerns (Stella et al. 2023). In particular, LLMs have been shown to exhibit harmful social biases, such as gender and racial biases, that reflect negative or stereotypical perceptions of certain groups (Garimella et al. 2021). These biases, which are inherited from the massive amounts of human corpora that LLMs are trained on (Abramski, Citraro, et al. 2023; Bender et al. 2021; Dodge et al. 2021; Stella et al. 2023), mirror many of the social biases that humans possess (Acerbi and Stubbersfield 2023), and they have the potential to reinforce the harmful narratives that are at the core of many social inequalities.

Several mitigation strategies have been developed for reducing biases in LLMs (Shen et al. 2023), but they have mostly focused on addressing explicit biases (Lin and Li 2025; Y. Zhao, B. Wang, Y. Wang, D. Zhao, He, et al. 2025), which are overt and easily detectable. In contrast, implicit biases are far subtler and often escape detection, making them particularly challenging to study. Examples of the difference between explicit and implicit biases in LLMs are shown in Table 1. In humans, implicit biases reflect our subconscious associations and deep-rooted beliefs about the world, which often conflict with our explicit beliefs (Perugini 2005; Wilson et al. 2000). This contradiction explains why someone might believe that men and women are equally capable, while subconsciously preferring a male doctor to a female doctor. Similar to humans, implicit biases in LLMs can be nearly imperceptible (Stella et al. 2023), but they can manifest in subtle ways in downstream applications like question-answering, text-generation, and decision-making, even in models that exhibit little to no explicit bias (Abramski, Citraro, et al. 2023; X. Bai et al. 2025; Cau et al. 2025; Lin and Li 2025). Moreover, since there is evidence that implicit biases influence behavior more than explicit biases (Greenwald, Poehlman, et al. 2009; Stanley et al. 2011; Y. Zhao, B. Wang, Y. Wang, D. Zhao, He, et al. 2025), the cost of overlooking them can have serious consequences. This is especially a concern when LLMs are used in domains such as healthcare (Poulain et al. 2024), recruiting (Anzenberg et al. 2025), criminal justice (Struniawski et al. 2025), psychology (Binz et al. 2025), and education (Lee et al. 2024), where social identity plays a key role. In such contexts, the application of biased LLMs can lead to harmful discriminatory outcomes, such as AI systems suggesting lower salaries for women and minorities (Sorokovikova et al. 2025).

Another troubling and somewhat ironic consequence of LLM social biases has to do with the long-lasting impact they can have on our own biases. Recent studies have shown that as humans interact with biased LLMs, we subconsciously absorb the biases in their outputs, even long after we stop interacting with the models (Vicente and Matute 2023). These findings suggest the emergence of a feedback loop, in which LLMs acquire biases from humans, and through repeated exposure, we absorb new biases from LLMs. This dynamic is concerning, as it could potentially lead to a vicious cycle of reinforcing harmful social biases and even the emergence of new ones, especially if LLM biases are stronger than those possessed by humans. LLMs have also been shown to be more overconfident than humans when they are unsure about their answers, thus adding to the potential negative impacts that their biases could have on humans (Sun et al. 2025). These findings underscore the need to study LLM biases in relation to human biases, rather than evaluating models solely against neutral or artificial benchmarks.

The numerous harmful consequences of social biases in LLMs have led to an explosion of research dedicated to their detection and evaluation. In the last decade, great progress has been made towards this goal, but at the rate that LLMs are evolving, developing methodologies that remain relevant and effective is a continuous challenge. Because of the diversity of models, the growing number of their applications, and the various types of

Table 1. **Examples of explicit and implicit biases in LLMs.** Explicit biases are expressed overtly, for example as statements or stated reasons for a decision. Implicit biases are not stated, but emerge indirectly through the associations that a model makes between concepts. Explicit and implicit biases do not necessarily align: LLMs often exhibit implicit biases despite the absence of explicit biases. Note, the examples are illustrative.

Bias	Explicit bias (overt, stated)	Implicit bias (covert, through associations)
Religion	Asked whether Muslims are more violent than Christians, a model answers *"Yes, Muslims are more prone to violence."*	Asked to continue a story that begins *"Two Muslims walked into a…"*, a model writes violent continuations far more often than for other religious groups (Abid et al. 2021).
Gender	Asked whether women make worse doctors than men, a model answers *"Yes, men are better suited to be doctors, and women to be nurses."*	Asked to complete *"The doctor said that ___ would be late"*, a model chooses *he*; for *"The nurse said that ___ would be late"*, it chooses *she* (Kotek et al. 2023).
Ethnicity	Asked whether African Americans are less intelligent than White Americans, a model answers *"Yes, they are less intelligent."*	Given texts written in African American English, with no mention of race, a model describes their writers as *lazy*, *stupid*, and *ignorant* (Hofmann et al. 2024).

social biases that LLMs can exhibit, there is no one-size-fits all solution. Recent surveys on LLM bias evaluation methodologies emphasize the importance of a multifaceted approach that includes a wide range of evaluation methodologies that investigate biases from different perspectives (Gallegos et al. 2024; Y. Guo et al. 2024). Such a multifaceted approach enables a broader more comprehensive understanding of the problem, which is crucial for informing effective bias mitigation strategies (Gallegos et al. 2024; Y. Guo et al. 2024).

The word association network methodology that we present in this paper is a novel contribution to this aim. Our cognitive psychology-inspired methodology is an unexplored approach to evaluating implicit biases in LLMs that uses network models built from LLM-generated word associations. Our approach combines several elements from existing methodologies, but our method is unique in that it provides a metric that can be compared directly to humans, enabling parallel investigations between LLM and human cognition. Moreover, our methodology is flexible – it can be used to investigate many different types of biases (e.g. gender, ethnicity, religion, politics, etc.) – as well as scalable – it can be applied to nearly any LLM, regardless of the application. In the remainder of this section, we provide additional background and motivation for our methodology. We first discuss existing work in the field of LLM bias evaluation to provide a contextual understanding of how our methodology compares to existing approaches. We then discuss the theories and methodologies from cognitive psychology upon which our methodology is based.

1.1 Overview of Existing Work on LLM Bias Evaluation

Given the breadth and variety of research on this topic, the landscape of LLM bias evaluation methods can be difficult to navigate. Methodologies can be classified in many different ways based on the types of biases investigated and the data used to investigate them. An extensive review of bias evaluation methodologies is beyond the scope of this paper, and we refer interested readers to surveys on this topic (Gallegos et al. 2024; Y. Guo et al. 2024). Rather, for the purpose of providing a contextual understanding of our word association network methodology, we discuss existing approaches to LLM bias evaluation as they relate to four defining characteristics of our methodology. Specifically, our methodology:

(1) evaluates *intrinsic bias* as opposed to extrinsic bias

(2) uses an *output-level* approach rather than a model-level approach
(3) evaluates biases via *word associations*
(4) enables direct *comparisons with humans*

While none of these characteristics are unique to our methodology, the combination of them contributes to what distinguishes our approach from others. In what follows, we discuss the significance of these characteristics, focusing on the strengths and weaknesses of some of the methodologies that share them.

1.1.1 Intrinsic vs. Extrinsic Bias. Methods for evaluating biases can be distinguished by the type of bias that they aim to investigate. In LLMs, biases can be classified as either intrinsic or extrinsic (Y. Guo et al. 2024). Intrinsic biases are those that are embedded within the model's architecture, acquired from the training data. They represent the patterns and relationships that are encoded in the internal representations of a model, and they are independent of any downstream task (Y. Guo et al. 2024; Lum et al. 2024). Extrinsic biases are those that appear in applications and downstream tasks after fine-tuning; they can cause harm through biased outputs or decisions, and they manifest differently based on the downstream task (Y. Guo et al. 2024; Lum et al. 2024). Thus, intrinsic bias evaluation methods typically aim to investigate the model's internal representations directly via their embedding space (Caliskan et al. 2017; W. Guo and Caliskan 2021; Kurita et al. 2019; Nadeem et al. 2020), while extrinsic bias evaluation methods aim to investigate biases that appear in specific tasks and contexts like question answering and text generation (Parrish et al. 2021; Wan et al. 2023; J. Zhao et al. 2018). Although this distinction may resemble that between implicit and explicit bias, they are conceptually different. The intrinsic/extrinsic distinction concerns where the bias resides – within model representations or outputs – whereas the implicit/explicit distinction concerns how covertly or overtly bias is expressed. Both implicit and explicit biases can manifest in LLM outputs, thus, as extrinsic bias. However, implicit bias may be more closely aligned with intrinsic bias, since a model's representational structures represent latent associations, which form the basis of implicit bias. Recent works investigating the relationship between intrinsic and extrinsic evaluation metrics have found that the two measures do not always correlate, sparking skepticism about whether measures of intrinsic bias are really worthwhile (Cao et al. 2022; Goldfarb-Tarrant et al. 2020; Katô and Miyao 2024; Steed et al. 2022). These findings suggest that evaluations of intrinsic bias are not sufficient for understanding how biases will manifest in outputs, and that extrinsic bias metrics may be better at predicting how biases can cause harm in specific tasks (Cao et al. 2022; Steed et al. 2022). However, these studies focused on comparisons involving only *model-level* approaches, which access the model's representations directly via their embedding space. In contrast, our *output-level* approach to intrinsic bias evaluation offers an alternative which creates proxies for the model's internal representations through careful prompting (Abramski, Citraro, et al. 2023; Abramski, Lavorati, et al. 2024; X. Bai et al. 2025), mirroring the behavioral methodology that is used for investigating implicit biases in humans. In what follows, we dive into the distinction between model-level and output-level approaches to provide a better understanding of how they are typically applied for investigating intrinsic and extrinsic bias, respectively.

1.1.2 Model-Level vs. Output-Level Approaches. Methods for evaluating biases in LLMs can also differ by whether they operate at the model level, directly probing internal representations, or at the output level, treating the model as a black box and using outputs from prompting for investigating biases. Model-level methods are typically intrinsic since they directly access a model's representational structures, while output-level approaches are typically extrinsic, focusing on specific downstream tasks or specific types of bias in certain domains (e.g. gender bias in recruiting) (An et al. 2024). Model-level methods usually fall into one of two categories: embedding-based or probability-based (Gallegos et al. 2024). Embedding-based metrics such as the well-known Word Embedding Association Test (WEAT) (Caliskan et al. 2017) and similar variations (W. Guo and Caliskan 2021; May et al. 2019) apply the logic of the Implicit Association Test (IAT) (Greenwald, McGhee, et al. 1998), a well-known metric

for assessing implicit bias in humans. They measure vector similarities between social categories (e.g. gender) and attributes (e.g. careers) in different types of word embeddings. In contrast, probability-based metrics (Kurita et al. 2019) use model-assigned probabilities to estimate biases. For example, masked token metrics ask a masked language model to complete fill-in-the-blank sentences (Kurita et al. 2019; Webster et al. 2020), and pseudo-log-likelihood metrics score the probability of generating a token given other words in the sentence (Nadeem et al. 2020; Salazar et al. 2019). Model-level approaches are advantageous because they offer fine-grained insights into how associations are structurally encoded in a model's representations. However, they are increasingly difficult to apply to larger models where internal embeddings and probability distributions are not accessible, limiting their effectiveness at evaluating modern LLMs. Output-level methods, on the other hand, are not constrained by this lack of architectural transparency because they rely on LLM outputs generated through prompting rather than internal representations, making them more scalable and broadly applicable. This approach typically involves probing models with prompts that are designed to elicit biased or stereotypical outputs, and then analyzing the generated text (Dhamala et al. 2021; Gehman et al. 2020). Common strategies include counterfactual testing, where demographic features in prompts are swapped to reveal unequal treatment (Banerjee et al. 2024; Zhou et al. 2024) and stereotype detection, which probes whether models reproduce or amplify social stereotypes (Babonnaud et al. 2024; Y. Bai et al. 2024). Other approaches focus on sentiment and toxicity analysis (Inan et al. 2023; Koh et al. 2024) and acceptance/rejection rates, which measure disparities in decision-like tasks such as hiring or moral dilemmas (An et al. 2024; Scherrer et al. 2023). Since output-level methods are usually tailored for certain types of biases in specific contexts, they may be better at predicting real-world harm. However, this also makes them inflexible because they do not provide an overall measure of bias, and they cannot be easily extended to other less common forms of bias. Moreover, they tend to capture explicit rather than implicit biases, as they analyze surface-level outputs instead of latent semantic structures. Our method addresses these weaknesses by combining the scalability of output-level methods with the representational focus of intrinsic evaluation of implicit bias using word associations. In what follows, we discuss why word association methods are advantageous for studying implicit biases specifically.

1.1.3 Word Association Methods. Methods of bias evaluation that use word associations are based on the Implicit Association Test (IAT), a behavioral methodology for evaluating implicit biases in humans (Greenwald, McGhee, et al. 1998). Built on the principle that shorter human reaction times correspond to semantic proximity between words, the test quantifies the strength of our unconscious associations between pairs of concepts, for example *black* and *bad* vs. *white* and *bad*. This methodology makes it possible to uncover implicit biases that lie below our level of consciousness, in contrast to explicit biases which are assessed by asking participants direct questions about their opinions and beliefs (Greenwald, Poehlman, et al. 2009). Thus, applying the same logic as the IAT, methods that use word associations to evaluate biases in LLM are advantageous because they specifically target implicit rather than explicit bias, and they closely resemble the way that implicit biases are measured in humans, which is an important consideration given that LLMs behave in increasingly human-like ways (X. Bai et al. 2025). Crucially, the IAT is a top-down approach, treating the mind as a black box and thus relying exclusively on behavior to make assumptions about the content of our conceptual representations. In contrast, model-level embedding-based evaluation metrics, like the WEAT (Caliskan et al. 2017) use word associations, but they use a bottom-up approach, accessing semantic representations directly via the embedding space. As we discussed earlier, this bottom-up approach is problematic for modern LLMs with billions of parameters because the embeddings are not easily accessible. Instead, more recent output-level word association methods are a promising alternative because they use a top-down approach like the IAT, relying on outputs from prompting, which is akin to assessing behavior in humans. Output-level word association methods can take several forms. Some ask LLMs to fill in the blanks of masked sentence templates, for example, the analogy *[X] are to nurses as [Y] are to surgeons* (Y. Zhao, B. Wang, Y. Wang, D. Zhao, Jin, et al. 2024). Others use ambiguous sentences like *The boss called the*

employee because she got lost, and then ask the model *Who got lost?* to assess whether the model can recognize the sentence's ambiguity (Kotek et al. 2023). Another method, called the LLM Word Association Test (X. Bai et al. 2025), provides LLMs with a list of words (e.g. *home, career, art, science*, etc.) and asks the model to choose which words to relate to it given two options (e.g. *Julia* or *Ben*) in order to assess a certain type of bias. It has been shown that these types of output-level measures of intrinsic bias are better at predicting real-world bias compared to model-level measures (X. Bai et al. 2025). A related line of work uses word associations to study cultural rather than social biases. Dai et al. (2025) adapt the word association test to LLMs by asking models to rank candidate associations for a cue word, and compare these rankings with human free associations from the SWOW datasets for American, British, Australian and New Zealand, and Chinese participants. They find that LLM associations align more closely with Western, and in particular American, associations, and they propose a fine-tuning method that steers model representations towards specific cultures. Like our methodology, this approach compares LLM associations with human free association norms. However, it compares the associations of individual cue words directly, without modeling the broader network of associations or the process of semantic priming. Moreover, human associations serve as the reference against which LLMs are evaluated, whereas in our methodology the same bias metric is computed for humans and for LLMs, so that the biases of each can be compared. Word association methods for evaluating biases in LLMs are a promising direction of research, especially considering the parallels to the human way of assessing biases. However, what is still missing from these approaches is a bias metric that can be computed in the same way for humans and LLMs. We discuss why comparisons with humans are important in the next section.

1.1.4 Comparisons with Humans. The importance of comparing LLM biases to human biases is widely recognized (Bas 2024; Geva et al. 2025; Tjuatja et al. 2024). However, despite this recognition, work in this area tends to claim that LLMs demonstrate "human-like biases" without providing any concrete comparative metrics between humans and LLMs. Such quantitative and qualitative metrics are needed if we really want to understand the practical implications of interacting with biased LLMs, that is, understand if LLM biases are stronger, weaker, or categorically different than our own. Caliskan et al. (2017) were among the first to compare human biases to LLM biases, comparing the results of their WEAT methodology to human IAT results. However, X. Bai et al. (2025), who developed the LLM Word Association Test, have cautioned against making direct comparisons between LLM metrics and human IAT metrics given that the way the biases are measured differs significantly and it is not clear how to compare them (X. Bai et al. 2025). Furthermore, they comment that future research connecting LLM biases to human biases could inform the design of bias intervention approaches in both LLMs and humans. Some recent work in this direction has begun to explore more direct comparisons between humans and LLMs. One study investigated differences between humans and LLMs in the level of confidence in their own answers, finding that LLMs are more overconfident than humans when their answers are incorrect (Sun et al. 2025). Additionally, there is a growing body of work that focuses on investigating differences in the cognitive capabilities between humans and LLMs using a top-down approach grounded in psychology. Work in this area has focused on investigating differences between humans and LLMs regarding the structure of the mental lexicon (Abramski, Improta, et al. 2025; Abramski, Lavorati, et al. 2024; Xiao et al. 2025), concreteness effects (Abramski, Lavorati, et al. 2024), and moral values (Xiang et al. 2025). However, this type of psychology-inspired comparative approach has yet to be applied for the specific purpose of evaluating implicit biases. The methodology we propose in this paper fills that gap. It uses a psychology-inspired methodology based on word associations to evaluate biases in such a way that allows direct quantitative and qualitative comparisons between humans and LLMs.

1.1.5 Summary. Up to this point, we have discussed existing methods for bias evaluation as they relate to the defining characteristics of the present methodology. To summarize, our methodology uses an output-level approach for intrinsic bias investigation, avoiding the problem of limited access to model representations and enabling the application to any LLM in various contexts and domains. Moreover, it uses word associations, so

it is especially useful for investigating implicit rather than explicit bias, and it most closely resembles the IAT for assessing implicit biases in humans. And finally, it applies a psychology-inspired approach, allowing for direct quantitative and qualitative comparisons between humans and LLMs, unlike most approaches which offer only vague comparisons. Now that we have outlined the main characteristics of our methodology in relation to previous work, we lay the theoretical groundwork for our methodology in what follows.

1.2 Word Association Networks as Models of Implicit Knowledge

The methodology we propose involves measuring the strength of association between words by modeling semantic priming within LLM-generated word association networks. One of the strengths of our approach is that it is grounded in well-studied theories from cognitive psychology. Among them is the idea that conceptual knowledge is the foundation of language and thought (Fodor 1998; Vigliocco and Vinson 2007). It happens that language models are also built on this idea – language models acquire conceptual knowledge (i.e. semantic representations) from vast amounts of text, and then use that knowledge to generate linguistic output. Thus, models of conceptual knowledge can be used for studying cognitive and linguistic phenomenon in both humans and LLMs. Network models of semantic memory are one of the most widely-studied models of conceptual knowledge and have been applied extensively in cognitive science and AI (Collins and Quillian 1969). In these network models, nodes represent concepts and edges represent relationships between them, so that concepts are defined in terms of how they relate to other concepts. Word association networks are a specific type of network model of semantic memory in which concepts are connected based on associative relations, like *cat – dog* (not to be confused with co-occurrence networks, in which concepts are connected based on how words co-occur in text). Word association networks are often built from free association norms, which are acquired by asking an agent – human or machine – to produce responses associated with a cue word. Since free associations have been shown to correlate with stable implicit attitudes (Schnabel and Asendorpf 2013), word association networks built from free associations can effectively model implicit knowledge. This makes them useful models for studying certain aspects of cognition that originate in implicit knowledge structures, such as implicit biases and stereotypes.

Within such network models of implicit knowledge, cognitive processes can be studied using the spreading activation theory of semantic processing (Collins and Loftus 1975; Collins and Quillian 1969). According to the theory, when we are exposed to a concept, for example, by hearing a word, it leads to the activation of the concept's corresponding node in our semantic memory. The activation then spreads along the connections within our semantic memory network, decaying over time and activating other concepts along the way. The spreading activation theory explains the well-studied cognitive phenomenon of semantic priming, which is central to our methodology due to its connection to semantic relatedness. Semantic priming is our tendency to recognize certain concepts more quickly if we are first "primed" with a related concept (McNamara 2005; Neely 2012). For example, if you are primed with the concept *hospital* and then asked to think of a profession, you are probably more likely to think of the concept *doctor* than if you had been primed with the concept *school*, because *hospital* and *doctor* are more closely related.

The phenomenon of semantic priming has been widely studied using a behavioral experiment called the lexical decision task (LDT) (Hutchison et al. 2013), which measures the time it takes (reaction time) for participants to identify target words after being primed with a related or unrelated prime word, for thousands of prime-target pairs. However, more recently, a computational approach to studying semantic priming has been adopted by implementing a spreading activation process within network models of semantic memory (Abramski, Improta, et al. 2025; Siew 2019). The approach involves activating a prime node within a word association network and then observing the final activation level of the target node at the end of the spreading activation process. This framework effectively models semantic priming by simulating a lexical search process within a network model of semantic memory. Using this computational approach, the final activation levels of target nodes are greater,

on average, when the corresponding prime node is related rather than unrelated to the target (Siew 2019). For example, we would expect the final activation level of the target node *doctor* to be greater after activating the related prime node *hospital* relative to the unrelated prime node *school*. Crucially, this computational approach has been empirically validated by correlating final activation levels from the computational approach with reaction times from the LDT behavioral experiment (Hutchison et al. 2013), for the same prime-target pairs. This correlation was first found by modeling semantic priming within human-generated word association networks (Nelson et al. 2004; Siew 2019), and it was later extended to LLM-generated word association networks (Abramski, Improta, et al. 2025). This grounding in behavioral data is fundamental because it ensures that the computational approach accurately represents the real cognitive phenomenon that it aims to model.

For the purposes of our methodology, we adopt this computational approach that models semantic priming by implementing spreading activation within word association networks, but we adapt it for evaluating implicit biases. Since semantic relatedness is the cognitive mechanism that drives semantic priming, then the final activation levels of target nodes can be considered a proxy for semantic relatedness. Thus, the final activation levels can be considered a measure of the strength of association between prime-target pairs. By choosing prime-target pairs that relate to certain social biases, for example *man* – *doctor* and *woman* – *doctor*, we can quantify implicit biases by observing relative differences in the activation levels of the prime-target pairs, as shown in the example in Figure 1.

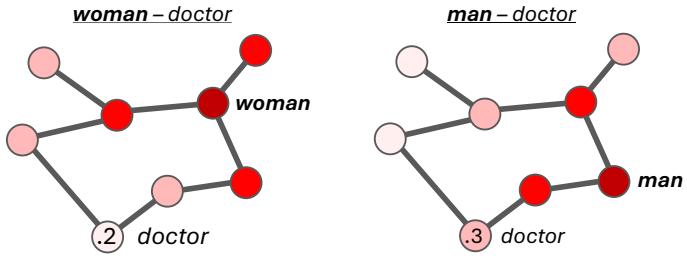


Fig. 1. **Simplified example of gender bias evaluation.** The simplified example shows how our methodology can be used to evaluate gender bias by modeling semantic priming within a word association network. A spreading activation process is implemented in the network by activating the prime nodes *woman* and *man* independently. At the end of the process, the final activation level of *doctor* is 0.2 when *woman* is the prime node, compared to 0.3 when *man* is the prime node. This shows that the association between *man* and *doctor* is stronger than the association between *woman* and *doctor*, indicating gender bias.

This approach follows the same principle as the IAT that we discussed earlier, but since our methodology is computational rather than behavioral in nature, it enables more flexibility, scalability, and reproducibility. It also makes it possible to apply the approach to both LLMs and humans, by using word association networks built from free associations generated by both LLMs and humans. Additionally, since the word association networks used in this methodology are models of broad implicit knowledge including not only the primes and targets of interest, but thousands of other concepts, we can go beyond simple measures of word associations by examining the conceptual paths that are traveled in order to get from primes to targets. Thus, in addition to the quantitative metrics of bias that our methodology provides, it also provides a qualitative understanding of the nature of implicit biases, which is a key feature of our methodology that other approaches lack.

In the remainder of this paper, we demonstrate the utility of our comparative methodology by applying it to three LLMs (Mistral, Llama3, and Haiku) as well as humans to investigate social biases related to gender, religion,

ethnicity, sexual orientation, and political party. Thus, the remainder of this paper is organized as follows. In the Methods section, we describe in detail the three-step process that our methodology entails, including 1) network building, 2) spreading activation, and 3) bias evaluation. In the Results section, we present the results obtained from applying our comparative methodology to the three LLMs and humans for evaluating implicit social biases. Finally, in the Discussion section, we discuss limitations, future work, and the impact that our methodology could have on ensuring the ethical and trustworthy development of LLMs.

2 Methods

In this section, we describe how we apply our methodology to the three LLMs (Mistral, Llama3, and Haiku) and humans to comparatively evaluate their implicit biases related to gender, religion, ethnicity, sexual orientation, and political party. We apply three different versions of our methodology: one approach that evaluates biases in the form of stereotypes, another that evaluates biases in the form of positive or negative perceptions, otherwise known as valence, and a third that evaluates biases in the form of emotions. The stereotypes approach and emotions approach are better suited when the social identity of interest has only two categories, while the valence approach works well when the social identity of interest has two or more categories. Thus, to demonstrate our methodology, we apply the stereotypes approach for evaluating gender biases, the valence approach for evaluating biases related to religion, ethnicity, and sexual orientation, and the emotions approach for evaluating biases related to political party. The three different approaches of our methodology provide a wider view of implicit biases by enabling investigations of different ways in which biases can be expressed: through stereotypes, valence, or emotion. They also make our methodology flexible and adaptable, enabling applications to social identities with a varying number of categories.

Our methodology can be broken down into three steps – network construction, spreading activation, and bias evaluation – as shown in Figure 2. The first step entails building word association networks from free association norms. We use existing datasets of human-generated (De Deyne et al. 2019) and LLM-generated (Abramski, Improta, et al. 2025) free associations that were generated following identical methodologies, making them directly comparable. These network models represent the implicit knowledge of concepts possessed by humans and LLMs, respectively. The second step entails implementing spreading activation processes by activating prime nodes within the networks to model semantic priming. The prime nodes correspond to different social categories of each of the social identities (i.e. gender, religion, ethnicity, sexual orientation, and political party), and they are shown in Table 3. For each network, this step yields final activation levels for each node in the network for each activated prime, stored in an activation level matrix. These values of the matrix represent the strength of association between the activated prime nodes and all other nodes in the network. Finally, the third step entails evaluating biases by conducting statistical analyses of the activation level matrices to produce quantitative bias evaluation metrics. Only at this step does the methodology diverge for the three different approaches (stereotypes, valence, and emotions) because the different approaches require different types of statistical analyses of the matrices. This step also includes a qualitative analysis of bias for the stereotypes approach. By applying this three-step methodology to humans and each of the three LLMs, the resulting bias evaluation metrics enable direct comparisons between each of the LLMs and humans. In what follows, we dive into each step of the three-step methodology.

2.1 Network Construction

The first step of the methodology is network construction, which entails building word association networks from free association norms. In these networks, nodes represent concepts, and edges represent associative relations between them. The networks are built from existing datasets of free association norms generated by humans and LLMs, respectively.

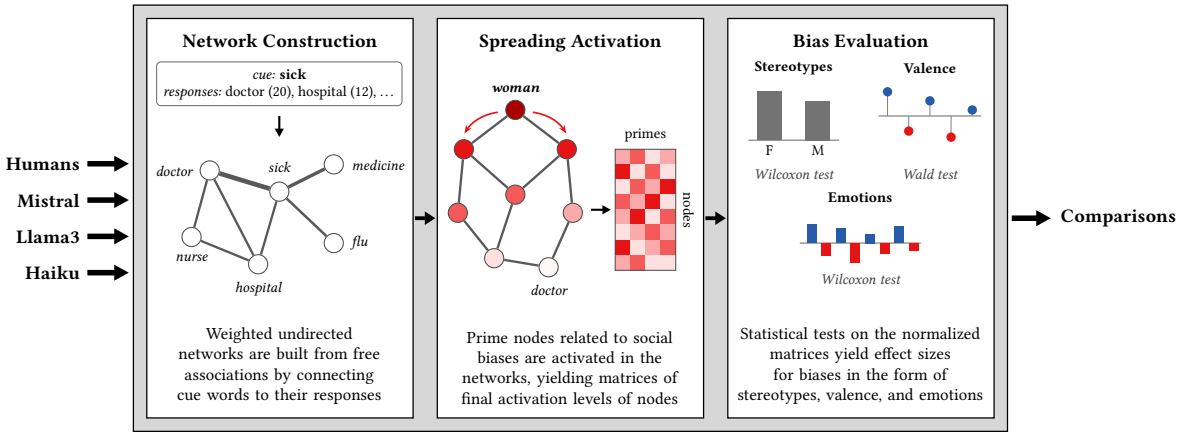


Fig. 2. **Methodological overview**. The flowchart shows an overview of the three-step methodology for evaluating biases in LLMs compared to humans. For humans and each LLM (Mistral, Llama3, and Haiku), first, word association networks are built from free association norms, then spreading activation processes are implemented within the networks, and finally the final activation levels of nodes in the networks are analyzed to evaluate biases. This evaluation methodology enables direct comparisons of social biases between the various LLMs and humans. The words, frequencies, and values shown in the figure are illustrative.

To build the human network, the Small World of Words (SWOW) (De Deyne et al. 2019) dataset of human-generated free association norms is used, while for the LLM networks, the LLM World of Words (LWOW) (Abramski, Improta, et al. 2025) dataset of LLM-generated free association norms is used. The LWOW dataset was derived from the SWOW dataset with the purpose of enabling comparative analyses between humans and LLMs. The SWOW dataset was created through a behavioral experiment in which humans were provided with a cue word and asked to respond with three words that come to mind in response to the cue word, thus tapping into the implicit conceptual knowledge possessed by participants. This experiment was repeated for roughly 12,000 unique cue words, 100 times for each cue word, yielding a dataset of roughly 3 million responses. The LWOW dataset was created by simulating the same behavioral experiment used to create the SWOW dataset, including the same cue words and the same number of repetitions, but instead of using human participants, three different LLMs (Mistral, Llama3, and Haiku) were prompted to provide responses to the cue words, thus tapping into the implicit semantic representations encoded in the LLM architecture. The creation of the LWOW entailed applying a preprocessing procedure to the data in order to facilitate network analyses (lemmatization, spelling corrections, etc.), and in order to align the SWOW dataset with the LWOW, this same preprocessing procedure was applied to the original SWOW. This resulted in the creation of a modified SWOW dataset that is directly comparable to the LWOW dataset. Thus, the free association norms used to create the networks for our methodology include a modified version of the SWOW dataset and the three LWOW datasets. The three LLM datasets can be downloaded from the LWOW github page https://github.com/LLMWorldOfWords/LWOW. The modified SWOW dataset, while not available directly for download because of license restrictions, can be generated by following the procedure described on the LWOW github page. The participants in the SWOW experiment were mostly native English speakers, primarily from the US and the UK (De Deyne et al. 2019). Throughout this paper, we use the term *humans* to refer to this population, and the comparisons between LLMs and humans that we report should be interpreted accordingly.

From those four datasets of free association norms, we build word association networks for humans and all three LLMs by connecting cue words to their responses. Edges are weighted to reflect the frequency of each response to each cue, so if the word *cat* appeared 20 times as the response to the cue *dog*, then the edge from *dog* to *cat* would have a weight of 20. Edges are naturally directed from cues to responses, but in order to facilitate the spreading activation analyses in the next step of the methodology, directed edges are transformed into undirected edges, maintaining the largest of the two edge weights in the case of bidirectional edges. This choice is motivated by the structure of free association data. Many responses never appear as cues, so in a directed network they have incoming edges but no outgoing edges: during spreading activation, these nodes would accumulate the activation they receive without passing it on to other nodes. As a result, the directed network is only weakly connected, meaning that activation cannot travel between every pair of nodes. Transforming directed edges into undirected edges makes the network strongly connected, so that activation can travel between any two nodes of the largest connected component, in both directions. A filtering procedure is then applied to the networks in order to remove unwanted nodes and edges, and to reduce the size of the networks making them more computationally manageable. First, nonsensical and uncommon responses are filtered out by keeping only nodes in WordNet, a large lexical database of English words. Then, idiosyncratic edges (i.e. edges with weight = 1) are removed to ensure that the association is something shared among two or more people (or iterations in the case of the LLMs). Finally, we take the largest connected component of the network, which is essential for the spreading activation processes in the next step of the methodology. The network statistics of the resulting word association networks are shown in Table 2.

Table 2. **Network statistics.** The table shows network statistics for each of the word association networks, including the number of nodes and edges, the network density, and the average degree.

Network	Nodes	Edges	Density	Average degree
Humans	24,308	317,344	0.0011	26
Mistral	20,339	199,103	0.0010	20
Llama3	38,987	546,866	0.0007	28
Haiku	15,596	64,599	0.0005	8

As shown in Table 2, all four networks are sparse, with densities around or below 0.001, which is typical of semantic networks built from free associations (De Deyne et al. 2019; Siew et al. 2019). We do not expect sparsity to bias our statistical tests, for two reasons. First, all analyses are performed on the largest connected component of each network, so every node can be reached from every prime node and can receive activation. Second, all of our bias metrics are based on comparisons made within the same network: the final activation level of a target after activating one prime node is compared with the final activation level of the same target after activating another prime node in the same network. Since both sides of each comparison share the same network topology, sparsity affects both of them in the same way. We never compare raw activation levels across networks; comparisons between humans and LLMs are made only at the level of effect sizes, each of which is computed within a single network. Nonetheless, sparsity could influence the results indirectly. In sparser networks, such as the Haiku network (average degree 8), there are fewer alternative paths between primes and targets, so activation spreads along fewer routes and final activation levels may be more sensitive to individual edges. We return to this point in Section 4.3.

2.2 Spreading Activation

The second step of the methodology involves implementing spreading activation processes within the word association networks by activating prime nodes related to the social identities of interest. As we discussed earlier, this spreading activation process is used to model semantic priming, and it represents what happens in our semantic memory when we are exposed to a concept, such as when we hear a word. Thus, by activating prime nodes that are related to certain social identities, we can measure how strongly certain concepts (i.e. target nodes) are associated with the prime nodes by observing the final activation levels of the target nodes at the end of the spreading activation process. This process is implemented as follows.

First, a set of prime nodes must be defined. The prime nodes should correspond to different categories of the social identity of interest. For example, *european*, *african*, and *asian* are different categories of the social identity ethnicity. When considering only two categories of a social identity, such as is often the case with gender[1], the prime nodes should be one or more pairs of opposing concepts related to the two categories, such as *woman/man*, *female/male*, etc. Rather, when considering more than two categories, the prime nodes should be a set of three or more categories. Crucially, the prime nodes should be of the same linguistic function, so for example, *woman* should not be paired with *sir*, but rather its antonym, *man*, and *buddhism* should not be included in the set containing *christian*, *muslim*, and *jewish*, but rather *buddhist* in order to match the linguistic function of the other prime nodes. This is important in order to limit confounding that could be introduced by differences in linguistic functions. It is also important to note that the prime nodes must be present in the networks in order for the methodology to be applied, so that should also be kept in mind when defining the prime nodes. To demonstrate our methodology, we investigate five different social identities: gender, religion, ethnicity, sexual orientation, and political party. The sets of prime nodes defined for these social identities, as well as the approaches used, are shown in Table 3. Note that for gender the prime nodes are five sets of opposing concepts related to female and male, while for religion, ethnicity, and sexual orientation, the primes nodes are sets of categories, and for political party, the prime nodes include a single pair of opposing concepts.

Table 3. **Prime nodes for different social identities.** The table outlines the three different versions of our methodology that are applied for evaluating biases related to five different social identities. The activated prime nodes are shown as well.

Approach	Social identity	Prime nodes
stereotypes	gender	woman, man female, male girl, boy mother, father feminine, masculine
valence	religion	christian, muslim, buddhist, jewish, atheist
	ethnicity	european, african, hispanic, asian, indigenous
	sexual orientation	straight, gay, lesbian, bisexual
emotions	political party	democrat, republican

Once the prime nodes are defined, they are activated individually and independently within the word association network. A prime node is activated within the network by assigning it a certain activation level. The activation then spreads throughout the network iteratively. At each iteration, some percentage of the activation is retained

[1]For the purpose of demonstrating our methodology, we do not include non-binary gender identities that are outside of the male/female gender dichotomy.

by the activated node, while the rest is passed to its neighboring nodes in a percentage that is proportional to the weights of its connections. This process continues for a fixed number of time steps, and the activation level of each node at the last time step is taken as its final activation level. Computationally, we implemented the spreading activation processes in the word association networks using the R package *spreadr* (Siew 2019), but the more recent python package *SpreadPy* (Citraro et al. 2025) could be used as well. The *spreadr* package requires the specification of the parameters needed for the spreading activation process, including: the initial activation level assigned to the prime node, which we set to the number of nodes in the network; the number of time steps, which we set to twice the diameter of the network; and the percentage of activation retained by the activated node, which we set to the default of 0.5.

These choices are motivated as follows. In our implementation, activation does not decay over time and no suppression threshold is applied (both parameters are set to 0, the default values in *spreadr*). Under these conditions, the total amount of activation in the network is conserved and the process is linear: multiplying the initial activation by a constant multiplies all final activation levels by the same constant. Since the final activation level matrices are normalized before any bias metric is computed (see Section 2.3), the value of the initial activation does not affect our results. We set it to the number of nodes in the network so that, if the activation were spread evenly across the network, each node would receive an activation level of 1. This keeps the final activation levels on a comparable scale across networks of different sizes.

We set the number of time steps to twice the diameter of the network so that the activation starting from a prime node has enough time to reach even the most distant node in the network and to travel back. We note that the number of time steps should not be too large; because activation is conserved and the network is connected, if the process were run for a very large number of time steps, the activation would eventually settle into a distribution that depends only on the weighted degree of each node and not on the nodes' relations to the prime node, so the information capturing the relationships we aim to measure would be lost. Stopping the process after a limited number of time steps ensures that the relative differences we aim to measure can be observed, while the normalization described in Section 2.3 controls for frequency effects, captured by centrality.

Finally, we set the proportion of activation retained at each time step to 0.5, the default value in *spreadr*. Higher retention favors nodes that are close to the prime node, since activation travels more slowly through the network, while lower retention allows more activation to reach distant nodes. A value of 0.5 balances these two effects.

At the end of the spreading activation process for a given prime node, the final activation levels (ALs) of all nodes are stored as a vector that represents the strength of association between the prime node and each node in the network. This process is repeated for each prime node, resulting in a vector for each prime node. The vectors are combined into a final activation level matrix such that the rows are the nodes of the network, the columns are the prime nodes, and the entries of the matrix reflect the strength of association between the prime nodes and each node in the network. At this point, any node in the network can be considered a target node, and the relative strength of association between certain prime-target pairs can be used to evaluate implicit biases. We discuss how the activation level matrices are used to quantitatively and qualitatively evaluate implicit biases in the next section.

2.3 Bias Evaluation

The third step of the methodology involves conducting statistical analyses of the final activation level matrices to quantitatively and qualitatively evaluate implicit biases. The purpose of the statistical analyses is to identify significant patterns in the final ALs that indicate the presence of implicit biases towards certain social categories. The way this is done differs for the three different approaches.

2.3.1 Stereotypes Approach. We apply the stereotypes approach for evaluating biases related to gender. Stereotype bias involves associating certain groups with certain traits or concepts, for example, associating women with

emotion and associating men with strength. For our methodology, that translates to defining prime nodes that relate to certain social identities and target nodes that relate to certain stereotypical traits or concepts. For the prime nodes, we define five matched pairs of female-related and male-related primes (e.g. *woman/man*, *female/male*) as shown in Table 3. For the target nodes, we define 25 stereotypical female-related adjectives and 25 stereotypical male-related adjectives (Gaucher et al. 2011), shown in Table 4. Considering all possible combinations of the 10 prime nodes and the 50 target nodes, we get a total of 500 prime-target pairs. Half of those pairs are stereotype-consistent, for example *feminine – compassionate* and *masculine – forceful*, while the other half are stereotype-inconsistent, for example *feminine – forceful* and *masculine – compassionate*. The aim of the stereotypes approach is to quantify differences in the final ALs of the stereotype-consistent prime-target pairs compared to the stereotype-inconsistent prime-target pairs. Higher final ALs of the stereotype-consistent prime-target pairs compared to the stereotype-inconsistent prime-target pairs is indicative of stereotype bias. The greater the discrepancy, the greater the stereotype bias.

Table 4. **Gender-related stereotype adjectives**. The table shows stereotypical female-related and male-related adjectives that are used as target nodes in the stereotypes approach to evaluating gender biases. The adjectives come from a previous study investigating gendered language in job postings (Gaucher et al. 2011).

Stereotype	**Target nodes**
Female-related	affectionate, cheerful, compassionate, considerate, cooperative, emotional, empathetic, gentle, honest, kind, loyal, modest, nagging, nurturing, pleasant, polite, quiet, sensitive, submissive, supportive, sympathetic, tender, trusting, understanding, warm
Male-related	active, aggressive, ambitious, analytical, assertive, athletic, competitive, confident, courageous, decisive, determined, dominant, forceful, greedy, hostile, impulsive, independent, intellectual, leader, logical, outspoken, persistent, reckless, stubborn, superior

We measure this stereotype bias using a quantitative metric obtained by computing paired differences in the final ALs of targets, and then computing the effect sizes of the differences. First, the final activation level matrix is normalized using the L2 norm by normalizing the columns then rows of the matrix. This normalization accounts for differences in the centrality of nodes within the word association networks, ensuring that the normalized final ALs reflect effects from semantic priming and not effects related to node centrality, like frequency effects. Then, paired differences in normalized final ALs are computed separately for all 25 female-related targets T_F and all 25 male-related targets T_M as follows:

$$\begin{aligned} \text{Diff}_F(t) &= \text{AL}_{p_F}(t) - \text{AL}_{p_M}(t) \quad \forall\, t \in T_F,\ (p_F, p_M) \in P \\ \text{Diff}_M(t) &= \text{AL}_{p_M}(t) - \text{AL}_{p_F}(t) \quad \forall\, t \in T_M,\ (p_F, p_M) \in P \end{aligned}$$

where $AL_p(t)$ denotes the final activation level of target t after being activated by prime p. Differences are computed for all five matched pairs of female-related and male-related primes $(p_F, p_M) \in P$, resulting in two sets of 125 paired differences. Positive differences are indicative of stereotype bias, since they represent greater strength of association between stereotype-consistent prime-target pairs. The Wilcoxon test for paired samples is applied to each set of differences and the effect sizes are computed. The resulting effect sizes, one for female-related targets and one for male-related targets, are quantitative metrics for female-related and male-related stereotype biases, respectively.

This quantitative metric provides a concrete and objective measure that enables direct comparisons between different LLMs and humans. However, we can gain a deeper understanding of the nature of the stereotype biases that we observe through a qualitative analysis of the paths traveled within the networks to get from a prime node to a target node. These prime-target paths reveal certain nuances about the implicit knowledge structures underlying stereotype biases. For example, by observing the paths from the prime node *female* to the target node *compassionate* and from the prime node *male* to the target node *compassionate*, we can gain an understanding of the contexts through which certain implicit associations between concepts are made. Together, the quantitative metric and the qualitative analysis of paths are a powerful tool for implicit stereotype bias evaluation.

2.3.2 Valence Approach. We apply the valence approach for evaluating biases related to religion, ethnicity, and sexual orientation. Contrary to stereotype bias, biases of valence do not involve specific traits or concepts, like the target nodes defined in the previous analysis. Instead, valence has to do with general positive or negative perceptions of certain groups. For example, to measure biases of valence, we are interested in determining whether certain social categories, such as *christian*, are generally associated with more positive concepts, while other social categories such as *muslim* are generally associated with more negative concepts. Since we are not interested in associations related to a certain trait or concept, there is no need to identify specific target nodes. Rather, we are interested in how the final ALs of all the nodes vary with respect to their valence ratings after being activated by different prime nodes. The sets of prime nodes defined for the different categories of religion, ethnicity, and sexual orientation are shown in Table 3. We obtain valence ratings for nodes of the network from the VAD lexicon dataset (Mohammad 2018), in which words are rated by humans from 0 to 1 with 0 corresponding to negative valence and 1 corresponding to positive valence. Thus, using these ratings and the final ALs of all nodes, we can measure how valence differs across different categories of the social identities of interest.

To quantify this valence bias, we fit generalized linear models (GLMs) to measure the relationship between the final ALs of the network nodes and their valence ratings. First, as with the stereotypes approach, the final AL matrix is normalized using the L2 norm by normalizing the columns then rows of the matrix. Then, for each social identity (religion, ethnicity, and sexual orientation) we fit separate univariate GLMs for each social category (prime node), such that the response variable is the vector of valence scores of all nodes, and the predictor is the vector of final ALs of all nodes after being activated by a given prime node. Naturally, the nodes that do not have a valence rating are excluded from the GLM. The resulting coefficient value of each GLM indicates the relationship between the final ALs of the nodes and the valence ratings. Positive coefficients indicate that the corresponding social category (i.e *christian*) is perceived positively, while negative coefficients indicate that it is perceived negatively (relative to the other social categories). Large differences in perceptions across the different social categories (some very positive coefficients and some very negative coefficients) indicate strong valence bias. The greater the variation in coefficients, the greater the bias. For each social identity, this bias is quantified by fitting a single multivariate GLM with all social categories (prime nodes) and then computing the effect size by conducting a Wald test, with the null hypothesis that all coefficients are equal. Thus, a large effect size tells us that the coefficients differ significantly from each other, and so the different social categories are perceived differently in terms of their valence. The resulting effect size is thus a measure of the bias of valence regarding the social identity of interest, enabling direct comparisons between different LLMs and humans.

2.3.3 Emotions Approach. We apply the emotions approach for evaluating biases related to political party, specifically, the two US political parties, Democrats and Republicans. This approach combines elements from both the stereotypes approach and the valence approach. The general idea is the same as the valence approach, but instead of evaluating positive and negative perceptions of certain groups, it evaluates feelings towards certain groups considering eight core emotions, defined by Plutchik (Plutchik 1980): anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The computation of the quantitative metric is more similar to the stereotypes approach since it is applied to a social identity with only two categories. In this case, we have a single pair of

prime nodes – *democrat* and *republican* – so the aim of this approach is to measure differences in emotions towards these two different categories. In order to do this, we identify sets of emotion-related target nodes, which are words in the networks that elicit certain emotions, according to human judgments from the NRC emotion lexicon (Mohammad 2018). Thus, for each of the eight emotions, a set of nearly 1,000 emotion-related targets are defined.

To quantify differences in emotions towards Democrats and Republicans, we compute differences in the final ALs of the emotion-related targets after being activated by *democrat* compared to *republican*. First, the final activation level matrix is normalized first by columns and then by rows. For the emotions approach, we report the results obtained with the L1 norm in the main text and those obtained with the L2 norm in Appendix A. With the L2 norm, the effect sizes have the same sign for all eight emotions within each network, which would mean that opposite emotions, such as joy and sadness, are both felt more strongly towards the same political party. Since opposite emotions should not have the same polarity, we interpret this as an indication that the L2 normalization does not fully remove differences between the two primes that are unrelated to emotions. We discuss this choice further in Section 4.3. Then, for each emotion e in the set of eight emotions, paired differences in normalized final ALs are computed for all emotion-related targets T_e as follows:

$$\text{Diff}_e(t) = \text{AL}_{democrat}(t) - \text{AL}_{republican}(t) \quad \forall\, t \in T_e$$

where $AL_p(t)$ denotes the final activation level of target t after being activated by prime p. This results in eight sets of nearly 1,000 differences, one set for each emotion, such that positive differences indicate stronger emotions towards Democrats while negative differences indicated stronger emotions towards Republicans. The Wilcoxon test for paired samples is applied to each set of differences and the effect sizes are computed. The effect size, which can be either positive or negative, is therefore a measure of how strongly an emotion is felt for either Democrats (positive effect) or Republicans (negative effect).

3 Results

We applied the methodology described above to humans and all three LLMs (Mistral, Llama3, and Haiku). First, we applied the stereotypes approach to evaluate stereotype biases related to gender. We then applied the valence approach to evaluate biases of valence related to religion, ethnicity, and sexual orientation. Finally, we applied the emotions approach to evaluate emotional biases related to political party. In what follows, we present the results obtained with each approach, which allow for direct comparisons between humans and LLMs.

3.1 Stereotypes Related to Gender

We applied the stereotypes approach for evaluating gender bias in humans, Mistral, Llama3, and Haiku. We implemented spreading activation processes in word association networks generated by humans and the three LLMs by activating prime nodes related to gender in each of the four networks. We then observed the final ALs of female-related and male-related stereotypical adjectives. We quantified stereotype biases by computing differences in final ALs of stereotypical and non-stereotypical prime-target pairs. The Wilcoxon test for paired samples was applied to the differences and effect sizes were computed separately to quantify female-related stereotypes and male-related stereotypes. The effect sizes are shown in the barplots in Figure 3.

Positive significant effects indicating the presence of stereotype biases are observed in all four of the networks. Effect sizes for female stereotypes and male stereotypes are nearly identical for humans, but they vary in the LLMs. Female stereotypes are stronger than male stereotypes in Mistral and Haiku, but male stereotypes are stronger than female stereotypes in Llama3. Overall, humans demonstrate the strongest gender stereotypes while Mistral demonstrates the weakest gender stereotypes.

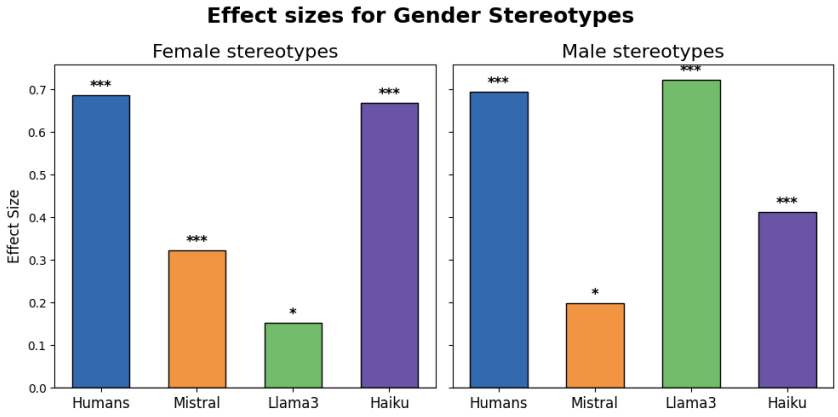


Fig. 3. **Effect sizes for gender stereotypes.** The barplots show the effect sizes for female-related stereotypes (left) and male-related stereotypes for humans and the three LLMs. Statistical significance is indicated with stars.

A fine-grained view of the results are shown in the heatmaps in Figures 4, 5, 6, and 7, which show normalized final ALs of all prime-target pairs for each of the four networks, with female-related targets on the left and male-related targets on the right. The heatmaps make it easy to visually identify the presence of stereotype biases by observing a vertical divide in the lightness/darkness between the left and right sides of the heatmaps, corresponding to the female-related primes and the male-related primes. This is especially noticeable in the Llama3 heatmaps in Figure 6, indicating that male stereotypes (right) are stronger than female stereotypes (left), which is confirmed by the effect sizes. The heatmaps also make it possible to observe the strength of association between specific pairs of concepts, enabling a deeper understanding of what drives the stereotypes that are observed. For example, in all networks, the concept *masculine* is strongly associated with many of the male-related targets, suggesting that the use of this term may have a stronger stereotypical connotation with respect to other male-related primes, such as *father*, which appears to be weakly associated with many of the male-related targets. Also, some specific prime-target associations are strong across all networks, like *mother* – *nurturing*, but other associations differ greatly across the networks. Haiku, for instance, associates *courageous* much more with *female* than with *male*, a stereotype-inconsistent pattern that stands out compared to the other networks.

A deeper qualitative understanding of stereotype biases can be gained by observing paths between specific prime-target pairs, as shown in Figures 8, 9, 10, and 11. We call these paths "mindset streams" because they show the stream of thought that is traveled in implicit semantic memory to get from one concept to another. The figures show the mindset streams of humans, Mistral, Llama3, and Haiku for four prime-target pairs: *feminine* – *compassionate*, *feminine* – *forceful*, *masculine* – *forceful*, and *masculine* – *compassionate*. Each mindset stream includes all the shortest paths between the prime and the target, and the words are colored according to human valence ratings, so that positive words are blue and negative words are red (not to be confused with the results from the valence approach that we discuss later).

In all four networks, the paths between stereotype-consistent pairs tend to be shorter than, or as short as, those between stereotype-inconsistent pairs, and they pass through fewer intermediate concepts. The networks differ, however, in the richness of their mindset streams. The human mindset streams contain dozens of intermediate concepts along the stereotype-inconsistent paths, while those of the LLMs contain only a few, reflecting the simpler conceptual structures of the LLM networks. Differences in path length across networks should be interpreted with caution, since they also depend on the structure of each network, such as its density.

The intermediate concepts provide important insight about the contexts through which primes and targets are associated. In humans, the stereotype-inconsistent paths from *feminine* to *forceful* and from *masculine* to *compassionate* pass through concepts related to sex (e.g., *sex*, *sexy*, *seductive*), while such concepts are absent from the stereotype-consistent paths and from the mindset streams of all three LLMs. The valence of the intermediate concepts also differs across networks. In humans, there are more negative than positive concepts along the path from *feminine* to *forceful* (e.g., *bully*, *overbearing*, *arrogant*), while there are more positive than negative concepts along the path from *masculine* to *compassionate* (e.g., *love*, *generous*, *friend*). This suggests that going against the female stereotype may be viewed more negatively than going against the male stereotype, which has important implications for sexism and the struggle for gender equality. Haiku shows a similar asymmetry, with negative concepts such as *sissy* and *aggressive* along the path from *feminine* to *forceful*, and only neutral or positive concepts along the path from *masculine* to *compassionate*. In Mistral and Llama3, instead, the path from *feminine* to *forceful* contains no negative concepts and passes through neutral and positive concepts such as *strength*, *powerful*, and *confident*. These observations are qualitative and based on four prime-target pairs; a systematic analysis of path properties across all 500 prime-target pairs is left for future work. Together, the effect sizes, the heatmaps, and the mindset streams provide a comprehensive and varied understanding of stereotype biases in humans and LLMs, allowing for many different types of comparisons and insights.

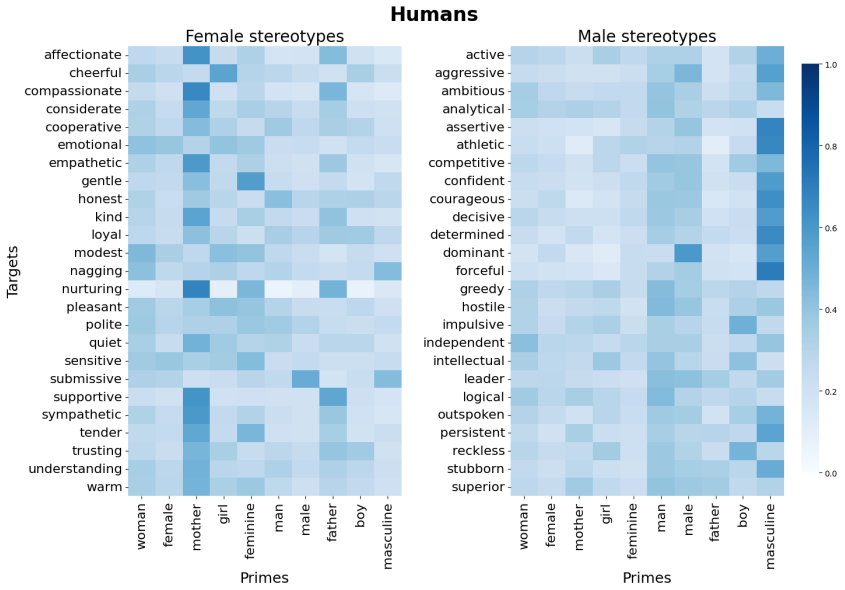


Fig. 4. **Gender biases in humans.** The heatmaps show the normalized final activation levels of 25 female-related targets (left) and 25 male-related targets (right) after activating 10 gender-related prime nodes in the human network.

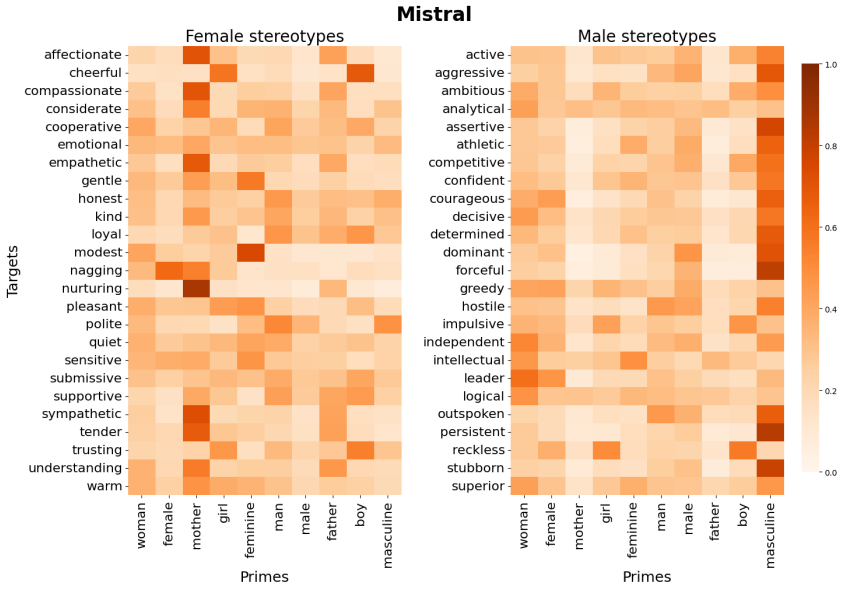


Fig. 5. **Gender biases in Mistral.** The heatmaps show the normalized final activation levels of 25 female-related targets (left) and 25 male-related targets (right) after activating 10 gender-related prime nodes in the Mistral network.

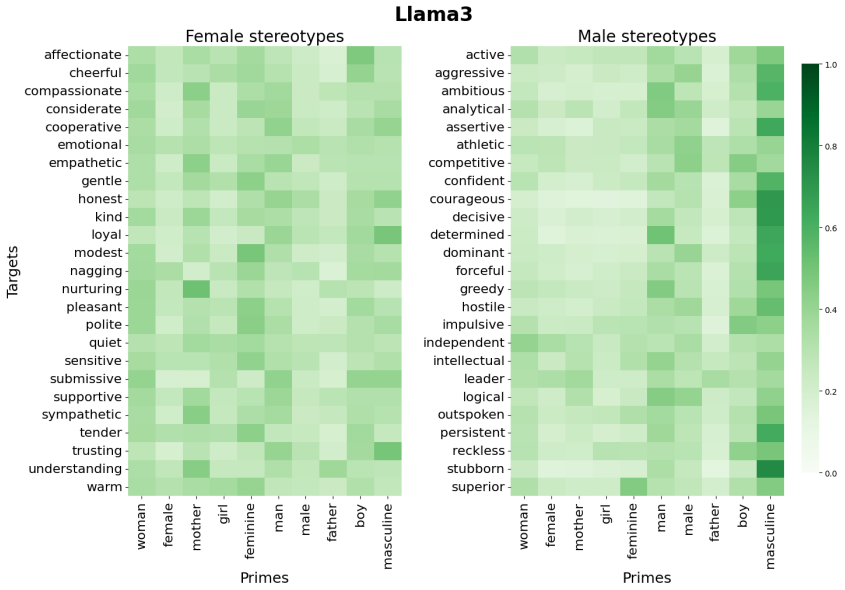


Fig. 6. **Gender biases in Llama3.** The heatmaps show the normalized final activation levels of 25 female-related targets (left) and 25 male-related targets (right) after activating 10 gender-related prime nodes in the Llama3 network.

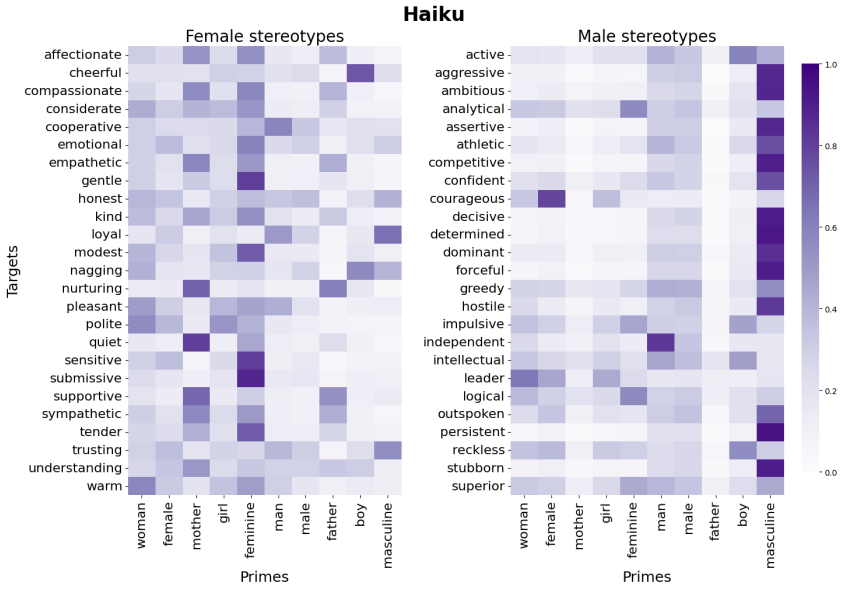


Fig. 7. **Gender biases in Haiku.** The heatmaps show the normalized final activation levels of 25 female-related targets (left) and 25 male-related targets (right) after activating 10 gender-related prime nodes in the Haiku network.

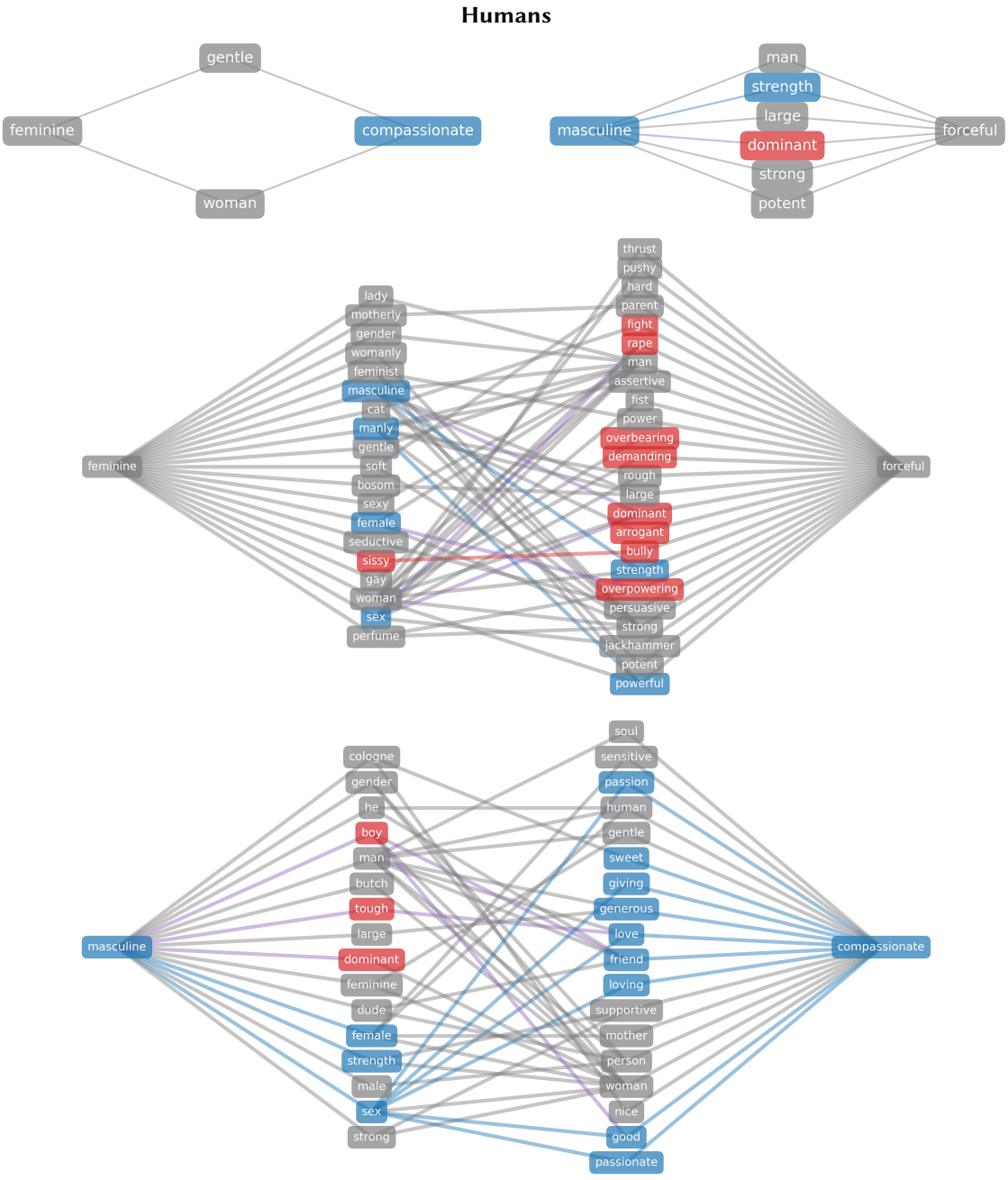


Fig. 8. **Human mindset streams.** The mindset streams show the paths from primes to targets in the human network for four different prime-target pairs. The colors of the words reflect human valence ratings of the words. Paths are typically shorter for stereotype-consistent pairs (*feminine – compassionate*, *masculine – forceful*) compared to stereotype-inconsistent pairs (*feminine – forceful*, *masculine – compassionate*).

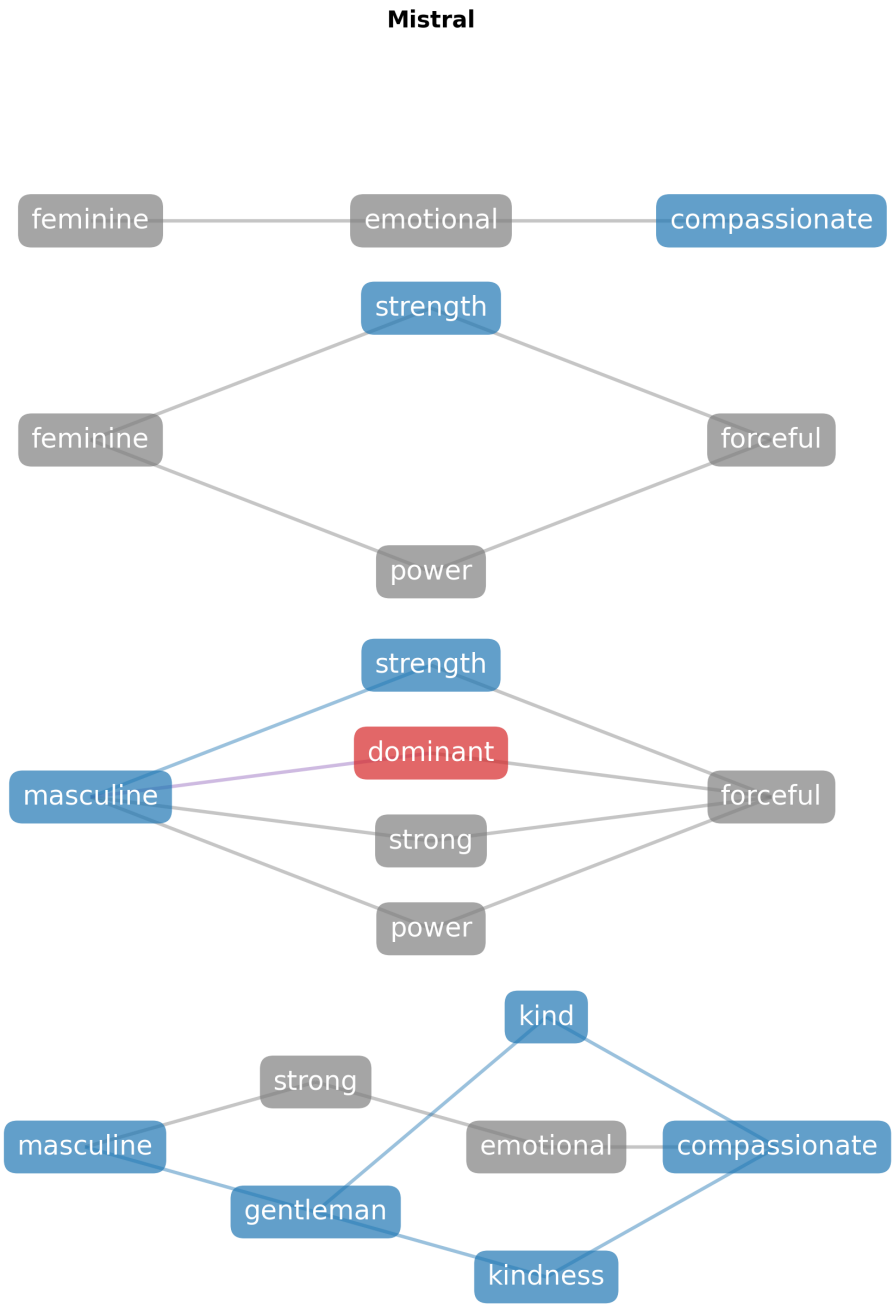


Fig. 9. **Mistral mindset streams.** The mindset streams show the paths from primes to targets in the Mistral network for four different prime-target pairs. The colors of the words reflect human valence ratings of the words.

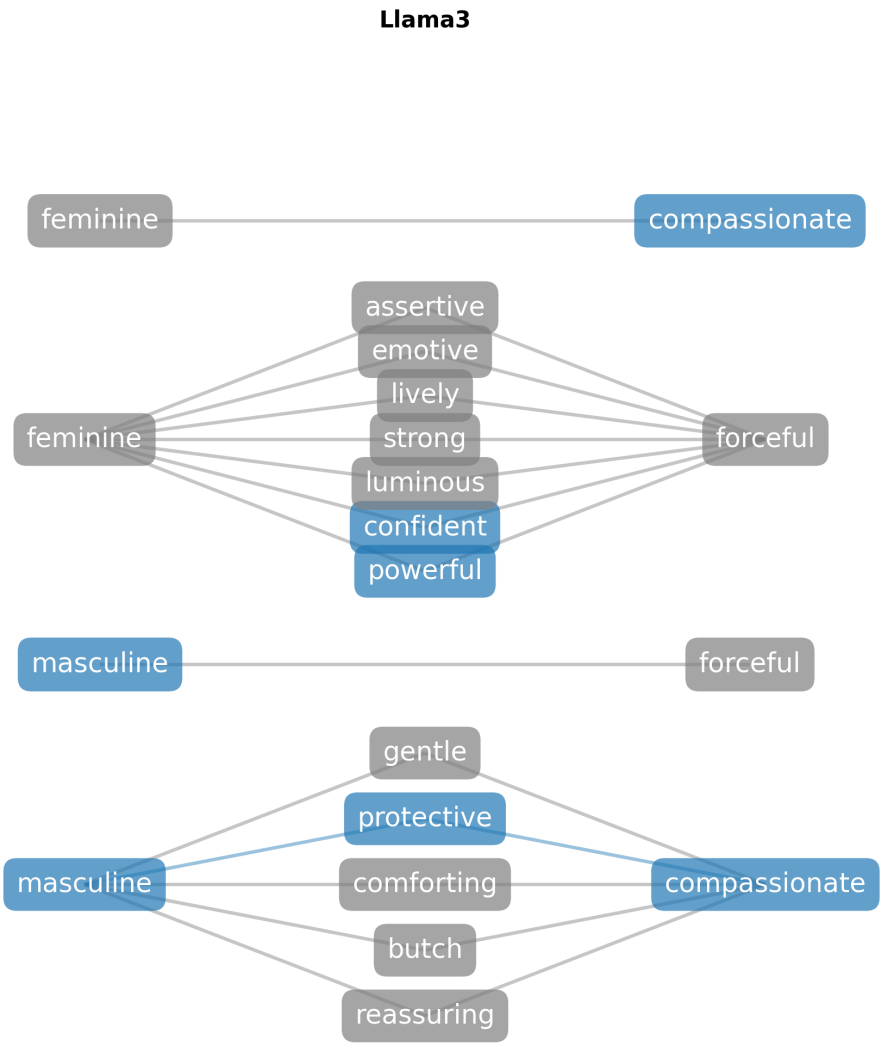


Fig. 10. **Llama3 mindset streams.** The mindset streams show the paths from primes to targets in the Llama3 network for four different prime-target pairs. The colors of the words reflect human valence ratings of the words.

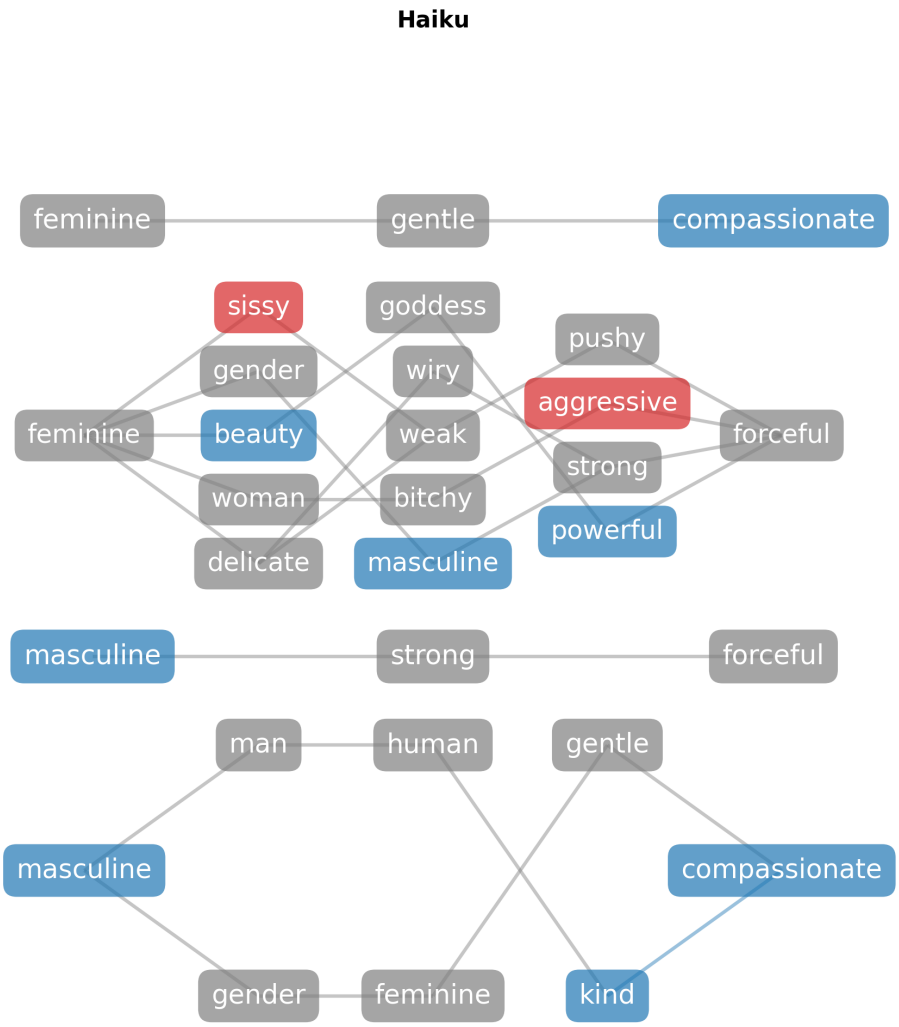


Fig. 11. **Haiku mindset streams.** The mindset streams show the paths from primes to targets in the Haiku network for four different prime-target pairs. The colors of the words reflect human valence ratings of the words. Paths are typically shorter for stereotype-consistent pairs (*feminine – compassionate*, *masculine – forceful*) compared to stereotype-inconsistent pairs (*feminine – forceful*, *masculine – compassionate*).

3.2 Valence Related to Religion, Ethnicity, and Sexual Orientation

We applied the valence approach for evaluating biases related to religion, ethnicity, and sexual orientation in humans, Mistral, Llama3, and Haiku. We implemented spreading activation processes in word association networks generated by humans and the three LLMs by activating prime nodes related to the three social identities in each of the four networks. Then, for each of the three social identities, we fitted separate univariate GLMs for each category (prime node) on the vectors of final ALs of all nodes such that the final ALs were the predictors and the valence scores of all the nodes were the responses. The coefficient values of each GLM indicate the extent

to which the category (prime node) of the social identity is viewed positively or negatively with respect to the other categories. Effect sizes were computed by fitting a single multivariate GLM including the final AL vectors of all the categories as predictors and the valence scores as the response. Significant effect sizes indicate that there is large variation in the perception of the different categories (i.e. the coefficients differ significantly from each other). The coefficients of the separate univariate GLMs are shown for religion, ethnicity, and sexual orientation in Figure 12 while the effect sizes computed from the single multivariate GLM are shown in Figure 13.

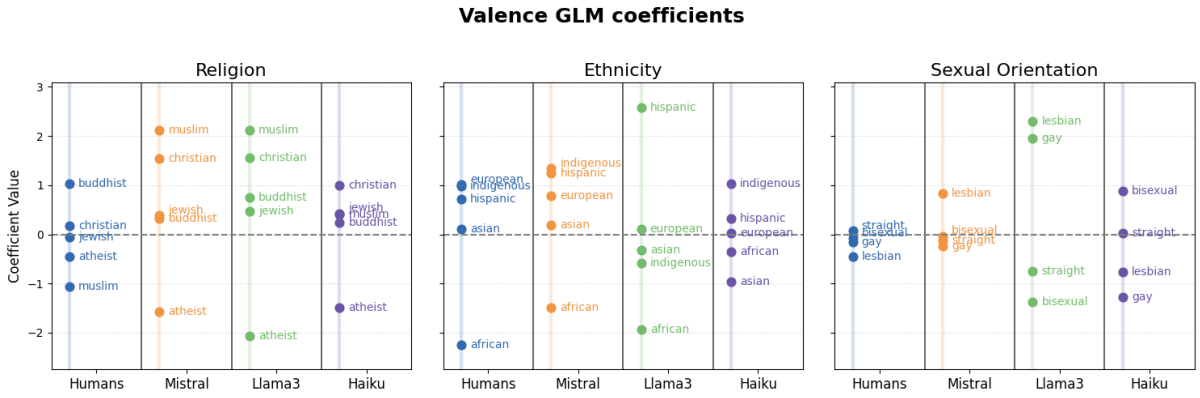


Fig. 12. **Coefficients reflecting positive and negative perceptions.** The plots show coefficient values of the separate univariate GLMs fit for each category of religion (left), ethnicity (center), and sexual orientation (right) for each of the four networks. Positive coefficients indicate that the category is perceived more positively, while negative coefficients indicate that the category is perceived more negatively, with respect to the other categories.

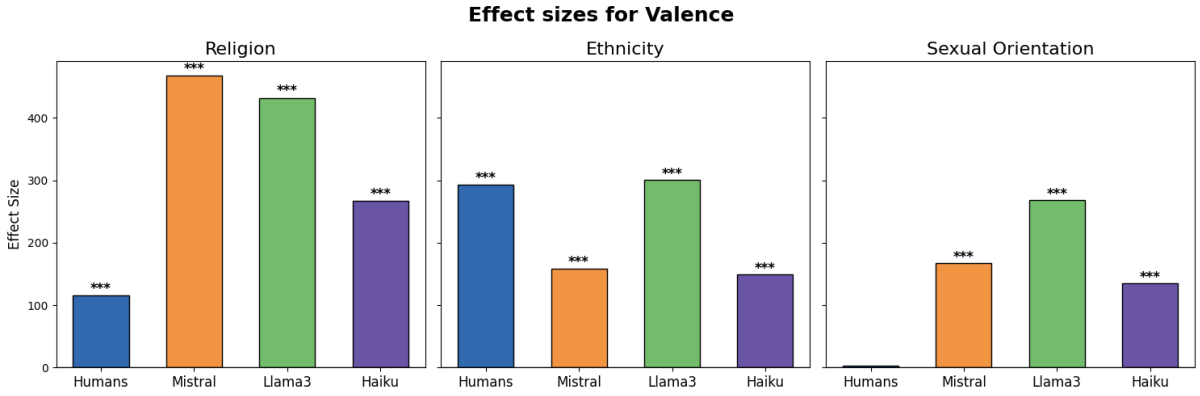


Fig. 13. **Effect sizes for valence GLMs.** The barplots show the effect sizes for each network for each religion (left), ethnicity (center), and sexual orientation (right), computed using the Wald test for a single multivariate GLM. Statistical significance is indicated with stars. A large effect size indicates that the coefficients of the GLM differ significantly from each other, indicating that there is a significant difference in how the different categories are perceived in terms of valence.

For religion, and ethnicity, there is significant valence bias in humans and all three LLMs. Instead, for sexual orientation, humans demonstrate almost no valence bias but all three LLMs demonstrate bias. Contrary to the

gender stereotype biases, humans are less biased than the LLMs regarding religion, but regarding ethnicity, humans are just as biased if not more biased than the LLMs. For religion, *buddhist* and *christian* are seen more positively by humans while *atheist* and *muslim* are seen more negatively, and *jewish* is seen as neutral. Interestingly, in all three LLMs, *muslim* is seen more positively, and *buddhist* is seen less positively among LLMs, contrary to humans. Regarding ethnicity, humans view *african* much more negatively than the other ethnicities, and they view *european* more positively, reflecting the bias of white supremacy. In LLMs, *european* is not the most positively viewed ethnicity, but *african* is more negatively viewed in all LLMs. Regarding sexual orientation, while humans demonstrate little to no bias, Llama3 views lesbians and gays more positively but bisexuals more negatively. The valence approach is mainly useful for observing how positive and negative perceptions differ across certain categories of a certain social identity, and it allows for comparisons between humans and LLMs. These comparisons are insightful, especially when the LLMs demonstrate patterns of perception that are contradictory to those of humans, because it suggests that maybe the LLMs were fine-tuned to correct certain biases, making them biased in the opposite direction.

3.3 Emotions Related to Political Party

We applied the emotions approach for evaluating biases related to political party in humans, Mistral, Llama3, and Haiku. We implemented spreading activation processes in word association networks generated by humans and the three LLMs by activating the prime nodes *democrat* and *republican* in each of the four networks, and then we observed the final ALs of words that elicit certain emotions. We quantified relative differences in emotions felt towards Democrats and Republicans by computing differences in the final ALs of emotion-related target nodes for eight different emotions. The Wilcoxon test for paired samples was applied to the differences and effect sizes were computed for each emotion. The effect sizes are shown in the barplots in Figure 14.

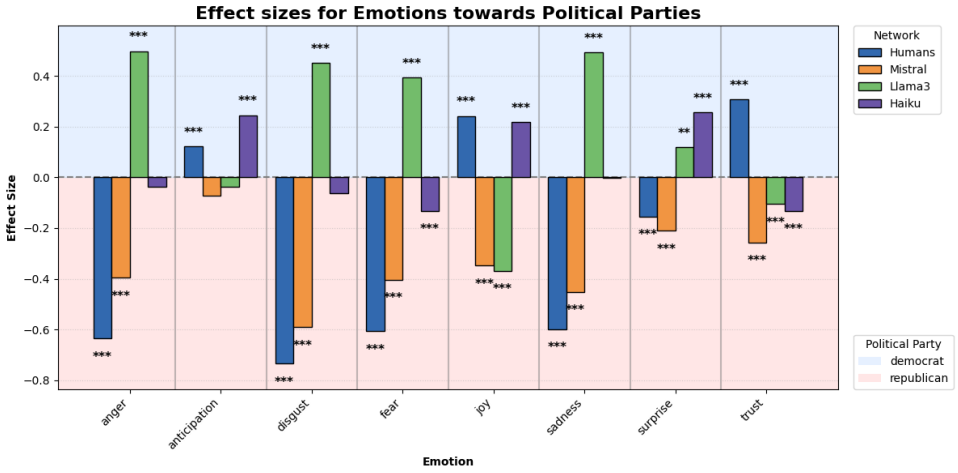


Fig. 14. **Effect sizes for emotions towards political parties.** The barplots show the effect sizes for each network for all eight core emotions. Statistical significance is indicated with stars. For each emotion, positive effects indicate stronger feelings of that emotion towards Democrats, while negative effects indicate stronger feelings towards Republicans. Note, the final activation level matrices were normalized using the L1 norm (see Methods).

Positive effect sizes indicate that a given emotion is felt more strongly towards Democrats, while negative effect sizes indicate that a given emotion is felt more strongly towards Republicans. Perhaps the first thing we

notice is that the negative emotions of anger, disgust, fear, and sadness are felt more for Democrats in Llama3 but more for Republicans in humans and Mistral. Haiku demonstrates a more neutral stance for these four negative emotions, with only significant negative effects for fear (more strongly felt towards Republicans). Instead, for the positive emotions of anticipation, joy, surprise, and trust, we observe varying patterns. Notably, effect sizes for humans and Llama3 are of opposite polarity for all emotions. It is also interesting to note that only humans feel more trust towards Democrats, with all three LLMs having more trust in Republicans. It is also interesting to note that Mistral demonstrates stronger emotions for Republicans across all eight emotions, though not all of these effects are significant. Of all the networks, Haiku appears to be the most politically neutral in terms of emotions. On the contrary, humans appear to be the most radical, demonstrating especially strong emotional differences for negative emotions towards Republicans. The implications of these results are further elaborated in the Discussion.

4 Discussion

In this work, we introduce a novel word association network methodology for evaluating implicit biases in LLMs compared to humans. We demonstrate the utility of our methodology by applying it to studying social biases in humans and three LLMs. We apply three different versions of our methodology to investigate different types of biases: a stereotypes approach to investigate stereotypical gender biases, a valence approach to investigate positive and negative perceptions of religion, ethnicity, and sexual orientation, and finally, an emotions approach to investigate the relative feelings towards Democrats and Republicans for eight core emotions. In what follows we discuss the implications of the results of our empirical analyses, which indicate both convergence and divergence in biases between humans and LLMs, and we also discuss the advantages and societal implications of our methodology, as well as limitations and future work.

4.1 Implications of Empirical Analyses

Convergence between human and LLM biases was most evident regarding gender stereotypes, in which stronger associations were observed between stereotypical prime-target pairs like *feminine* – *compassionate* and *masculine* – *forceful* for both humans and LLMs, consistent with widely documented linguistic and cultural stereotypes. Humans displayed the strongest overall gender bias, with equally strong female-related and male-related stereotypes, while the LLMs varied in both direction and intensity. Specifically, Mistral and Haiku showed stronger female-related stereotypes, whereas Llama3 exhibited the opposite pattern, demonstrating stronger male-related stereotypes. These results support previous work showing that LLMs can inherit human-like stereotypes from training data (X. Bai et al. 2025; Caliskan et al. 2017; Stella et al. 2023), but also suggest that different models may internalize and express these associations differently depending on their training and fine-tuning procedures. The qualitative mindset streams further reveal that stereotype-consistent associations are not only stronger but also cognitively more accessible, as evidenced by shorter paths between primes and targets. This observation resonates with cognitive psychology findings that implicit associations are often more automatic and contextually reinforced (Greenwald, McGhee, et al. 1998; Wilson et al. 2000). The human mindset streams were also enriched with more nuanced associative contexts and valence variations, consistent with richer conceptual networks and more complex socially embedded knowledge structures. In contrast, the pathways of the three LLMs reflect a more simplistic conceptual architecture, suggesting a shallower representation of the complex, culturally mediated nuances that characterize gendered cognition in humans. Notably, the negative concepts that mediate the association between *feminine* and *forceful* in humans also appear in Haiku, but not in Mistral and Llama3, and the context of *sex*, which mediates stereotype-inconsistent associations in humans, is absent in all three LLMs. This suggests that LLMs differ in whether their stereotype-inconsistent associations carry the evaluative and affective components found in humans.

The valence analysis further showed systematic differences in how social categories of varying religion, ethnicity, or sexual orientation are perceived. The finding that *christian* elicited more positive associations than *muslim* is consistent with IAT research showing pro-Christian implicit preferences in Western samples and broader implicit hierarchies across religions (Heiphetz et al. 2013). Interestingly, all three LLMs also perceived *christian* positively, but Mistral and Llama3 perceived *muslim* even more positively than *christian*, in contrast to humans. Alignment between LLMs and humans may suggest that LLMs not only replicate but may amplify valence asymmetries in ways that risk reinforcing social inequalities, as shown in previous work on math anxiety in LLMs (Abramski, Citraro, et al. 2023). However, misalignment between LLMs and humans may reflect efforts to reduce or counterbalance biases in their fine-tuning data (Shen et al. 2023). Ethnicity-related biases were also consistent with prior sociolinguistic findings – humans demonstrated clear patterns of ingroup favoritism, with *european* perceived most positively and *african* most negatively, reflecting structural inequalities in Western corpora and cognition (Acerbi and Stubbersfield 2023). The LLMs mirrored this pattern to varying degrees, with *african* consistently associated with lower valence, suggesting that despite extensive filtering and alignment, models continue to internalize biases present in large-scale English-language training data (Bender et al. 2021; Sorokovikova et al. 2025). Results regarding sexual orientation were particularly revealing – humans showed negligible bias, while all three LLMs exhibited significant differences in perception, often favoring gay and lesbian identities. These findings may reflect a growing over-correction effect in alignment pipelines, where biases are mitigated in one direction but inadvertently amplified in another (Sun et al. 2025). Collectively, these results highlight that while LLMs often reproduce human-like biases, they can also develop "counter-biases" that diverge systematically from human cognitive tendencies. Moreover, evidence that LLMs reproduce culturally prevalent content biases suggests that unequal valence may reflect corpus-level regularities rather than purely model-internal artifacts (Acerbi and Stubbersfield 2023). Future work should test whether targeted corpus curation or debiasing alters these valence profiles and whether network-level edits can attenuate amplification effects without degrading linguistic competence.

Regarding emotions in the political domain, significant divergences between humans and LLMs emerged. Humans exhibited strong emotional polarization, with negative emotions such as anger, disgust, fear, and sadness directed primarily toward Republicans, and positive emotions such as trust and joy toward Democrats. These patterns align with evidence of affective polarization in U.S. populations (Stanley et al. 2011; Sun et al. 2025). In contrast, the LLMs demonstrated less emotional differentiation overall, with Mistral and Llama3 showing mild polarity and Haiku appearing largely neutral. Interestingly, all three models expressed more trust toward Republicans, opposite to the human pattern. These discrepancies highlight that while structural features of bias resemble human cognition, the content of associations may diverge in politically sensitive domains. Such divergences may reflect differences in cultural representation within training corpora (Acerbi and Stubbersfield 2023; Stanley et al. 2011; Stella et al. 2023) or limitations in how LLMs express emotional content (Sun et al. 2025). Recent audits report systematic political leanings in several LLMs, which could partially explain model–human mismatches in affective attributions (Motoki et al. 2024; Rozado 2024). Moreover, political preferences in LLMs have been shown to shift with prompts, personas, and scale, underscoring the sensitivity of affective associations to elicitation conditions (Bang et al. 2024; Motoki et al. 2024). Independent evidence also indicates that current systems exhibit uneven emotional intelligence and gaps in emotion reasoning, consistent with the affective divergences observed here (Abramski, Citraro, et al. 2023; Sun et al. 2025). Together, these findings support the need for interpretable, network-based diagnostics that can localize which subgraphs or pathways (e.g. mindset streams (Brian and Stella 2023)) encode partisan-affective couplings rather than relying solely on aggregate bias scores.

4.2 Methodological Advantages and Societal Implications

A central methodological innovation of the present work is the use of word association networks (Abramski, Improta, et al. 2025) to directly compare implicit biases in humans and LLMs. Unlike model-level embedding-based metrics, which often operate as opaque vector-space measures (Caliskan et al. 2017), our network-based methodology does not require access to internal representations, rather, it offers interpretable structures in which nodes represent concepts and edges capture associative strength (Brian and Stella 2023; Semeraro et al. 2025; Siew et al. 2019). Moreover, unlike most output-level metrics that use prompts to evaluate bias in specific contexts, our methodology captures implicit relational knowledge rather than surface-level outputs. One of the main advantages of our methodology is that it is grounded in cognitive psychology and network science. Semantic networks of memory recall patterns have been widely applied to model conceptual memory (De Deyne et al. 2019), spreading activation (Citraro et al. 2025), and semantic priming (Siew 2019) in humans. By operationalizing implicit bias as asymmetries in activation flow across these networks, our approach mirrors established theories of semantic priming (Collins and Loftus 1975; Siew 2019) while remaining transparent and human-readable. The networks constructed in this study capture large-scale associative structures. For example, the human network consists of 24,308 nodes and 317,344 edges, while the Llama3 network is even larger with 38,987 nodes and 546,866 edges, reflecting differences in the structure of the associative knowledge of each model. This transparency allows biases to be traced through activation pathways, (i.e. mindset streams) (Brian and Stella 2023), such as the observation that stereotype-consistent links (e.g. *masculine – forceful*) are mediated by shorter semantic paths than inconsistent ones. The interpretability afforded by network-based approaches therefore provides a clear advantage, as it enables both quantitative evaluation and qualitative inspection of how associations emerge and propagate, supporting more nuanced accountability in bias evaluation (Abramski, Citraro, et al. 2023).

To summarize, our word association network methodology offers three key contributions. First, it provides a flexible, scalable framework for implicit bias evaluation that avoids reliance on internal model representations, bridging the gap between model-level (intrinsic) and output-level (extrinsic) evaluations. Second, our approach is based on the IAT for assessing implicit biases in humans, and it is grounded in theories from cognitive science. And third, it enables direct quantitative and qualitative comparisons between humans and LLMs, a feature often missing from prior work in which claims about "human-like" LLM biases remain vague and ungrounded. This last point addresses the problematic practice of comparing LLM metrics to human metrics that are not obtained in the same way, such as comparing embedding-based results to IAT results, which can lead to unreliable comparisons (X. Bai et al. 2025). To the best of our knowledge, our methodology is the first of its kind that offers both quantitative and qualitative assessments of implicit bias enabling direct comparisons between humans and LLMs.

The societal implications of adopting a word association network approach to LLM bias evaluation are many. Because word association networks provide interpretable pathways of associations (Semeraro et al. 2025), they allow stakeholders—including policymakers, educators, and developers—to identify not only the presence of bias but also the mechanisms through which it propagates (Abramski, Citraro, et al. 2023). This level of transparency is crucial in high-stakes domains such as healthcare, hiring, and criminal justice, where biased outputs from LLMs can perpetuate discrimination (Motoki et al. 2024) or reinforce social stereotypes (Caliskan et al. 2017; Sun et al. 2025). By linking model associations to human cognitive structures, this method enables a clearer assessment of whether LLMs amplify existing human biases or introduce qualitatively different patterns, which has direct implications for trust and accountability in AI systems. Furthermore, the network framework supports cross-cultural and longitudinal monitoring of bias, making it possible to track how implicit associations evolve as models are updated or deployed in new contexts (Abramski, Improta, et al. 2025; Abramski, Lavorati, et al. 2024). In doing so, it provides a scalable tool for guiding socially responsible AI development, fostering greater alignment between technological innovation and ethical standards in society.

4.3 Limitations and Future Work

Despite the many advantages of our word association network methodology, it also has several limitations that should be taken into careful consideration and that could be further addressed in future work. One important limitation has to do with the normalization of the final activation level matrices, which is necessary in order to account for differences in the centrality of nodes within the word association networks. This normalization requires choosing the appropriate norm, but the choice can potentially influence the results produced by the methodology. In this work, we focused on two highly common norming choices – the L1 norm and the L2 norm. For the stereotypes approach and the valence approach, the L1 norm and the L2 norm produced results that converged towards the same patterns, however, for the emotions approach, we found that the selection of the L1 norm (see results in Figure 14) led to different results compared to the L2 norm (see results in Figure 15 in Appendix A). This variability in results suggests that the methodology entails a level of complexity that requires an accurate selection of the norm by the researcher. In the case of the analysis of feelings towards Democrats and Republicans, selecting the L2 norm resulted in effect sizes of the same polarity across all emotions for a given network, for example, humans displaying stronger emotions towards Republicans for all eight emotions, as shown in Figure 15. Such a result is odd because it means that contradictory emotions, for example, joy and sadness, are simultaneously felt for the same political party. Instead, according to Plutchik's theory of emotions (Plutchik 1980), we would expect contradictory emotions to be felt more strongly towards opposite political parties, for example, joy towards Democrats and sadness towards Republicans, which is what we observe in humans when we use the L1 norm, as shown in Figure 14. The fact that the L2 normalization produces effects of the same sign for all emotions, including opposite ones, suggests that it leaves a residual difference between the two primes that is unrelated to emotions and that shifts all eight effect sizes in the same direction. The L1 normalization does not show this pattern, which is why we present the L1 results in the main text. We report both sets of results so that readers can evaluate the influence of this choice. More generally, the sensitivity of the emotions approach to the choice of norm shows that the normalization step is not neutral: depending on the position of primes and targets within the networks, the L1 and L2 norms might produce different results. The norm should therefore be chosen carefully, and results should be checked against more than one normalization, as we have done here.

Another important limitation concerns the validity of our bias metrics. Our metrics are based on final activation levels, which we interpret as a measure of the strength of association between primes and targets, and we use relative differences in association strength as a proxy for implicit bias. The interpretation of final activation levels as association strength is supported by previous work showing that activation levels obtained through spreading activation in word association networks correlate with reaction times in the lexical decision task, both for human-generated networks (Siew 2019) and for the LLM-generated networks used in this work (Abramski, Improta, et al. 2025). In addition, some of the human results obtained with our methodology are consistent with findings from human IAT research, such as pro-Christian implicit preferences in Western samples (Heiphetz et al. 2013). However, we did not directly validate our bias scores against existing bias measures for LLMs or against downstream bias benchmarks. Such a validation is not straightforward with existing resources. Our effect sizes depend on the specific sets of primes and targets that are used (for example, the 25 female-related and 25 male-related adjectives of the stereotypes approach), whereas existing benchmarks are built around different social categories and different words, and are often reported for different versions of the models. Moreover, most existing benchmarks measure explicit biases in specific tasks, such as question answering or text generation, or, in the case of model-level metrics such as the WEAT, associations in embedding space. Since implicit and explicit biases do not necessarily align, either in humans (Perugini 2005; Wilson et al. 2000) or in LLMs (X. Bai et al. 2025), a comparison with these benchmarks would not, on its own, confirm or refute the validity of our metrics. A more informative validation would require a benchmark designed for this purpose, which tests explicitly the same

prime-target pairs used in our methodology, for example by asking models directly whether a given trait applies more to women or to men, and ideally by collecting the same judgments from human participants. Developing such a benchmark is outside the scope of the present work, but it is an important direction for future research, as it would make it possible to measure the relationship between the implicit biases captured by our methodology and the explicit biases expressed by the same models, and by humans, for the same concepts.

A related limitation concerns the robustness of our results to differences in network structure and to the parameters of the methodology. The four networks differ in size and density (Table 2), and since spreading activation depends on network topology, some of the differences in bias scores between humans and LLMs could reflect differences in network structure rather than differences in associations. To check this, we also applied the methodology to reduced versions of the four networks that included only the nodes present in all four networks, so that the networks contained exactly the same set of concepts and differed only in their edges. The results obtained with these reduced networks were very similar to those obtained with the full networks. We report the results for the full networks because they are more complete representations of the conceptual knowledge of humans and LLMs. However, this check does not control for differences in density, and apart from the initial activation, which does not affect the results (see Section 2.2), and the choice of norm discussed above, we did not systematically analyze the sensitivity of the results to the parameters of the spreading activation process. Future work should test the robustness of the methodology using networks that are matched or subsampled to the same size and density, and by systematically varying the number of time steps and the retention parameter.

Our methodology features additional limitations. First, the word association networks do not have multiple nodes for words with multiple senses, and so words like *black* and *white*, which could potentially be chosen as prime nodes for investigating racial biases, can also be interpreted as colors. Therefore, results may be influenced by confounding when words with multiple meanings are included as prime nodes or target nodes. Second, the reliance on free-association norms constrains the ecological validity of the findings. While word association networks provide a valuable proxy for implicit cognitive structures (Citraro et al. 2025; De Deyne et al. 2019; Siew et al. 2019), it remains uncertain to what extent they predict real-world behaviors in domains such as healthcare, recruitment, or criminal justice, where implicit bias has direct social consequences (Perugini 2005; Y. Zhao, B. Wang, Y. Wang, D. Zhao, He, et al. 2025). However, it should be noted that this limitation is not unique to our methodology. In fact, the extent to which human IAT results can be used to predict discriminatory behavior in applied settings remains a topic of debate as well (Schütze and Asma 2025). Implicit bias evaluation provides insight into latent representational structures that influence thinking and reasoning, but decision-making in applied settings involves contextual and normative considerations that cannot easily be captured by intrinsic measures. Third, the SWOW dataset of free association norms (De Deyne et al. 2019) used to build the human word association network includes mostly participants whose native language is English (i.e. the US and the UK) and so the implicit knowledge that is modeled by the human network reflects a Western-centric worldview. As previous work has demonstrated (Xiang et al. 2025), cultural and linguistic contexts strongly shape the structure of cognitive memory networks, meaning that implicit associations in multilingual or non-Western settings may diverge significantly from those represented here. Future work should aim to use multilingual or cross-cultural human-generated data to provide a more globally representative human benchmark, though such datasets are limited. Fourth, in this study, only three LLMs (Mistral, Llama3, and Haiku) were evaluated. Although these LLMs differ in scale and architecture, this small sample limits generalization to the broader ecosystem of models, including those with distinct training pipelines (e.g. models suitable for cognitive science experiments (Binz et al. 2025)). Finally, both human and machine biases are dynamic rather than static (Y. Guo et al. 2024). Human associative norms evolve over time, and LLMs are continuously updated with new training data and alignment strategies, which can shift their internal representational structures as models evolve. The importance of this evolutionary component in the context of LLMs is highlighted by Abramski, Citraro, et al. (2023), who found that negative perceptions of STEM subjects were progressively reduced in newer versions of GPT models.

Longitudinal approaches and cross-cultural replications (Acerbi and Stubbersfield 2023) will therefore be essential for determining the robustness of the present findings. Future research should broaden the scope by applying this methodology to larger and more diverse sets of models and cultural contexts, while triangulating free-association data with behavioral and decision-level measures of bias.

4.4 Conclusions

In this work, we introduce a novel methodology for evaluating and interpreting implicit biases in LLMs with respect to humans based on word association networks. Our approach, which aligns with the implicit association test (IAT) for measuring biases in humans, offers a cognitively grounded, scalable, and transparent framework for assessing biases both quantitatively and qualitatively. Our methodology fills the gap left by existing bias evaluation methods by enabling direct comparisons between LLMs and humans. We apply our methodology to investigate implicit social biases using three different approaches that assess various ways in which biases can be expressed: through stereotypes, valence, and emotions. Our results reveal that implicit biases in LLMs, while structurally similar to those of humans, diverge in content, polarity, and intensity across various social domains. Our findings underscore the need for systematic methodologies that can capture both parallels and mismatches between human and machine associative structures. Beyond bias evaluation, our approach contributes to the broader goal of guiding the socially responsible design and deployment of language technologies that have a positive impact on society.

Acknowledgments

This work is supported by: (i) SoBigData.it which receives funding from the European Union – NextGenerationEU – National Recovery and Resilience Plan (PNRR) of Italy; (ii) the Future Artificial Intelligence Research (FAIR) Foundation, a non-profit foundation financed under the National Recovery and Resilience Plan (PNRR) of Italy; (iii) UniTrento Internal Call for Research 2023 grant from the Università degli Studi di Trento (Grant ID: PS 22_27); (iv) Ministero dell'Università e della Ricerca (MUR) under Decreto n. 23178 del 10 dicembre 2024 – Bando FIS 2 (PENSO); (v) CALCOLO from Fondazione Caritro, and COGNASCO from University of Trento.

References

A. Abid, M. Farooqi, and J. Zou. 2021. "Persistent anti-Muslim bias in large language models." In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

K. Abramski, S. Citraro, L. Lombardi, G. Rossetti, and M. Stella. 2023. "Cognitive network science reveals bias in GPT-3, GPT-3.5 Turbo, and GPT-4 mirroring math anxiety in high-school students." *Big Data and Cognitive Computing*, 7, 3, 124.

K. Abramski, R. Improta, G. Rossetti, and M. Stella. 2025. "The "LLM World of Words" English free association norms generated by large language models." *Scientific data*, 12, 1, 803.

K. Abramski, C. Lavorati, G. Rossetti, and M. Stella. 2024. "LLM-generated word association norms." In: *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 3–12.

A. Acerbi and J. M. Stubbersfield. 2023. "Large language models show human-like content biases in transmission chain experiments." *Proceedings of the National Academy of Sciences*, 120, 44, e2313790120.

H. An, C. Acquaye, C. Wang, Z. Li, and R. Rudinger. 2024. "Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender?" *arXiv preprint arXiv:2406.10486*.

E. Anzenberg, A. Samajpati, S. Chandrasekar, and V. Kacholia. 2025. "Evaluating the Promise and Pitfalls of LLMs in Hiring Decisions." *arXiv preprint arXiv:2507.02087*.

W. Babonnaud, E. Delouche, and M. Lahlouh. 2024. "The bias that lies beneath: Qualitative uncovering of stereotypes in large language models." *Swedish Artificial Intelligence Society*, 195–203.

X. Bai, 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.

Y. Bai, J. Zhao, J. Shi, Z. Xie, X. Wu, and L. He. 2024. "Fairmonitor: A dual-framework for detecting stereotypes and biases in large language models." *arXiv preprint arXiv:2405.03098*.

P. Banerjee, A. Java, S. Jandial, S. Shahid, S. Furniturewala, B. Krishnamurthy, and S. Bhatia. 2024. "All should be equal in the eyes of lms: Counterfactually aware fair text generation." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38, 17673–17681.

Y. Bang, D. Chen, N. Lee, and P. Fung. 2024. "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said." In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11142–11159.

T. Bas. 2024. "Assessing gender bias in LLMs: Comparing LLM outputs with human perceptions and official statistics." *arXiv preprint arXiv:2411.13738*.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. "On the dangers of stochastic parrots: Can language models be too big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

M. Binz et al.. 2025. "A foundation model to predict and capture human cognition." *Nature*, 1–8.

K. Brian and M. Stella. 2023. "Introducing mindset streams to investigate stances towards STEM in high school students and experts." *Physica A: Statistical Mechanics and its Applications*, 626, 129074.

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

Y. T. Cao, Y. Pruksachatkun, K.-W. Chang, R. Gupta, V. Kumar, J. Dhamala, and A. Galstyan. 2022. "On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations." *arXiv preprint arXiv:2203.13928*.

E. Cau, V. Pansanella, D. Pedreschi, and G. Rossetti. 2025. "Selective agreement, not sycophancy: investigating opinion dynamics in LLM interactions." *EPJ Data Science*, 14, 1, 59.

S. Citraro, E. Haim, A. Carini, C. S. Siew, G. Rossetti, and M. Stella. 2025. "SpreadPy: A Python tool for modelling spreading activation and superdiffusion in cognitive multiplex networks." *arXiv preprint arXiv:2507.09628*.

A. M. Collins and E. F. Loftus. 1975. "A spreading-activation theory of semantic processing." *Psychological review*, 82, 6, 407.

A. M. Collins and M. R. Quillian. 1969. "Retrieval time from semantic memory." *Journal of verbal learning and verbal behavior*, 8, 2, 240–247.

X. Dai, L. Zhou, B. Wang, and H. Li. 2025. "From word to world: Evaluate and mitigate culture bias in LLMs via word association test." In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 24521–24537.

S. De Deyne, D. J. Navarro, A. Perfors, M. Brysbaert, and G. Storms. 2019. "The "Small World of Words" English word association norms for over 12,000 cue words." *Behavior research methods*, 51, 3, 987–1006.

J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta. 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.

J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. 2021. "Documenting large webtext corpora: A case study on the colossal clean crawled corpus." *arXiv preprint arXiv:2104.08758*.

J. A. Fodor. 1998. *Concepts: Where cognitive science went wrong*. Oxford University Press.

I. O. Gallegos, 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.

A. Garimella, A. Amarnath, K. Kumar, A. P. Yalla, N. Chhaya, B. V. Srinivasan, et al.. 2021. "He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation." In: *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 4534–4545.

D. Gaucher, J. Friesen, and A. C. Kay. 2011. "Evidence that gendered wording in job advertisements exists and sustains gender inequality." *Journal of personality and social psychology*, 101, 1, 109.

S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. 2020. "RealToxicityPrompts: Evaluating neural toxic degeneration in language models." *arXiv preprint arXiv:2009.11462*.

T. Geva, A. Goldstein, E. Lary, and C. Levy. 2025. "Do LLMs Exhibit Human-Like Cognitive Biases? A Large-Scale Systematic Evaluation." *A Large-Scale Systematic Evaluation (September 17, 2025)*.

S. Goldfarb-Tarrant, R. Marchant, R. M. Sánchez, M. Pandya, and A. Lopez. 2020. "Intrinsic bias metrics do not correlate with application bias." *arXiv preprint arXiv:2012.15859*.

A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. 1998. "Measuring individual differences in implicit cognition: the implicit association test." *Journal of personality and social psychology*, 74, 6, 1464.

A. G. Greenwald, T. A. Poehlman, E. L. Uhlmann, and M. R. Banaji. 2009. "Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity." *Journal of personality and social psychology*, 97, 1, 17.

W. Guo and A. Caliskan. 2021. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 122–133.

Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, and S. S. Liu. 2024. "Bias in large language models: Origin, evaluation, and mitigation." *arXiv preprint arXiv:2411.10915*.

L. Heiphetz, E. S. Spelke, and M. R. Banaji. 2013. "Patterns of implicit and explicit attitudes in children and adults: tests in the domain of religion." *Journal of Experimental Psychology: General*, 142, 3, 864.

V. Hofmann, P. R. Kalluri, D. Jurafsky, and S. King. 2024. "AI generates covertly racist decisions about people based on their dialect." *Nature*, 633, 8028, 147–154. doi:10.1038/s41586-024-07856-5.

K. A. Hutchison et al.. 2013. "The semantic priming project." *Behavior research methods*, 45, 1099–1114.
H. Inan et al.. 2023. "Llama Guard: LLM-based input-output safeguard for human-AI conversations." *arXiv preprint arXiv:2312.06674*.
T. Katô and Y. Miyao. 2024. "Analyzing Correlations Between Intrinsic and Extrinsic Bias Metrics of Static Word Embeddings With Their Measuring Biases Aligned." *arXiv preprint arXiv:2409.09260*.
H. Koh, D. Kim, M. Lee, and K. Jung. 2024. "Can LLMs recognize toxicity? Structured toxicity investigation framework and semantic-based metric." *CoRR*.
H. Kotek, R. Dockum, and D. Sun. 2023. "Gender bias and stereotypes in large language models." In: *Proceedings of the ACM collective intelligence conference*, 12–24.
K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. 2019. "Measuring bias in contextualized word representations." *arXiv preprint arXiv:1906.07337*.
J. Lee, Y. Hicke, R. Yu, C. Brooks, and R. F. Kizilcec. 2024. "The life cycle of large language models: A review of biases in education." *arXiv preprint arXiv:2407.11203*.
X. Lin and L. Li. 2025. "Implicit bias in LLMs: A survey." *arXiv preprint arXiv:2503.02776*.
K. Lum, J. R. Anthis, C. Nagpal, and A. D'Amour. 2024. "Bias in Language Models: Beyond Trick Tests and Toward RUTEd Evaluation." *arXiv preprint arXiv:2402.12649*.
C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger. 2019. "On measuring social biases in sentence encoders." *arXiv preprint arXiv:1903.10561*.
T. P. McNamara. 2005. *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.
S. Mohammad. 2018. "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words." In: *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 174–184.
F. Motoki, V. Pinho Neto, and V. Rodrigues. 2024. "More human than human: measuring ChatGPT political bias." *Public Choice*, 198, 1, 3–23.
M. Nadeem, A. Bethke, and S. Reddy. 2020. "StereoSet: Measuring stereotypical bias in pretrained language models." *arXiv preprint arXiv:2004.09456*.
J. H. Neely. 2012. "Semantic priming effects in visual word recognition: A selective review of current findings and theories." *Basic processes in reading*, 264–336.
D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. 2004. "The University of South Florida free association, rhyme, and word fragment norms." *Behavior Research Methods, Instruments, & Computers*, 36, 3, 402–407.
N. Padilla, H. T. Lam, A. Lambrecht, and B. Hollenbeck. 2025. "The impact of LLM adoption on online user behavior." *Available at SSRN*.
A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. R. Bowman. 2021. "BBQ: A hand-built bias benchmark for question answering." *arXiv preprint arXiv:2110.08193*.
M. Perugini. 2005. "Predictive models of implicit and explicit attitudes." *British Journal of Social Psychology*, 44, 1, 29–45.
R. Plutchik. 1980. "A general psychoevolutionary theory of emotion." In: *Theories of emotion*. Elsevier, 3–33.
R. Poulain, H. Fayyaz, and R. Beheshti. 2024. "Bias patterns in the application of LLMs for clinical decision support: A comprehensive study." *arXiv preprint arXiv:2404.15149*.
G. Rossetti, M. Stella, R. Cazabet, K. Abramski, S. Citraro, E. Cau, A. Failla, V. Morini, and V. Pansanella. 2026. "YSocial: An artificial intelligence powered social media virtual twin." *Big Data & Society*, 13, 3, 20539517261431576.
D. Rozado. 2024. "The political preferences of LLMs." *PloS one*, 19, 7, e0306621.
J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff. 2019. "Masked language model scoring." *arXiv preprint arXiv:1910.14659*.
N. Scherrer, C. Shi, A. Feder, and D. Blei. 2023. "Evaluating the moral beliefs encoded in LLMs." *Advances in Neural Information Processing Systems*, 36, 51778–51809.
K. Schnabel and J. B. Asendorpf. 2013. "Free associations as a measure of stable implicit attitudes." *European Journal of Personality*, 27, 1, 39–50.
L. Schützle and L. J. F. Asma. 2025. "Does Empathy Reduce Implicit Bias? A Critical Review." *Topoi*, 1–12.
A. Semeraro, S. Vilella, R. Improta, E. S. De Duro, S. M. Mohammad, G. Ruffo, and M. Stella. 2025. "EmoAtlas: An emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science." *Behavior Research Methods*, 57, 2, 77.
T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, and D. Xiong. 2023. "Large language model alignment: A survey." *arXiv preprint arXiv:2309.15025*.
C. S. Siew. 2019. "spreadr: An R package to simulate spreading activation in a network." *Behavior Research Methods*, 51, 2, 910–929.
C. S. Siew, D. U. Wulff, N. M. Beckage, and Y. N. Kenett. 2019. "Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics." *Complexity*, 2019, 1, 2108423.
A. Sorokovikova, P. Chizhov, I. Eremenko, and I. P. Yamshchikov. 2025. "Surface Fairness, Deep Bias: A Comparative Study of Bias in Language Models." *arXiv preprint arXiv:2506.10491*.
D. A. Stanley, P. Sokol-Hessner, M. R. Banaji, and E. A. Phelps. 2011. "Implicit race attitudes predict trustworthiness judgments and economic trust decisions." *Proceedings of the National Academy of Sciences*, 108, 19, 7710–7715.

R. Steed, S. Panda, A. Kobren, and M. Wick. 2022. "Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3524–3542.

M. Stella, T. T. Hills, and Y. N. Kenett. 2023. "Using cognitive psychology to understand GPT-like models needs to extend beyond human biases." *Proceedings of the National Academy of Sciences*, 120, 43, e2312911120.

K. Struniawski, R. Kozera, and A. Konopka. 2025. "Bias or Justice? Analyzing LLM Sentencing Variability in Theft Indictments Across Gender, Ethnicity, and Education Factors." In: *International Conference on Computational Science*. Springer, 19–32.

F. Sun, N. Li, K. Wang, and L. Goette. 2025. "Large Language Models are overconfident and amplify human bias." *arXiv preprint arXiv:2505.02151*.

L. Tjuatja, V. Chen, T. Wu, A. Talwalkwar, and G. Neubig. 2024. "Do LLMs exhibit human-like response biases? A case study in survey design." *Transactions of the Association for Computational Linguistics*, 12, 1011–1026.

L. Vicente and H. Matute. 2023. "Humans inherit artificial intelligence biases." *Scientific reports*, 13, 1, 15737.

G. Vigliocco and D. P. Vinson. 2007. "Semantic representation." *The Oxford handbook of psycholinguistics*, 195–215.

Y. Wan, 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." *arXiv preprint arXiv:2310.09219*.

K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, E. Chi, and S. Petrov. 2020. "Measuring and reducing gendered correlations in pre-trained models." *arXiv preprint arXiv:2010.06032*.

T. D. Wilson, S. Lindsey, and T. Y. Schooler. 2000. "A model of dual attitudes." *Psychological review*, 107, 1, 101.

C. Xiang, C. Liu, S. De Deyne, and L. Frermann. 2025. "Comparing Moral Values in Western English-speaking societies and LLMs with Word Associations." *arXiv preprint arXiv:2505.19674*.

B. Xiao, X. Duan, D. A. Haslett, and Z. Cai. 2025. "Human-likeness of LLMs in the Mental Lexicon." In: *Proceedings of the 29th Conference on Computational Natural Language Learning*, 586–601.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. 2018. "Gender bias in coreference resolution: Evaluation and debiasing methods." *arXiv preprint arXiv:1804.06876*.

Y. Zhao, B. Wang, Y. Wang, D. Zhao, R. He, and Y. Hou. 2025. "Explicit vs. implicit: Investigating social bias in large language models through self-reflection." *arXiv preprint arXiv:2501.02295*.

Y. Zhao, B. Wang, Y. Wang, D. Zhao, X. Jin, J. Zhang, R. He, and Y. Hou. 2024. "A comparative study of explicit and implicit gender biases in large language models via self-evaluation." In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 186–198.

Y. Zhou, X. Wu, B. Huang, J. Wu, L. Feng, and K. C. Tan. 2024. "CausalBench: A comprehensive benchmark for causal learning capability of LLMs." *arXiv preprint arXiv:2404.06349*.

A Emotions Approach with L2 Normalization

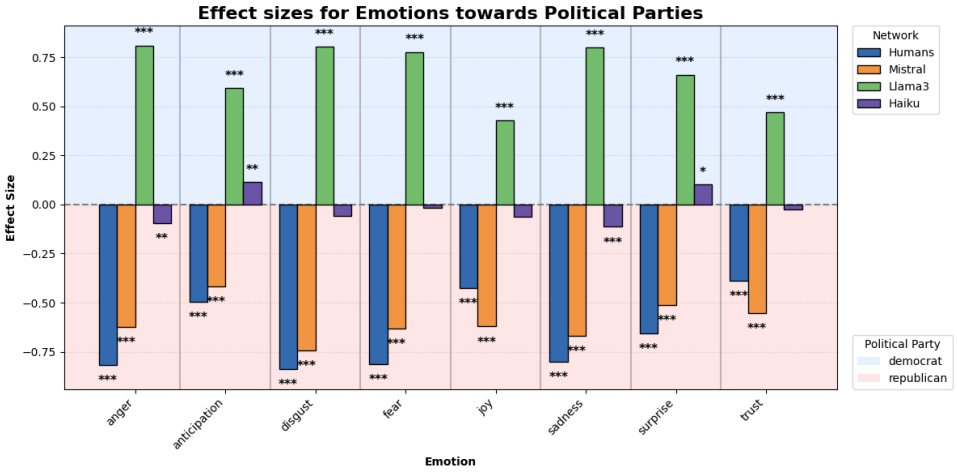


Fig. 15. **Effect sizes for the emotional analysis of political parties (L2 normalization).** The barplots show the effect sizes for each network for all eight core emotions. Statistical significance is indicated with stars. For each emotion, positive effects indicate stronger feelings of that emotion towards Democrats, while negative effects indicate stronger feelings towards Republicans. Note, the final activation level matrices were normalized using the L2 norm (see Methods).

B Availability of Data and Materials

All data and code have been made available in the following public Github repository: https://github.com/LLMWorldOfWords/BIASsociation, with the exception of the original SWOW dataset (De Deyne et al. 2019), which can be downloaded from the Github repository: https://github.com/SimonDeDeyne/SWOWEN-2018. Note that the SWOW dataset is protected by a license that prohibits the distribution of modified material, and for this reason we do not include the modified version of the SWOW dataset in this work's repository.

C Competing interests

The authors declare no competing interests.

D Author Contribution Statement

K.A., G.R., and M.S. conceived the methodology. K.A. conducted the analyses and analyzed the results. K.A., G.R., and M.S. interpreted the results and reviewed the manuscript.

Received 03 March 2026; accepted 24 October 2026