Modelling Implicit Bias in Gender–Career Associations: A systematic comparison of language models

Alexander Porshnev (Maynooth University)*

Kevin Dirk Kiy (Maynooth University)

Diarmuid O'Donoghue (Maynooth University)

Manokamna Singh (Maynooth University)

Cai Wingfield (University of Birmingham)

Dermot Lynott (Maynooth University)*

Keywords

Linguistic distributional models, computational modelling, large language models, implicit association test, bias, gender-career, implicit attitudes, behavior

Abstract

Biases in language and their reflection in language models have attracted researchers' attention, particularly with the growth of large language models (LLMs). However, many questions on the links between language models and people's biased attitudes remain unanswered. In the current study we focus on gender–career bias to examine the extent to which language models can be used to model behavioural responses in the Gender–Career Implicit Association Test (IAT). We provide a systematic evaluation of a range of language models, including n-gram, count vector, predict (word2vec), and Large Language Models (LLMs), to determine how well they capture people's behaviour in the IAT. We compared response time data from over 800,000 participants against 25 language models, with a total of 675 model variants. We find that many language models, including large language models (LLMs), correlated well with human behavior. While results support previous findings for both *predict* and *count* model families, we observed that performance of LLMs was consistently different from that of simpler *predict* models, particularly in terms of the direction and strength of correlations with reaction time and bias. This divergence may indicate successful attempts to mitigate bias in LLMs while preserving other aspects of linguistic information. Our findings reinforce the idea that societal biases are generally encoded in language, but that large language models can exhibit behaviors different to classical language models.

* Corresponding authors: alexander.porshnev@mu.ie (A. Porshnev); dermot.lynott@mu.ie (D.Lynott)

1. Introduction

Tina is less likely to be called for an interview than Tom. Darius is less likely to get a job than Dylan, and obese applicant Jenna is less likely to be shortlisted than her perceived healthy colleague Claire. Implicit biases (ones that we are not consciously aware of) are pervasive in our society (Staats, 2016) and can lead directly to prejudicial decision-making (e.g. Chang, 2011; Moss-Racusin et al., 2012; O'Brien et al., 2013). Biases linked to gender, race, perceived health status, and many other characteristics are seen in employment, education, criminal justice, politics, and healthcare (Greenwald & Krieger, 2006). For example, in employment contexts, prospective female employees are rated as less competent and hireable than (identical) male applicants, and are less likely to be offered a job (Moss-Racusin et al., 2012). Similarly, while women engineers publish in journals with higher Impact Factors than their male peers, they receive fewer citations from the scientific community (Ghiasi et al., 2015). In these ways gender-bias in career progression is readily visible, in both empirical studies and in the workplace.

Such systemic patterns make the issue of bias one of global importance, and one with significant economic and societal costs. For example, employees who perceive bias are more than three times as likely to quit their jobs, with an estimated annual cost of up to $550 billion in the US alone (O'Boyle & Harter, 2013). Yet despite the acknowledged prevalence of such biases, we still do not fully understand where these biases come from or how they are transmitted (Nosek et al., 2012).

Even with great advancements in AI and machine learning, new technologies continue to reflect underlying societal biases, with numerous examples of bias evident in AI models, across a range of settings. For example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system was developed to provide estimates of the likelihood of convicted criminals reoffending (see Brennan & Dieterich, 2018). Unfortunately, the model systematically overestimated the reoffending risk of black criminals, who did not go on to reoffend, and systematically underestimated the reoffending risk of white criminals, who did go on to reoffend (Angwin et al., 2022). In healthcare, an AI model called Impact Pro was used by UnitedHealth to determine patient eligibility for certain high-risk treatment pathways. It was found that the model categorised the health needs of black patients as being of lower risk than those of white patients, and therefore denied them access to appropriate specialised treatments (Takshi, 2020). In a study of a range of LLMs used in medical contexts, Omar et al., (2025) found that patients who were categorised as being Black, homeless, or identifying as LGBTQIA+, were more likely than other groups to be directed towards urgent care, invasive interventions, or mental health evaluations (Omar et al., 2025). In financial services, evidence of racial bias in the availability of financial products has been well-established, including in the denial of home loans and the level of calculated interest (e.g. Bartlett et al., 2022; Wachter & Megbolugbe, 1992). What's more, Zou and Khern-am-nuai (Zou & Khern-am-nuai, 2023; see also Bowen III et al., 2024) found that using off-the-shelf AI models exacerbated such bias, increasing the rate at which black applicants were denied loans from 54% to 67%. Thus, the adoption of AI models, may not only inherit biases ingrained in training data, but can also amplify the effect of those biases. Bias in such models can therefore have real, tangible effects on people in the world, making it critical that we understand both the origins and mechanisms that lead to such bias.We might ask first, where do these implicit biases come from? To look at the literature, it has variously been suggested that implicit attitudes may stem from early experiences (Rudman et al., 2007), automatic biases towards outgroup members (Dasgupta & Greenwald, 2001; Gonzalez et al., 2017), from a process of socialisation with others (Yasui, 2015), or that implicit attitudes can be acquired through learning of cultural norms (e.g., being told a specific positive or negative opinion about a particular group – Hudson et al., 2024; Kurdi & Dunham, 2020). While each of these routes can provide a source for attitude development, some have also suggested that implicit attitudes stem from a cumulative exposure to evaluative associations in our environment (McConnell et al., 2018). Thus, language, as part of our cultural life, provides a unique means of expressing attitudinal information, since, through language we frequently encounter positive and negative evaluations of things in the world. Furthermore, language can also provide a mirror to people's non-linguistic experience (Lynott & Connell, 2013), and provide a mechanism (e.g., via Hebbian or reinforcement learning) for how linguistic experiences may be linked to the development and transmission of biased attitudes (Hudson et al., 2024; Kurdi & Charlesworth, 2023). Surprisingly, even in some discussions on the role of culture as a source of implicit attitudes, the specific role of language is often given scant attention or only passing reference (see e.g., Rudman, 2004; Kashima et al., 2015). Despite such treatments, we suggest that language may have a fundamental role to play in the development of implicit attitudes, with recent evidence supporting such a proposal. Indeed, recent work in computational modelling suggests that implicit biases may be captured by the latent statistical patterns in language (Lynott et al., 2012; Caliskan et al., 2017; Onnis & Lim, 2024). In other words, the way words appear together in language may influence our unconscious, and potentially prejudicial, attitudes towards others. However, we still do not have a good handle on whether these patterns can be adequately captured by existing language models.

Nevertheless, recent work has found that the linguistic distributional patterns of words closely correspond to people's biases and their prejudicial judgements (Caliskan et al., 2017; Lynott et al., 2012, 2019). For example, using implicit association tests (IAT), we can see that positive or negative biases are reflected in how often positive or negative words are associated with a particular concept. In an IAT, participants classify stimuli into categories as quickly as possible, where faster responses indicate stronger associations between concepts (Arkes & Tetlock, 2004). Displaying a greater degree of bias results in a higher D score for a participant. For example, in a Gender–Career IAT which contrasts male and female names, people consistently respond more quickly when male names are paired with career-related concepts (e.g., "John" and "Management"), compared to male names paired with non-career-related concepts (e.g., "John" and "family"), and vice versa for pairings with female names and concepts. This pattern indicates stronger negative associations for women and career concepts. Thus, participants tend to respond more quickly to congruent stimuli pairings (i.e., male names and career concepts, or female names and family concepts) compared to incongruent stimuli pairings (i.e., male names and family concepts, or female names and career concepts). This is not to say that every individual participant follows this pattern, but over a large sample of participants this is the pattern that emerges (e.g. Nosek et al., 2007).

A huge benefit to using the IAT in the present context is that Project Implicit, which hosts multiple IATs on their platform, openly shares their extensive datasets that includes data ranges from 2005 to the present day, meaning that there are responses from tens of millions of participants on various topics. Using such large data sets means that even after filtering out participants according to various selection criteria or using different covariates, there is still ample statistical power to provide good estimates of any underlying effects. The same level of data is simply not available for other measures of implicit attitudes, meaning they are less amenable to statistical modelling or machine learning techniques.

Researchers may disagree over what exactly the IAT measures – does it measure associations? Bias? Unconscious attitudes? Or something else? For our purposes, these differences are not as important as trying to model the documented effects of the IAT. For example, knowing that a particular language model can capture the behavioural data of the IAT is not diminished by our understanding of whether the underlying effect is driven by conscious or unconscious processing.

In terms of the properties of the IAT, it has been found to have generally strong internal reliability, with Cronbach's Alpha in the range (0.7–0.9) and demonstrates good construct validity (Nosek, Greenwald, et al., 2007; Nosek, Smyth, et al., 2007; Schimmack, 2021). It should be noted that some have observed less-than-satisfactory test-retest reliability levels (Lai & Wilson, 2021), but nonetheless, on this dimension the IAT still outperforms many other measures of implicit attitudes, such as the Go–No Go Association Test, Sorting Paired-Features task, Evaluative-Priming, and many others (Bar-Anan & Nosek, 2014; LeBel & Paunonen, 2011). Thus, despite some shortcomings, the positive aspects of using the IAT have resulted in it being the most used measure of people's automatic associations between concepts (e.g., Greenwald et al., 2003), and make the IAT an excellent candidate for modelling in the current context.

1.2 Background on language models, linguistic distributional knowledge, and modelling biased attitudes

Prior to describing the current study, we first describe three examples that are representative of how language models can be used to capture the extent of human biases measured by IATs.

Using n-gram co-occurrence counts, Lynott et al., (2012) found a strong correlation between the implicit biases predicted by distributional properties of the linguistic stimuli used in 16 IATs (related to a broad range of topics, including: race, gender, obesity, drug use etc.) and the actual degree of bias indicated by the human behavioural data. Caliskan and colleagues (2017) extended this logic using GloVe (Pennington et al., 2014) to create a Word Embedding Association Test (WEAT), a linguistic analogue of the IAT. They found that the biases observed in the language model reflected known behavioural biases (e.g. people's preferences for flowers over insects, the negative bias towards stereotypical Black names, and gender bias observed in various professions).

More recently, Bhatia and Walasek (2023) showed that distributed semantic representations not only predict data from an IAT, but that language biases more strongly correlate with implicit attitudes than explicit ones (i.e., where people are asked directly about specific biases). Building upon Caliskan et. al.'s (2017) work, Bhatia and Walasek compared the predictive power of the WEAT with their own Valence Estimation Model (VEM; based on Warriner et al., 2013), which includes both linguistic distributional information (as in the WEAT) and valence information about stimuli (i.e., measures of how positive or negative those stimuli are perceived). Bhatia and Walasek found that biases measured in both the WEAT (R^2 = 0.13, r = 0.361) and the VEM (R^2 = 0.26 ,r = 0.51) significantly correlated with human IAT data, but that the valence model provided a significantly better fit to the human data (Bhatia & Walasek, 2023). The authors suggest that implicit attitudes can be better predicted by combining psychological data (like valence norms) and large-scale language data (word embeddings), rather than relying on linguistic data alone.

While the above examples are suggestive of the power of language models to capture human implicit attitudes and biases, they are limited in other regards. For example, the work of Lynott et al., uses relatively small sample sizes to provide the human behavioural data, while Caliskan et al., does not directly link language model data to human data at all. Instead, Caliskan et al., show that different biases exist within language models that *resemble* those seen in human behavioural data. The work of Bhatia and Walasek goes further in this regard, using much larger samples of human participants, which provides the much-needed statistical power for these kinds of analyses, and directly links human behaviour with language model performance.

However, a drawback in all three cases is that they rely on a very small number of models to determine if language models can reflect human attitudes and biases, while others have suggested that certain language models may not be capable of robustly detecting certain forms of bias. For example, Zhang and colleagues (e.g. Zhang et al., 2020) suggest that bias measures using word embeddings and based on word pairs lack robustness and reliability. Given the vast number of language models currently available, and the range of parameters that can be tweaked for any given model, it remains unclear how well language models *in general* do at this task, or whether good performance is limited to a small subset of models, within very specific parameter settings.

1.3 The current study

To address this issue, in the current study we conduct a systematic analysis of a range of language models, parameters and semantic distance measures to determine a) whether language models generally capture behavioral responses to gender–career concepts as measured by the implicit association test, and consider which models show greatest alignment with people's behavior, b) what distance measures show the best relationship between language models and human behaviour, and c) whether Large Language Models (LLMs) consistently outperform non-LLM models, such as count-vector, predict, and n-gram.

While much work in AI assumes that larger and more computationally-intensive models will perform better in most tasks, findings in the cognitive modelling literature suggest that this is not always the case. For example, as tasks become more complex, simpler models often do as well as, if not better than, more complex models (see Wingfield & Connell, 2022; Grinsztajn et al., 2024). Thus, a priori we might expect LLMs to perform very well, but their greater computational complexity might not buy them as much of an advantage as one might expect. As well as modelling performance, there are of course additional reasons for exploring simpler approaches than LLMs, including the high energy and financial costs, the environmental impact of LLM training and use (Luccioni et al., 2024), possible exposure of sensitive information (Jaff et al., 2024), connectivity issues, potential reliance on un-governed corporations, and the questionable traceability, explainability and reproducibility of results. Thus, we hope that the current study will provide some insights into how well LLMs and other language models can capture human behaviour linked to implicit bias.

2. Method

2.1 Families of models and specific model implementations

We examined four families of language models that ranged considerably in their complexity: n-gram, count, predict (Word2vec), and LLM. The first three families have been used extensively in previous research, particularly in cognitive and psycholinguistic work (e.g. Connell & Lynott, 2014; Landauer & Dumais, 1997; Wingfield & Connell, 2022), which has found that larger models do not necessarily lead to better performance in modelling cognitive tasks (see e.g., Wingfield & Connell, 2022). We also included 12 publicly available large language models, which were primarily developed with a focus on text generation from Meta (Llama) and Mistral.AI (Mistral). We provide an overview of each model family and its associated measures below.

N-gram models. These are relatively simple models where text is represented by chunks of *n* words. In these models, each word is characterized by a vector of co-occurrence frequencies between the target word and the *n-1* surrounding words. We used n-gram models to calculate the following measures: conditional probability (a measure of the probability of a word, given that another word has already occurred (conditional_probability), logarithm of n-gram frequency (log_ngram_freq, the log transformation of the frequency of a given n-gram with Laplace smoothing), pointwise mutual information (pmi_ngram – a measure of association that compares the probability of two words occurring together to the probability of the words appearing independently), positive pointwise mutual information (ppmi_ngram – similar to PMI, but focuses only on whether pairs of words co-occur more frequently than would be expected given their independent frequencies), and probability ratio (probability_ratio – compares the probability of finding a context and target together to the probabilities of finding the context and target separately). The details and calculation process are thoroughly described in Wingfield and Connell (2022).

Count-Vector Models: Context-counting vectors are more advanced models that consider words within a specified context window (i.e., words to the left and right of the target word), resulting in a co-occurrence frequency vector with dimensions equal to the number of unique words in the corpus. To create vector representations, we used the same measures of conditional probability, logarithm of word co-occurrence, pointwise mutual information, positive pointwise mutual information, and probability ratio. We also include the GloVe model (Pennington et al., 2014), which has been used extensively, including in studies specifically examining biased attitudes (Bhatia & Walasek, 2023; Caliskan et al., 2017). GloVe is an openly available model that uses an unsupervised learning algorithm performed on aggregated global word-word co-occurrence statistics from a large corpus of approximately 42 billion words. Unlike the other count-vector models, GloVe is an "off-the-shelf" model, and is not customisable in terms of parameter settings such as embedding size, radius etc.

Predict Models: Predict or Word2vec models are those where word vectors are generated by artificial neural networks trained on corpus data with selected word-window sizes. Mikolov et al. (2013) suggested two model architectures to produce vector representations of words: Continuous Bag of Words (CBOW), which aims to predict the target word from an unordered collection of context words, and Skip-gram, which focuses on predicting each of the context words separately from the target word. Predict models vary in their complexity in terms of the embedding sizes, ranging here from 50 to 500. The details and calculation process are fully described in Wingfield and Connell (2022), so are not reiterated here.

Large Language Models: LLMs are the next generation of artificial neural networks trained with a large number of units in each of many hidden layers used for training. In the current work, we selected models from Meta and MistralAI with a demonstrated long-term commitment to open-source research (Jiang et al., 2023; Touvron et al., 2023). There are several reasons for studying Llama and Mistral, as opposed to focussing on other available LLMs. First, Mistral and Llama are two of the leading open source LLM architectures. By contrast, the more famous ChatGPT is not open source. Second, Meta and Mistral offered a range of generational variants allowing us to study both newer and older models (three generations of Llama and Mistral models). Various model sizes (from 7bn to 70bn parameters) allow exploration of the fidelity of the learned information, along with quantization and other model variants. Third, the selected models are non-stochastic[1], unlike other available models. Finally, both Llama and Mistral models are both open source and open weights, which enables researchers to have a much clearer view of the inner-workings and representations of such models, compared to other more black-box approaches.

Specifically, in the present study, we utilized nine models from the LLaMA family (first generation: 7B, 30B, 65B; second generation: 7B, 13B, 70B; third generation: 8B, 70B* (Meta and Hermes)) and three models from Mistral family (v0.1, v0.2, v.03) accessing their quantized transformer models via the Hugging Face service (see Supplementary materials, for full description of the models). Together, these models provide exemplars of currently used transformer LLMs.

While there are clear algorithmic differences between the families of models outlined above, it is important to also highlight differences in the usage and customizability between the "custom" trained models – n-gram, count vector, and predict (excluding GloVe) – and "off the shelf" models –LLMs and GloVe. For the custom models, we had full control over the training corpora (using three different corpora: UK Web As Corpus – UKWAC, British National Corpus – BNC, and the BBC Subtitle Corpus; details provided below), context window sizes (1, 3, 5, 10), and embedding sizes (for predict models: 50,

[1] *A concern was that the LLMs tested here might also respond in a stochastic manner, similar to ChatGPT and other commercial models. Therefore we tested this explicitly by doing multiple runs of the models and examining the outputs. We established that the token embeddings are fixed, so the results do not differ over multiple runs, and distance measures were consistent across all runs to at least 4 decimal places.*

100, 200, 300, 500). For GloVe, information about training corpora used and embedding sized is documented and available (https://nlp.stanford.edu/projects/glove/), but for LLMs we have significantly less information about the corpora used and other training features (Touvron et al. 2023, Jiang et a. 2023).

An additional difference between “classical” and large language models is in the unit of analysis. In the smaller, classical models the basic units are words, but LLMs typically use tokens that include subwords. For example, in the Mistral LLM the word “Michelle” is tokenized as three tokens: “Mic”, “hel”, “le” and also with different tokens for lower- and upper-case uses, that leads to different vectors and therefore different distances between words in the embedding space. For our purposes, as our stimuli include male and female names, with LLMs we calculated distances for two sets of stimuli; one lowercase (as used in “classical” models) and another with capitalized names, which better reflects the presentation of the stimuli to participants.

2.2 Distributional measures

Different models use different representations for words, and therefore allow different measures of association and similarity between words within their semantic spaces. These measures of association can be taken as the extent to which concepts are related, and so capture bias in terms of the degree of association between concepts. For example, in an n-gram model, words are represented as vectors of variable length which represent how often the target word has appeared in the context of other words in the corpus. The similarity between words is compared by simply looking up one word’s distributional score in the context of the other. N-gram models therefore capture direct co-occurrences (or first-order relations) between words, with more frequent co-occurrences associated with higher similarity or relatedness. In this way, “dog” and “cat” might receive a high similarity score because they are often mentioned together in contexts about pets. By contrast, “dog” and “gorilla” would receive lower similarity scores because they rarely appear together within the same narrow context window.

Measures such as conditional probability and PPMI use this basic co-occurrence information to provide different estimates of the similarity of pairs of words. These calculations are provided in more detail in Section 2.1. By contrast, in a vector-based representations (count-vector, predict, and LLM families) two words are compared by selecting their respective fixed-dimensionality vector representations in the model, and calculating the distance between them using vector metrics (Euclidean distance, Cosine distance, Correlation distance). For example, the word “nurse” might be closer in space to the word “woman” than the word “man”, and in this way, bias is reflected in terms of distance between concepts in a high-dimensional semantic space. Wingfield and Connell provide a detailed description of how different types of model extract different types of distributional relations, depending on the model’s particular architecture (Wingfield & Connell, 2022).

2.3 Training corpora

While we used pre-trained models such as GloVe and the LLMs, for other models we were able to vary the training corpus to include corpora of different characteristics. The GloVe corpus contains approximately 42 billion words (plus Gigaword 5), with the Llama 2 corpus containing 2 trillion tokens, and the Mistral corpus containing at least several hundred billion words, although Mistral have been reluctant to share exact details (see e.g., Bommasani et al., 2024). While the assumption might be that the larger the training corpus the better, this is not always necessarily the case, with smaller, but less noisy training data sometimes outperforming models trained on larger, but noisier data sets (Wingfield & Connell, 2022). Thus, we use three corpora that vary in terms of size and quality: The British National Corpus (BNC, 100 million words; (BNC Consortium, 2007)), The United Kingdom Web as Corpus (UKWAC, 2 billion words; (Ferraresi et al., 2008; Baroni et al., 2009)) and subtitles of British television programmes SUBTLEX-UK (Subtitles, , and 200 million words, (Van Heuven et al., 2014)). . For further details of preprocessing of these corpora, we refer the reader to Wingfield & Connell (2022, pp 15–16). Table 1 provides a summary of the different measures, distances and other parameter settings used with each model, and the total number of models used from each of the four model families.

Table 1. Summary of all models, including variants by corpus, window radius, and embedding size. Custom models are those where various parameters have been manipulated by the researchers, while Off-the-Shelf Models are those that, while transparent, have fixed sets of parameters, including training corpus, embedding size, and so on.

Model family	Model	Window radius	Embedding size	N total models
		Custom models		
count[a,b]	Conditional probability	1,3,5,10		36
count[a,b]	Log cooccurrence frequency	1,3,5,10		36
count[a,b]	PMI	1,3,5,10		36
count[a,b]	PPMI	1,3,5,10		36
count[a,b]	Probability ratio	1,3,5,10		36
n-gram[a]	Conditional probability	1,3,5,10		12
n-gram[a]	Log n-gram frequency	1,3,5,10		12
n-gram[a]	PMI n-gram	1,3,5,10		12
n-gram[a]	PPMI n-gram	1,3,5,10		12
n-gram[a]	Probability ratio n-gram	1,3,5,10		12
predict[a,b]	Skip-gram	1,3,5,10	50, 100, 200, 300, 500	180
predict[a,b]	CBOW	1,3,5,10	50, 100, 200, 300, 500	180
		Off-the-Shelf Models		
count[b]	GloVE 42B	Global	300	3
llm[b]	Llama-7b, 30b, 65b	4k	4k	18
llm[b]	Llama-2 (7b, 13b, 70)	4k	8k	18
llm[b]	Llama-3.1 (7b, 70B Hermes, 70B Meta)	128k	4k	18
llm[b]	Mistral-7b-v0.1	4k	4k	6
llm[b]	Mistral-7b-v0.2 (Dolphin 2.8)	32k	4k	6
llm[b]	Mistral-7B-v0.3	32k	4k	6

[a] Each of “custom” models have three modifications related to corpora on which it was trained (BNC, Subtitles, UKWAC)

[b] Three distances were calculated Euclidean, Cosine, Correlation for each vector model

2.4 Behavioral data and evaluating model performance

Human behavioral data was obtained from the Project Implicit Gender & Career IAT study, via an Open Science Framework repository (Nosek et al., 2015), including people’s response times (RTs) for blocks with "congruent" and "incongruent" stimuli.

We included only participants who chose UK or USA as their current residence and country of their origin, reflecting the English-language nature of the corpora used to generate the language models. We selected participants >18 years of age, and who had completed the common version of the IAT (without additional stimuli), and who had not previously completed an IAT task (Cochrane et al., 2023; Röhner & Lai, 2021). We ensured that raw trial data was available, and that our calculation of the bias effect size D was equal to the effect size in the preprocessed dataset (avoiding discrepancies between raw and preprocessed data). Table 2 provides a summary of data preprocessing.

Table 2 The table provides a summary of data preprocessing, indicating the total number and percentage of participants excluded at each stage of preprocessing. N is the number of participants at each stage of preprocessing; "N filtered out" is the number of participants removed at each stage, and "% of filtered out from Total N" is the number of removed participants expressed as a percentage of the Total Number of Participants

Preprocessing stage (2005–2021)	N	N filtered out	% of filtered out from Total N
Participants from countries USA and UK older than 18 (Total N)	1,425,903		
Participants without discrepancies between raw and preprocessed data	1,369,188	-56,715	3.98%
Participants with less than 30% errors per session	1,356,442	-12,746	0.93%
Participants with mean response time within 3 standard deviations from overall mean response time (by year)	1,351,418	-5,024	0.37%
Participants without prior experience with IAT tasks (within 15 minute timeframe)	806,983	-544,435	40.29%
Participants not registered for other IAT tasks	802,071	-4,912	0.61%
Participants with standard IAT tasks	802,070	-1	0.00%

After data cleaning, the final dataset comprised 802,070 participants from the USA and UK, spanning the period from 2005 to 2021. Next, we randomly split the whole sample into two subsamples (Sample A, Sample B) of equal size and for each sample data we calculated mean response time (m_RT) for each pair of stimuli in each condition (congruent vs incongruent), separated by country (USA, UK). Descriptive statistics for both samples are provided in Supplementary materials, Tables A.1.1-A.1.4.

2.5 Preregistration and analysis

We preregistered our approach to data handling and our planned analyses (https://aspredicted.org/ZWZ_9RV). Here we investigated three primary hypotheses: first, that the linguistic distributional relationship between stimuli words (e.g. Ben, Rebecca) and category related words (e.g. Family, Career) will be a significant predictor of the response times (RT) for participant responses on in the implicit association test. Specially, we expected that the smaller the distributional distance between words (using cosine, correlation, and Euclidian distances) in word vectors (count vector and predict vector models), the faster reaction times will be. Second, we expected that linguistic features (vector model, distance, etc.) will provide more information in comparison with the baseline model (containing variables such as country, word frequency and other lexical variables). Third, we expected to observe similar patterns of effects over multiple years. As such, we reran our main analysis separately for the years 2009, 2014, and 2019. Finally, we report an additional set of robustness checks, also preregistered (see below).

Our primary analysis involves conducting multiple regression analyses to model the mean response time (RT) of participants in the IAT. First, using Subsample A, we established a baseline regression model containing the factors of *country, log word frequency, number of letters* and *number of syllables* as predictors, to account for important predictors of reading and processing time (see e.g. Dymarska et al., 2023), but which are not of theoretical importance in this case. Subsequently, we add a single distance measure from one language model as a regressor to determine if this model improves fit over the baseline model, using $p < .05$ as a threshold for significance. For each distance measure, we calculate the Pearson correlation coefficient, r, and the change in Bayesian Information Criterion (BIC) to provide complementary measures of how well the model fits the RT data (with a BIC of 3 as a threshold for positive support of additional information provided by a given language model (as suggested at Raftery, 1995). Because the sign of correlations indicating a good fit between model values and behavioural data differed between models (Wingfield & Connell, 2022), we report absolute Pearson's correlation values for ease of cross-comparison. We ran all regression models in R (R Core Team, 2023) and flexmix (Grün & Leisch, 2023) as well as multiple helper packages (Haghish, 2017; Lüdecke, 2024; Lüdecke et al., 2021; Revelle, 2024; Wickham et al., 2023; Xie, 2021).

To test for the robustness these findings, we ran the same models on the second half of the data (subsample B), and then compared the performance across Subsamples A and B to determine if there was consistency across the samples using a number of different measures: correlation direction between model estimates in A and B samples, inclusion of B sample correlation estimate within 95% Confidence Intervals of the A sample correlation, whether the direction of the regression coefficient is the same for A and B samples, whether significant models in A are also significant in B, and whether the change in BIC is similar (i.e., > 3) for both samples. We first report the results of our preregistered analyses, followed by the results of the robustness analysis outlined above.

3. Results

Overall, there was considerable variability in model performance, but the best-performing models in each family of models did very well in their ability to reflect human performance in the Gender–Career IAT. Using absolute correlation strength from sample A, 48.24% of all language models lead to significant improvements over the baseline model. Figure 1a summarises the model correlations: mean performance was best for LLMs (Mean *r* = .459, 47.22% significant models, best-model *r* = .659), followed by count vector (Mean *r* = .216, 30.05% significant models, best-model *r* = .535), predict (Mean *r* = .208, 31.38% significant models, best-model *r* = .597), and then n-gram models (Mean *r* = .177, 35.00% significant models, Best-model *r* = .491). In terms of correlation strength, 9 of the top 10 performing models were LLMs (see Figure 2a). However, large variability was also visible, with the best performing n-gram models outperforming more than 66.66% of the large-language model variants.

When we examined model performance using BIC as a measure of model fit, we saw a slightly different picture (Figure 1b). Using BIC change from the baseline model, only one of the top 10 models was LLMs, with the other 9 coming from the predict (Word2vec) family of models, including a mix of CBOW and Skip-gram variants (Figure 2b). From each model family, we found that the best model fit for predict models had a BIC change = 28.41, followed by LLMs (best model BIC change = 27.46), count-vector (best model BIC change = 19.61) and lastly n-gram (best model BIC change = 9.81). Figures 2a and 2b in particular highlight how considering correlation strength or change in BIC reveals different patterns, favouring LLMs in the former and predict models in the latter.

In terms of comparing Euclidean, cosine, correlation and association (from n-gram models) measures, we find that all measures perform reasonably well, but that mean performance for Euclidean (M = 0.269, SD = 0.173) is greater than that of cosine (M = 0.223, SD = 0.168) and correlation (M = 0.225, SD = 0.168), which in turn are greater than association strength of n-gram models (M = 0.177, SD = 0.114). If we consider the top performing models (Tables 3 and 4), we can see that it is dominated by models using Euclidean distance (see Figures 4, 6).

For the three different corpora that were used for the customizable models, we found that models using the Subtitle corpus performed best on average (M = 0.297, SD = 0.138), followed by those using the BNC (M = 0.182, SD = 0.142), and then those using UKWAC (M =0.227, SD = 0.188). It is perhaps surprising to see that the relatively small, but high-quality subtitle corpus outperforms the UKWAC corpus which is an order of magnitude larger (see Figures 4,5,6).

Figure 1a Distribution of correlations between each language model and mean RT in the four model families. Each dark circle represents an individual model instantiation, while the violin outline areas represent the density of correlations in the overall distribution

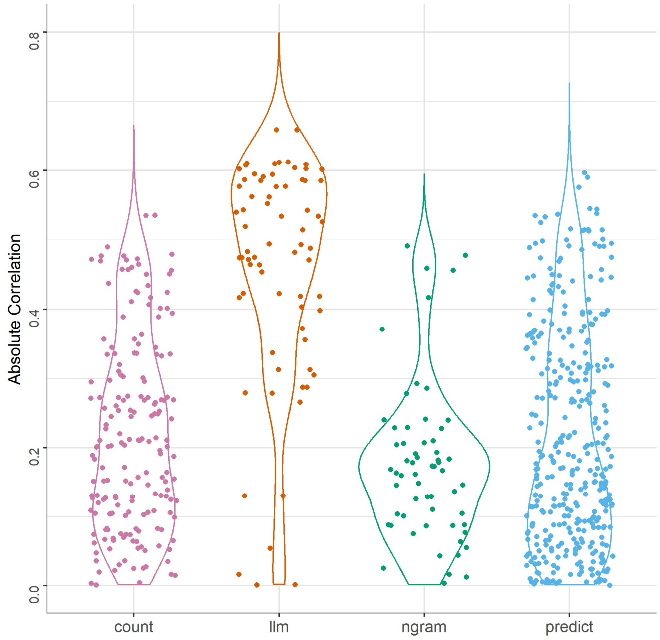


Figure 1b Distribution of BIC change scores between each language model and mean RT in the four model families. Each dark circle represents an individual model instantiation, while the violin outline areas represent the density of BIC change values in the overall distribution. The dashed line indicates a threshold of BIC = 3.

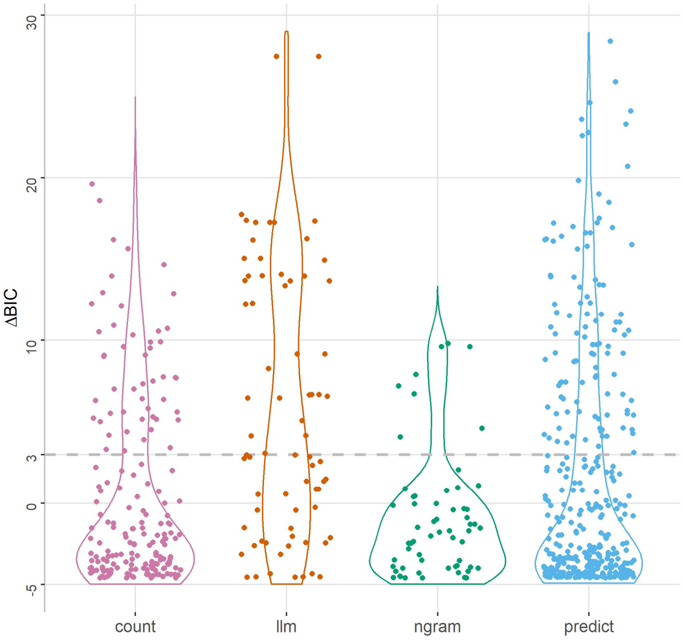
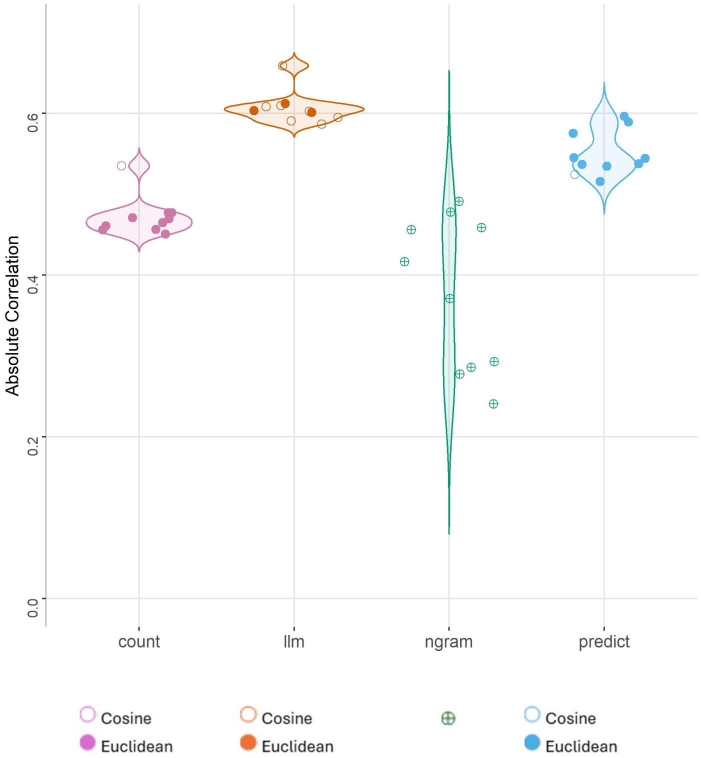


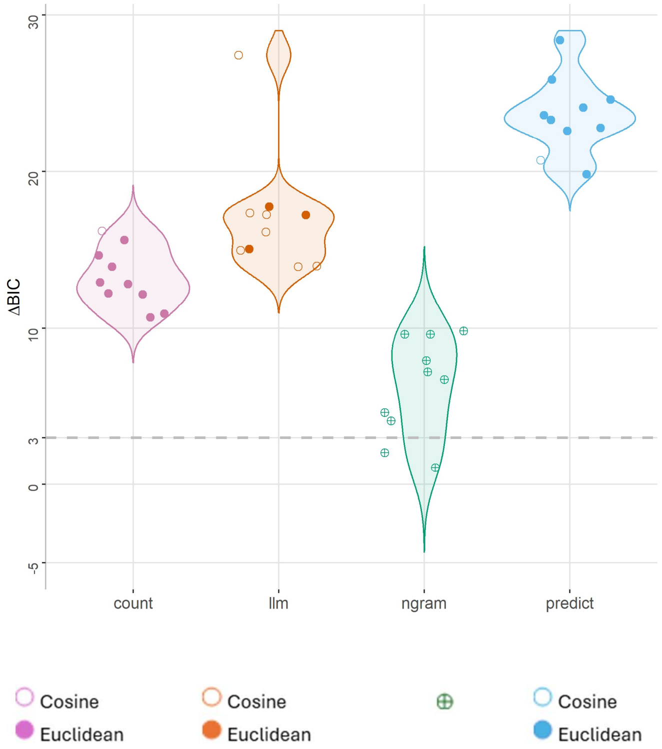
Figure 2a Distribution of correlations between language models and mean RT for the top 10 best performing models in each of the four model families. Each circle represents an individual model instantiation (Euclidean, cosine and ngram measures), while the violin outline areas represent the density of correlations in the overall distribution*



**Correlation distances measures were identical to cosine measures for top 10 models, so we provide here results only for the cosin\e distance (for full results see Supplementary Materials. Table X1)*

For embedding sizes, there was some improvement in average performance for as embedding sizes increased, but not consistently so. In fact, there are individual models with extremely good performance at almost all embedding sizes. For example, the best-performing predict model with embedding sizes of only 100 has a correlation of 0.597 (Table 3).

Figure 2b Distribution of BIC change scores between each language model and mean RT for the top 10 best performing models in each of the four model families. Each circle represents an individual model instantiation (Euclidean, cosine and ngram measures), while the violin outline areas represent the density of BIC change values in the overall distribution*



**BIC changes for correlation and cosine are almost similar, so we provide here results only for the cosine distance (for full results see Supplementary Materials. Table X1). The dashed line indicates a threshold of BIC = 3.*

A similar pattern was evident for context window radius sizes. While some larger radius models did well, there was also wide variability, with some small radius models performing just as well. For example, the best-performing predict models with a radius of 5 and 10 had model performances of 0.576 and 0.597 respectively. Even the best performing predict model with a radius of only 3 had a correlation of 0.545 with the behavioural data.

Table 3 Top ten models ordered by decreasing value of the absolute value of correlation with mean response time. Language model names provide information on some model features (e.g., embedding size, window radius, corpus used), followed the distance measure used (Euclidean, cosine), and whether stimuli were capitalised or lowercase.*

Language Model	Distance	Stimulus Case	Correlation r
Llama_2_7b	Cosine	capitalized	0.659
Llama_7b	Cosine	capitalized	0.612
Llama_65b	Cosine	capitalized	0.61
Mistral_7b_v0.2 (Dolphin 2.8)	Cosine	capitalized	0.608

Llama_3_1_70B (Hermes 3)	Euclidean	capitalized	0.604
Mistral_7b_v0.1	Euclidean	capitalized	0.603
Mistral_7b_v0.3	Cosine	capitalized	0.602
Cbow (embed.size 100, window radius 10, subtitles)	Euclidean	lowercase	0.597
Mistral_7b_v0.1	Cosine	capitalized	0.595
Llama_3.1_70B (Meta)	Euclidean	capitalized	0.591

**Correlation distances measures were identical to cosine measures for top 10 models, so we provide here results only for the cosine distance (for full results see Supplementary Materials. Table X1)*

Table 4 Top ten best-fitted regression models ordered by decreasing BIC change from the baseline regression model to the model containing the individual language model. BIC values of >10 are considered to provide strong evidence in favour of a given model. BIC_A and BIC_B indicate the BIC for subsample A and B respectively, following the addition of the language model. ΔBIC gives the change in BIC from the baseline model for each sample.

	Distance	Case of stimuli	ΔBIC	
Language Model			A	B
Cbow (emb.size 100, window 10, Subtitles)	Euclidean	lower	28.41	28.76
Llama_2_7b	Cosine	capitalized	27.46	27.93
Cbow (emb.size 50, window 10, Subtitles)	Euclidean	lower	25.91	26.16
Skipgram (emb.size 50, window 3, Subtitles)	Euclidean	lower	24.61	24.76
Cbow (emb.size 50, window 5, Subtitles)	Euclidean	lower	24.11	24.06
Skipgram (emb.size 100, window 10, Subtitles)	Euclidean	lower	23.61	23.76
Skipgram (emb.size 50, window 10, Subtitles)	Euclidean	lower	23.31	23.26
Skipgram (emb.size 100, window 5, Subtitles)	Euclidean	lower	22.81	22.96
Skipgram (emb.size 50, window 5, Subtitles)	Euclidean	lower	22.61	22.56
Skipgram (emb.size 50, window 3, Subtitles)	Cosine	lower	20.71	20.66

**Correlation distances measures provide identical BIC to cosine measures for top 10 models, so we provide here only results for cosine distance (for full results see Supplementary Materials. Table X1)*

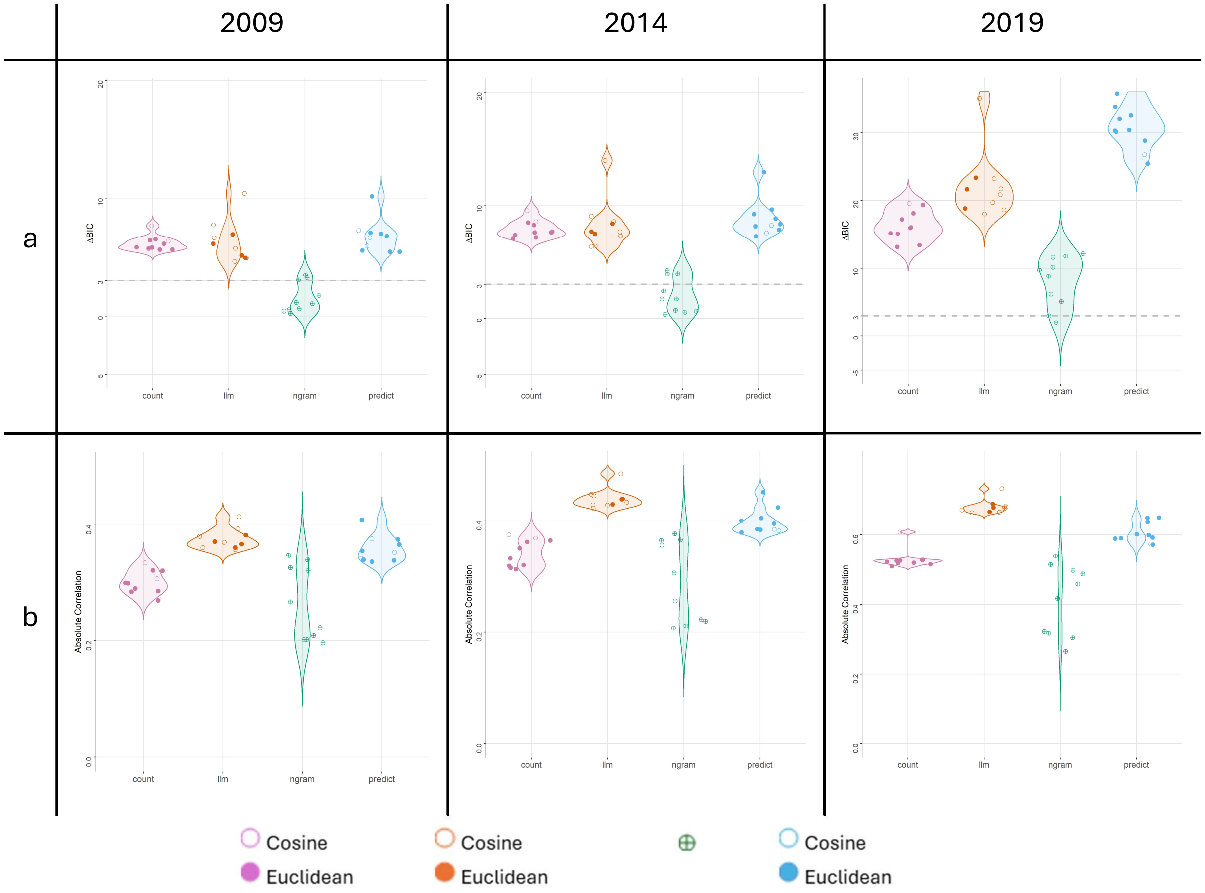
3.1 Robustness Analysis

We compared the results observed for Sample A and Sample B and found very similar overall patterns. We found that the correlation between model performance between the two *r* > .99, with 100% of Sample B correlations being within the 95% confidence intervals for the correlations of Sample A, demonstrating extremely high consistency across both samples. With more conservative checks, some differences emerge across model families. Looking at the consistency of the sign of beta coefficients across samples, almost all models do well (>99%), with the exception of the count probability ratio model (67%) and several n-gram models being less than 95% consistent. Looking at whether significant regression models are consistent across both samples, we find that certain models within each model family perform well, and overall LLMs and predict models are more consistent. Considering the change in BIC in both samples, LLMs show reasonable robustness, with 50% of models showing consistency across both samples, followed by predict (29%), count vector (26%), and lastly n-gram models trailing with 13%. Table 5 summarises the findings of the robustness and consistency checks across the samples.

Table 5 Outcome of robustness check comparisons between sample A and Sample B. The columns indicate if the regression beta-coefficients for the language model showed the same sign for both samples, if the effect of adding the language model is significant in both samples, and if there is a decrease in BIC > 3 for the language model in both samples. The final column indicates the total and percentage of models from each model family that meet all 3 criteria.

		b has same	Model is	decrease in BIC	
family	model	sign	significant	higher than 3	Total
count	Conditional probability	35 (97%)	4 (11%)	2 (6%)	
count	GloVe 42B 300d	3 (100%)	1 (33%)	0 (0%)	
count	Log cooccurrence	36 (100%)	20 (56%)	17 (47%)	
count	PMI	36 (100%)	14 (39%)	12 (33%)	
count	PPMI	36 (100%)	12 (33%)	12 (33%)	
count	Probability ratio	24 (67%)	14 (39%)	6 (17%)	49 (26%)
ngram	Conditional probability ngram	12 (100%)	6 (50%)	4 (33%)	
ngram	Log ngram frequency	12 (100%)	4 (33%)	4 (33%)	
ngram	PMI ngram	12 (100%)	2 (17%)	0 (0%)	
ngram	PPMI ngram	11 (92%)	4 (33%)	0 (0%)	
ngram	Probability ratio ngram	11 (92%)	5 (42%)	0 (0%)	8 (13%)
predict	Cbow	180 (100%)	78 (43%)	50 (28%)	
predict	Skipgram	179 (99%)	76 (42%)	56 (31%)	106 (29%)
llm	llama_30b	6 (100%)	(0%)	(0%)	
llm	llama_65b	6 (100%)	6 (100%)	3 (50%)	
llm	Llama_7b	6 (100%)	6 (100%)	4 (67%)	
llm	llama_2_13b	6 (100%)	5 (83%)	2 (33%)	
llm	llama_2_70b	6 (100%)	3 (50%)	3 (50%)	
llm	llama_2_7b	6 (100%)	6 (100%)	6 (100%)	
llm	Llama_3_1_70B (Hermes 3)	6 (100%)	3 (50%)	3 (50%)	
llm	Llama_3_1_70B (Meta)	6 (100%)	3 (50%)	3 (50%)	
llm	Llama_3_1_8b	6 (100%)	3 (50%)	2 (33%)	
llm	Mistral_7b_v0.1	6 (100%)	6 (100%)	4 (67%)	
llm	Mistral_7b_v0.2 (Dolphin 2.8)	6 (100%)	6 (100%)	5 (83%)	
llm	Mistral_7B_v0.3	6 (100%)	6 (100%)	3 (50%)	36 (50%)

Figure 3. Top 10 models for mean reaction time by year with (a) largest decrease of BIC and (b) largest absolute correlations. We show the top models for count, LLM, ngram, and predict (Word2vec) models.



In analysing robustness across years, Figure 3 shows the performance of the top 10 models for the four model families in terms of BIC change (Panel a) and correlation strength (Panel b) for the years 2009, 2014, and 2019. While of course there are some differences, the overall patterns are remarkably consistent, supporting our preregistered hypothesis. It's also evident that if one were being highly selective, you would be able to select a high performing model from any model family in any given year, so testing a range of models and examining the overall performance gives a much clearer picture of what's going on.

Figure 4. BIC change for baseline model for predict models for mean reaction time calculated on data from 2005-2021. Models are separated by distance measure used and by radius size, and plotted separately for each of the source corpora. We set a cut off of BIC > 3, so only those models that provide a better fit to the data are visible in the figure, emphasising the differences according to the underlying corpus used to train the models, with the Subtitles corpus showing better fit to the data across a range of models. In the key, distance measures are cos (cosine), cor (correlation), and euc (Euclidean), with r_x representing the variants in radius size.

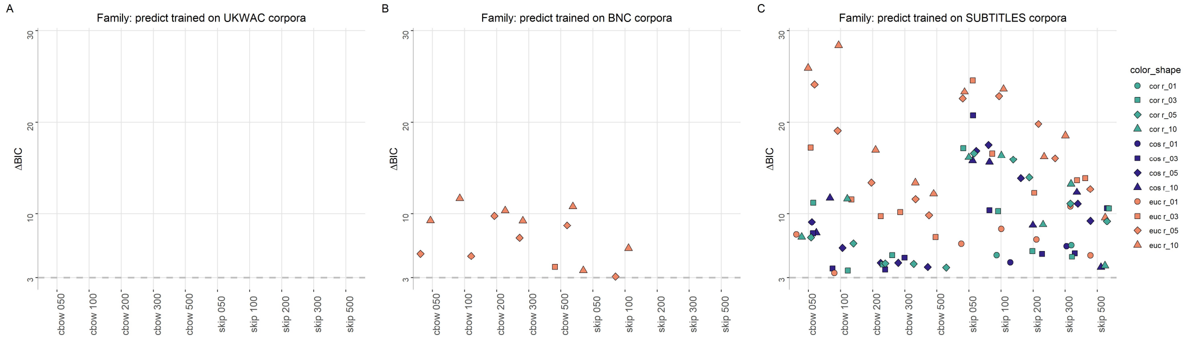


Figure 5. BIC change for baseline model with distance measure for N-gram models for mean reaction time calculated on 2005-2021. We include each of the n-gram model variants, with models further separated by source corpus and by window radius size. As above, we set a cut off of BIC > 3, so only those models that provide a better fit to the data are visible in the figure. This highlights the better performance of conditional probability and log models compared to PMI, PPMI and probability ratio models. In the key, r_x represents the variants in radius size.

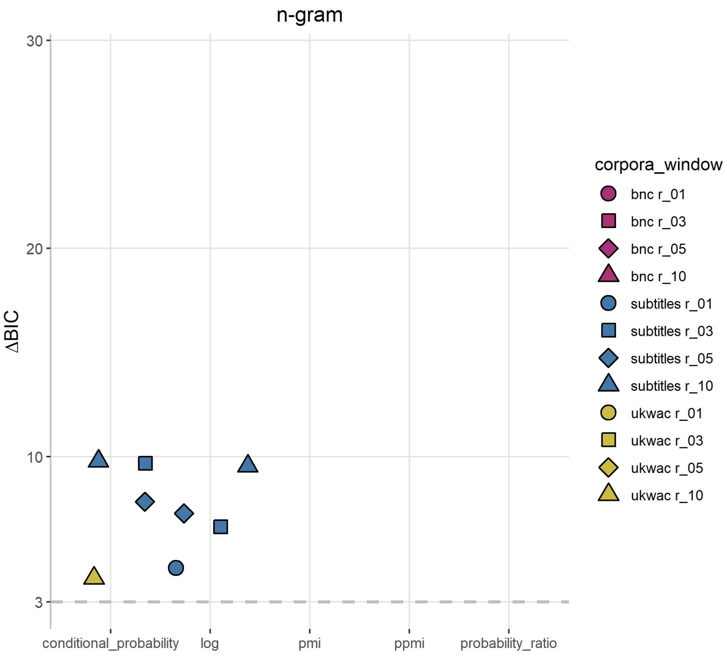


Figure 6. BIC change for baseline model with distance measure from count models for mean reaction time calculated on 2005-2021. We include each of the count model variants, with models further separated by source corpus and by window radius size. As above, we set a cut off of BIC > 3, so only those models that provide a better fit to the data are visible in the figure. This highlights the better performance of models using euclidean distance, and those with a larger window radius. In the key, distance measures are cos (cosine), cor (correlation), and euc (Euclidean), with r_x representing the variants in radius size.

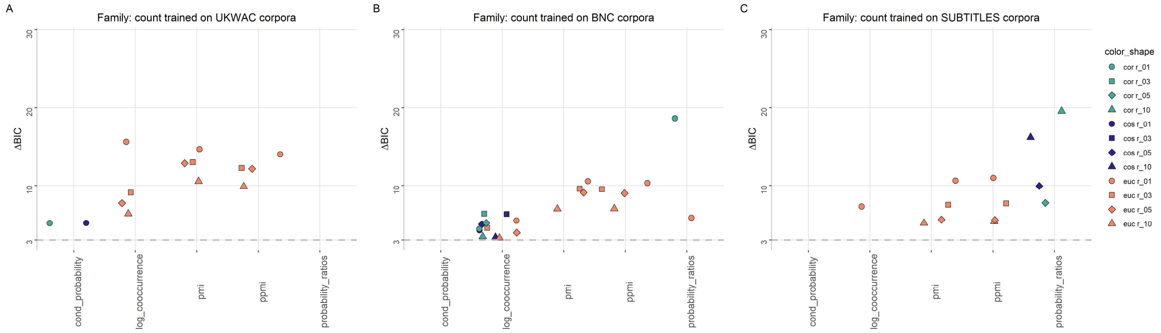
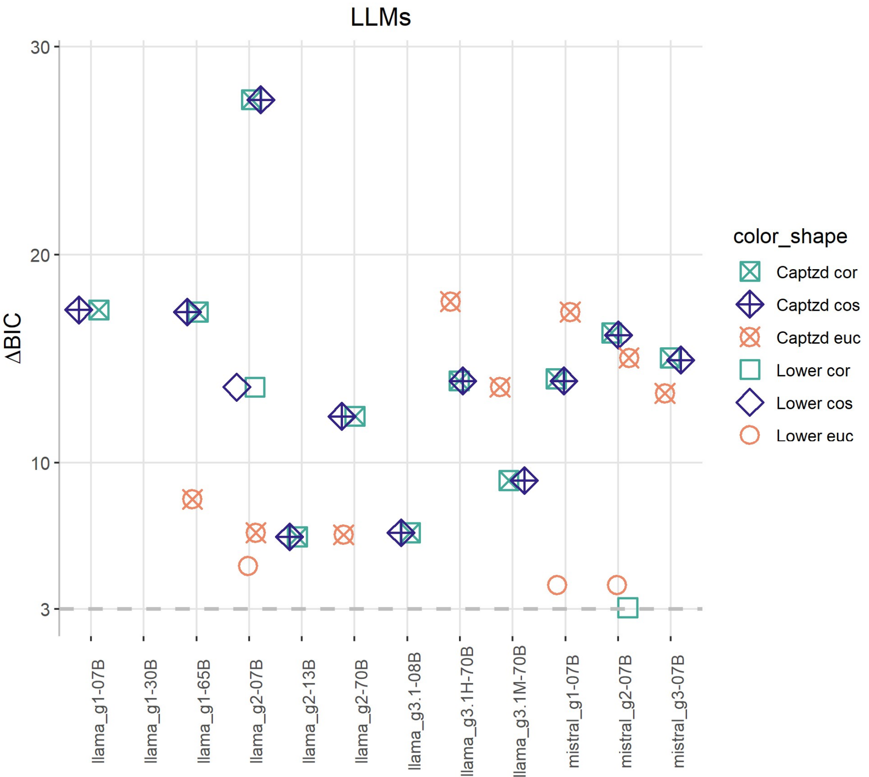


Figure 7. Distance measure from LLMs for mean reaction time calculated on 2005-2021. Models are separated by whether they include uppercase (captzd) or only lowercase (Lower) name variants, and by the distance measure used cos (cosine), cor (correlation), and euc (Euclidean). On the x-axis, models are ordered grouped by creator family (Llama, Mistral), and in order of generation, with earlier generation models to the left. As above, only those models that provide a better fit to the data (BIC > 3) are visible in the figure.



To summarise these findings - in our study preregistration we included three hypotheses regarding the relationship between linguistic features and response times in the Implicit Association Test (IAT), with all three at least partially supported by the results. As expected, there is a set of language models from ngram, predict and count families which demonstrate a strong relationship with reaction time, where the greater the cooccurrence frequency or the smaller the distributional distance between words, the faster the reaction times will be. Furthermore, a large number of models provided more information and better fits to data in comparison with the baseline model, and for the most part, these results are robust across multiple years of data. While these findings aligned with our expectations, some surprising patterns emerged, particularly with respect to LLMs, and these unexpected results are explored in more detail in the following section.

4. Discussion

Language models and linguistic distributional information more generally have previously been used to demonstrate associations between the statistical regularities in language and people's implicit biases. In this paper, we report the findings of a systematic analysis of a large range of language models in modelling human behaviour in a Gender–Career implicit association test. While large language models perform well in some cases, by other measures, they are outperformed by less resource-intensive predict models, such as Word2vec's CBOW and Skip-gram models. Furthermore, in comparison to the recent work by Bhatia and Walasek (2023) who used an off-the-shelf GloVE implementation, the correlations we observed here are considerably stronger, even outperforming models that had been supplemented with additional valence information.

It is worth noting that LLM values may be slightly elevated generally given that these models cannot be customised to the same extent as the other models, meaning we cannot create LLM variants with much smaller embedding sizes or context windows, as we can with the other model families. Results also reveal that high-quality, smaller corpora can outperform larger, but noisier corpora. For example, the predict models trained on relatively small Subtitle corpus (200 million words), resulted in better model performance than UKWAC (2 billion words), and even larger corpora like those used in GloVe (42 billion words) and the LLMs (trillions of words). We observed a similar pattern for the n-gram family (Figure 5), but not for the count-based family where the influence of corpora was less pronounced (Figure 6).

For distance measures, we found that there were good performing models with all measures (Euclidean, cosine and correlation measures), but that the best forming models tended to use Euclidean distance. For predict and count models, Euclidean distance demonstrated better performance in terms of added information relative to the baseline model of mean reaction time (Figures 4 and 6). The situation with distance measures and LLMs is not as straightforward. As shown in Figure 7, for some LLMs, the performance of different distance measures is quite similar; for others, Euclidean distance provides more added value over the baseline, while for others still, cosine or correlation distances perform better. It is worth noting that the distance measures for all investigated LLMs are highly correlated (min = 0.547, mean = 0.833).

While there was a general trend for larger embedding and context window radius sizes to do better, we found that there are models with very strong performance at even the smallest of embedding and radius sizes.

Our robustness analyses found that there was generally very good correspondence with model performance across years (2009, 2014, 2019) and with the A and B samples of the dataset, with very high correlations between observed effects. More conservative measures highlight some differences across model families, with LLMs tending to show more robust patterns across samples.

Thus, although LLMs do well in the current study, despite their massively greater complexity and resource requirements (e.g., Luccioni et al., 2024) (Luccioni et al., 2024), they do not do consistently better than the leaner, more efficient, less resource-intensive predict and even some count-vector models. If one is additionally concerned with the cognitive plausibility of the models being examined, then LLMs are also left wanting in terms of plausible learning mechanisms, training data that is orders of magnitude greater than what people can experience during a lifetime, and they generally lack grounding in broader sensorimotor experience, which is also critical to people's acquisition of semantic knowledge (Connell & Lynott, 2024).

4.1 Some observed differences between classical and large language models

During the process of testing the large range of models, we also observed some surprising patterns in terms of the relationship between model output and the behavioural responses in the IAT. Most models produced results in an intuitive and expected way, with higher model values showing a positive relationship with the behavioural responses (i.e., greater semantic distances corresponding to increased response times in the IAT). However, for the large language and GloVe models, we consistently observed negative correlations. Thus, while the relationships between model and behaviour were often strong, they were in the opposite direction to those observed for most other models. It was initially unclear exactly why this is the case, but we investigated a number of possible explanations.

First, we considered whether observed negative correlations between response times and similarity measures from the language models was something specific to LLMs or not. This was not definitively the case, since some classical models show the same pattern, as do GloVe model variants. Second, the majority of LLMs, including the Llama2 and Mistral models used here, include not only a language-model component, but also a further reinforcement learning from human feedback component (RLHF) that fine-tunes the original language model. This is similar to the approach used with the ChatGPT family of models. One of the aims of the reinforcement learning stage of LLMs is to counteract known biases, such as those associated with race and gender, thereby potentially altering the internal representations and leading to unexpected negative correlations. Again, however, this cannot be the primary reason for the negative correlations because Glove, while being a direct precursor of LLMs, shows the same pattern, but it does not include reinforcement learning from human feedback.

Finally, we considered whether there was something about the nature of the stimuli for the Gender–Career IAT that might be contributing to the effect. For example, in this IAT, there is a mix of both proper names (e.g., Rebecca, John) and common nouns (e.g., business, career). In psychological and psycholinguistic terms, it has been found that people process proper names differently to other words. For example, people often have poorer recall for personal names and proper nouns (e.g., country names) compared to common nouns (e.g., Evrard, 2002 (Evrard, 2002)). In other work, it was found that personal names often have a processing advantage compared to common object names (e.g., Hollis & Valentine, 2001; Peressotti, Cubelli & Job, 2003) . Indeed, when we categorise the stimuli from the current IAT by whether they entail responding to a personal name or a common object, we find that response times to personal names are significantly faster than responses to object names, mirroring this processing advantage from the behavioural literature. Similarly, when we look at the semantic distances extracted from LLMs, we find that LLMs are sensitive to the distinction between the proper names and objects in a way that classical models generally are not. More concretely, for LLMs there is a very strong correlation between semantic distances and whether those distances related to proper names or common objects. Specifically, semantic distances for proper names are significantly lower than those for common objects. In contrast, the direction and strength of this relationship is much reduced or completely absent in other models. On the other hand, with LLMs we can see that using uppercase rather than lowercase versions of proper names (Rebecca vs rebecca) results in higher correspondence with the stimulus category (personal name vs object). That is, uppercase versions of the personal names exhibit consistently smaller semantic distances than lowercase versions of the names. Although, it should be noted that even for lowercase stimuli, some models continue to demonstrate a high correlation with stimulus category (e.g. Llama_2_7b_Q8_0 (cosine distance, lowercase pairs) – 0.82). Nevertheless, despite LLMs generating quite different representations of proper names compared to other models, the contribution of the linguistic information still contributes strongly to accounting for the response time data, as indicated by high correlation and BIC values. Thus, by being sensitive to differences between proper names and common objects, the LLMs can capture more variance in the human behavioural data than many other models. However, the direction of these relationships is not always as intuitive as one might expect, and going forward, it will be important for researchers interested in behavioural modelling in this domain to also be sensitive to stimuli characteristics that may influence model performance.

4.2 Relationship between language models and human biases

A finding that language models can predict human behaviour in IATs demonstrates that cultural biases are reflected in the statistical properties of language. However, the relationship between language models and human biases is complex and nuanced, and it is reasonable to ask, firstly whether this means that language models actually understand biases, and secondly, whether language models simply reproduce existing biases present in their training data or whether they might actively generate biases (Guo & Caliskan, 2021)?

On the first point, we would argue that language models do not understand bias in the way that humans do, but rather what is captured is down to statistical regularities extracted from the training corpora. For example, we can see evidence for this when we consider words that are highly similar, but that might differ in terms of their frequency of use. If language models understood the underlying concept that linked these words, then the words would essentially occupy the same region of semantic space. However, that's usually not the case, with highly similar words often having very different distributions, and therefore providing different semantic distances when compared to other concepts.

On the second point, our sense is that these biases must be present in the training data for the language models to extract them. However, even if language models simply reproduce existing biases, it is possible that the widespread usage of algorithmic systems may also reinforce preexisting biases and inequalities in people (Kordzadeh & and Ghasemaghaei, 2022; Lynott et al., 2019; Noble, 2020; O'Neil, 2017). Because LLMs play an ever-larger role in mediating information, and in the production of new linguistic material, biases in statistical language models risk entering a self-reinforcing cycle with potential real-world consequences (Gallegos et al., 2024; Wilson & Caliskan, 2024).

A further issue is that, with new advances in LLMs, for example, in using synthetic data, reusage of previously created annotations and embeddings (e.g., distillation), application of a system designed for one context to another context (Bender & Friedman, 2018), and the inclusion of AI generated content in text production, may ultimately lead to emergent biases. A specific case might be where a bias, evident in one language is propagated to other languages by being embedded withing a multilingual semantic space (Rogers, 2025). Nonetheless, there is currently limited evidence for such novel biases, although some researchers have identified what they consider intersectional biases in some models, where existing biases are meshed into something new (Guo & Caliskan, 2021).

4.3 Limitations, challenges, and deviations from preregistration

Despite the general trend for good model performance across a range of model families and parameter settings, there are of course important limitations to highlight in the current work. Furthermore, we would like to note some deviations from the original preregistration of this study.

First, although we were able to use a large sample of participant data from Project Implicit (>800K participants, following preprocessing), our focus was only on one specific IAT topic, and therefore only on one set of stimuli. While the current findings are suggestive of the capacity of language models generally to reflect human behavioural biases, our future work will need to consider extending the stimuli and range of topic areas addressed in our modelling work. Given the logic of our approach, and the potential role for linguistic distributional patterns in representing and transmitting bias, we would expect that similar correspondences would emerge between models and human behavior if we considered other forms of bias, such as racial or religious bias. Furthermore, given findings elsewhere (e.g., Caliskan et al., 2017; Lynott et al., 2019), where language models have been applied to different topic areas (e.g., immigration, race), we are hopeful that these findings will extend well to other areas.

It is worth noting that other findings using LLMs have shown some surprising gender biases that also run counter to our expectations. For example, Fulgu and Capraro found examples in ChatGPT of stereotypically masculine roles, such as playing football, being consistently attributed to a female writer (Fulgu & Capraro, 2024). Similarly, ChatGPT agrees with *a woman using violence against a man to prevent a nuclear apocalypse*, but disagrees with *a man using violence against a woman* for the same purpose. The distributed nature of the representations in LLMs make this a challenging issue to unpack, but it is clear that additional work needs to be done in this area. In particular, future work with LLMs should aim to separate the contribution of the linguistic distributional information from that of the reinforcement learning aspects of the training, to be more confident about the language-specific impact on bias formation.

Related to this issue, in the current study, we focus only on the semantic similarity of individual *words* within language models. However, Zhang and colleagues (2020) suggest that focussing on the word level can be problematic and give rise to anomalous results. Considering metrics of gender bias specifically, Zhang et al. (2020) suggest that in addition to semantic similarity, two other factors can be responsible for smaller distances in embeddings space: sociolinguistic factors and mathematical properties of vectors. Zhang et al. argue that cases where there are very high similarity scores between vectors, can make it very difficult to properly evaluate bias. For example, words and their plurals can be assigned opposite bias directions due to vector multiplication, even though they should be considered conceptually in the same way. In our case, we can speculate that in cases of a very high cosine similarity between base words (e.g., in Glove 42b model cosine similarity between "female" and "male" equals to 0.894) smaller distances can be observed by chance, rather than due to gender bias per se. Additionally, we could also expect some sociolinguistic factors to impact the representation of bias within models. For example, we might expect that distance between "family" and "business" will be lower not because of any Gender–Career bias, but because of the relatively high frequency of co-occurrence phrases like "family business". For example, in the GloVe model used here, the cosine distance between "business" and "family" is lower (.632) than that between "business" and "career" (.673). Zhang and colleagues suggest that focusing on the conceptual level (e.g., examining clusters of concepts related to the core concepts of "family" and "career") may give rise to more robust results when considering bias, although the feasibility of this approach may depend on the modelling context.

An important point to consider is that many of the language models we tested here are fully transparent and fully customisable. However, LLMs rarely offer the same levels of transparency, which is problematic for researchers. We overcome this issue somewhat by using open-source models like Llama and Mistral, which for LLMs, offer some of the greatest visibility into their behaviour and underlying representations. However, even with these models we don’t have complete information on constituency and size of the training data used. In an ideal world, researchers would have complete access to all aspects of these models in order to fully and fairly assess their performance.

An additional consideration is the rapid pace at which the field of LLMs continues to develop. Thus, while we include a wide array of models and LLMs with up to 70 billion parameters, even these models could be considered “medium sized” LLMs. While we did not observe dramatic differences in accounting for the behavioral data between smaller and larger LLMs (with the smaller 7 billion parameter LLMs showing the strongest correspondence with the behavioral data) it remains to be seen whether such a pattern of results will continue with an extended set that includes even larger LLMs. However, the current patterns we observe certainly pose challenges for those seeking to develop LLMs that remove biased or stereotyped information. Given that every LLM included in this study shows a correlation with human performance in the implicit association test, it suggests that current attempts using fine-tuning and reinforcement learning with human feedback are not fully capable of addressing this issue. It is possible that focusing on better quality corpora might be a more fruitful approach, rather than trying to remove bias after the fact.

In terms of preregistration deviations, our primary analyses and treatment of the data follow our original plan very closely. However, in our original plan we included a smaller number of classical language models and model families. This was primarily because this work was first proposed more than 2 years ago, meaning that we originally included only n-gram, count vector and predict models. However, given the pace of change in the world of AI and language models, we felt it was important to include newer large language models, as well as GloVe, to provide a more complete picture of this domain.

Finally, it is worth mentioning that "gender" is descriptive of a broad phenomenon which extends beyond simply "male" and "female". Existing gender-based IATs employ a strict male/female binary, which we have therefore followed in the present analysis. Gender-based stereotypes and biases relating to transgender and non-binary identities, and particularly their reflection language, remains an under-studied topic (Hansen & Żółtak, 2022; McCarty & Burt, 2024). In further research additional *contextual information* could be provided related to IAT stimuli to enhance or reduce the focus on gendered elements of the stimulus profile, thereby highlighting gender associations with a particular name, or providing contrasting information to the model. Furthermore, use of larger contexts might provide an opportunity to overcome the binary treatment of gender in relation to the IAT, which would certainly be fruitful avenue for future research.

5. Conclusions

Overall, we find that a range of language models can capture human behavioural performance in relation to Gender–Career implicit biases. While LLMs perform well, their additional resource requirements may not be warranted as they do not reliably outperform much simpler and more cost-effective models.

The current work makes contributions in several respects. In all prior work, efforts to model the IAT and bias have operated at the group or aggregate level. The current analysis goes deeper by implementing modelling at the level of stimuli, which permits more fine-grained analysis, but also allows us to identify effects that are simply not detectable at the aggregate level. For example, operating at the group or aggregate level would mask differential effects for proper names compared to common nouns.

We also extend work comparing data from human behavioral experiments to language models in novel ways. For example, in analysing a large number of models and parameters we can observe patterns that also not visible when using one or two model variants. In particular, we can see that smaller, but well-curated, corpora give rise to closer correspondence with behavioural data (i.e., with the subtitle corpus outperforming the significantly larger UKWAC corpus), and that Euclidean distance may be a better candidate measure than correlation or cosine distance, which are the more traditionally adopted measures. Lastly, by using variants of large language models, we also see great consistency across generations of these models. For example, there is little variation in the correspondence with human data in going from a 7bn parameter model up to a 70bn parameter model. Similarly, we see that that the different representation of proper names and common names exists in almost all large language model variants. Such patterns also challenge assumptions about scale-dependent bias mitigation. In other words, larger models do not necessarily equate to reductions in the presence of bias. Furthermore, by analysing both traditional and large language models, we demonstrate the importance of transparency, where traditional models provide a opportunity for better understanding their inner workings, compared to the black box approach associated with most emerging LLMs.

On a practical level, revealing the relationship between model outputs and actual human behavior highlights limitations of current "bias scores" used in LLM research (Cao et al., 2022; Zhang et al., 2020), and supports calls for multi-method evaluation of such models (Bai et al., 2025). Our study also supports the idea that developers of language agents and models could adopt hybrid metrics, combining static embedding analysis (cost-effective screening) and behavioral probes in bias evaluation (Bai et al. 2025). Lastly, our work speaks to current debates on whether language models are "bias mirrors" or “bias generators", by showing empirically that biases in human behavior have a strong correspondence to those extracted from the models themselves. Such observations provide actionable steps to improve real-world AI systems, since appropriate treatment of training data is likely to be the greatest source of biases that are ultimately observed in deployed models.

In our future research we plan to examine conceptual clusters, rather than isolated word stimuli, consider techniques such as using enhanced contextual information to give a better sense of how bias is captured in language models, and also to continue with fine-grained analysis at the participant level. Another challenge is to understand better the propagation of bias through models, e.g., applying techniques such as layer-wise representation analysis (Pasad et al., 2023). We can expect that this approach could show not only how bias evolves across layers of a model during training, but also how bias trends specifically relate to model performance.

Thus, while current work demonstrates a clear link between people’s behavioral responses and the statistical regularities captured in language models, there remains much work to be done.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland/Research Ireland under Grant number 21/FFP-P/10118.

We would like to express our sincere gratitude to Dr. Vanush Paturyan for his invaluable support and expertise in managing and running the LLMs - we truly appreciate his support and contributions.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine Bias. In Ethics of Data and Analytics. Auerbach Publications.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" Psychological Inquiry, 15(4), 257–278.

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. Proceedings of the National Academy of Sciences, 122(8), e2416228122. https://doi.org/10.1073/pnas.2416228122

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. Behavior Research Methods, 46(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, 43(3), 209–226. https://doi.org/10.1007/s10579-009-9081-4

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech Era. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

Bhatia, S., & Walasek, L. (2023). Predicting implicit attitudes with natural language data. Proceedings of the National Academy of Sciences, 120(25), e2220726120. https://doi.org/10.1073/pnas.2220726120

BNC Consortium. (2007). British National Corpus [Oxford Text Archive]. XML edition. http://hdl.handle.net/20.500.12024/2554.

Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Xiong, B., Maslej, N., & Liang, P. (2024). The Foundation Model Transparency Index v1.1: May 2024.

Bowen III, D. E., Price, S. M., Stein, L. C. D., & Yang, K. (2024). Measuring and Mitigating Racial Bias in Large Language Model Mortgage Underwriting. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4812158

Brennan, T., & Dieterich, W. (2018). Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). In J. P. Singh, D. G. Kroner, J. S. Wormith, S. L. Desmarais, & Z. Hamilton (Eds.), Handbook of Recidivism Risk/Needs Assessment Tools (1st ed., pp. 49–75). Wiley. https://doi.org/10.1002/9781119184256.ch3

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cao, Y., Pruksachatkun, Y., Chang, K.-W., Gupta, R., Kumar, V., Dhamala, J., & Galstyan, A. (2022). On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 561–570. https://doi.org/10.18653/v1/2022.acl-short.62

Chang, R. (2011). Preliminary report on race and Washington's criminal justice system.

Cochrane, A., Cox, W. T. L., & Green, C. S. (2023). Robust within-session modulations of IAT scores may reveal novel dynamics of rapid change. Scientific Reports, 13(1), 16247. https://doi.org/10.1038/s41598-023-43370-w

Connell, L., & Lynott, D. (2014). Principles of Representation: Why You Can't Represent the Same Concept Twice. Topics in Cognitive Science, 6(3), 390–406. https://doi.org/10.1111/tops.12097

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. Journal of Personality and Social Psychology, 81(5), 800–814. https://doi.org/10.1037/0022-3514.81.5.800

Dymarska, A., Connell, L., & Banks, B. (2023). More is not necessarily better: How different aspects of sensorimotor experience affect recognition memory for words. Journal of Experimental Psychology: Learning, Memory, and Cognition, 49(10), 1572–1587. https://doi.org/10.1037/xlm0001265

Evrard, M. (2002). Ageing and Lexical Access to Common and Proper Names in Picture Naming. Brain and Language, 81(1), 174–179. https://doi.org/10.1006/brln.2001.2515

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. Proceedings of the 4th Web as Corpus Workshop.

Fulgu, R. A., & Capraro, V. (2024). Surprising gender biases in GPT.

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

Ghiasi, G., Larivière, V., & Sugimoto, C. R. (2015). On the compliance of women engineers with a gendered scientific system. PLOS ONE, 10(12), e0145931. https://doi.org/10.1371/journal.pone.0145931

Gonzalez, A. M., Steele, J. R., & Baron, A. S. (2017). Reducing Children's Implicit Racial Bias Through Exposure to Positive Out-Group Exemplars. Child Development, 88(1), 123–130. https://doi.org/10.1111/cdev.12582

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. California Law Review, 94(4), 945–967.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. Journal of Personality and Social Psychology, 85(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2024). Why do tree-based models still outperform deep learning on typical tabular data? Proceedings of the 36th International Conference on Neural Information Processing Systems, 507–520.

Grün, B., & Leisch, F. (2023). flexmix: Flexible mixture modeling (Version 2.3-19) [R package]. https://CRAN.R-project.org/package=flexmix

Guo, W., & Caliskan, A. (2021). Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 122–133. https://doi.org/10.1145/3461702.3462536

Hansen, K., & Żółtak, K. (2022). Social perception of non-binary individuals. Archives of Sexual Behavior, 51(4), 2027–2035. https://doi.org/10.1007/s10508-021-02234-y

Hudson, S. K. T. J., Kurdi, B., Lai, C. K., Johnson, J., & Banaji, M. R. (2024). Implicit Attitudes Evoked by a Singular American Slur: Experiments on N***er and N***a in Samples of Black and White Americans. Social Cognition, 42(1), 161–197. https://doi.org/10.1521/soco.2024.42.3.161

Jaff, E., Wu, Y., Zhang, N., & Iqbal, U. (2024). Data exposure from LLM apps: An in-depth investigation of OpenAI's GPTs.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Singh Chaplot, D., de las Casas, D., & Bressand, F. (2023). Mistral 7B.

Kashima, Y., Laham, S. M., Dix, J., Levis, B., Wong, D., & Wheeler, M. (2015). Social transmission of cultural practices and implicit attitudes. Organizational Behavior and Human Decision Processes, 129, 113–125. https://doi.org/10.1016/j.obhdp.2014.05.005

Kordzadeh, N., & and Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Kurdi, B., & Charlesworth, T. E. S. (2023). A 3D framework of implicit attitude change. Trends in Cognitive Sciences, 27(8), 745–758. https://doi.org/10.1016/j.tics.2023.05.009

Kurdi, B., & Dunham, Y. (2020). Propositional Accounts of Implicit Evaluation: Taking Stock and Looking Ahead. Social Cognition, 38(Supplement), s42–s67. https://doi.org/10.1521/soco.2020.38.supp.s42

Lai, C. K., & Wilson, M. E. (2021). Measuring implicit intergroup biases. Social and Personality Psychology Compass, 15(1), e12573. https://doi.org/10.1111/spc3.12573

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

LeBel, E. P., & Paunonen, S. V. (2011). Sexy But Often Unreliable: The Impact of Unreliability on the Replicability of Experimental Findings With Implicit Measures. Personality and Social Psychology Bulletin, 37(4), 570–583. https://doi.org/10.1177/0146167211400619

Luccioni, S., Gamazaychikov, B., Hooker, S., Pierrard, R., Strubell, E., Jernite, Y., & Wu, C. J. (2024). Light bulbs have energy ratings—So why can't AI chat-bots? Nature, 632(8026), 736–738.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. Behavior Research Methods, 45(2), 516–526. https://doi.org/10.3758/s13428-012-0267-0

Lynott, D., Kansal, H., Connell, L., & O'Brien, K. (2012). Modelling the IAT: Implicit Association Test Reflects Shallow Linguistic Environment and not Deep Personal Attitudes. Proceedings of the Annual Meeting of the Cognitive Science Society, 34(34). https://escholarship.org/uc/item/5fj441tg

Lynott, D., Walsh, M., McEnery, T., Connell, L., Cross, L., & O'Brien, K. (2019). Are You What You Read? Predicting Implicit Attitudes to Immigration Based on Linguistic Distributional Cues From Newspaper Readership; A Pre-registered Study. Frontiers in Psychology, 10, 842. https://doi.org/10.3389/fpsyg.2019.00842

McCarty, M. K., & Burt, A. H. (2024). Understanding perceptions of gender non-binary people: Consensual and unique stereotypes and prejudice. Sex Roles, 90(3), 392–416. https://doi.org/10.1007/s11199-024-01449-2

McConnell, A. R., Rydell, R. J., Leibold, J. M., Hugenberg, K., & Czopp, A. M. (2018). Adjusting implicit measures of racial prejudice: Insights from research on validity and change. Social Cognition, 36(4), 373–399. https://doi.org/10.1521/soco.2018.36.4.373

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1301.3781

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. Proceedings of the National Academy of Sciences, 109(41), 16474–16479.

Noble, S. U. (2020). Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. In Automatic processes in social thinking and behavior (J.A.Bargh (Ed.), pp. 265–292). Psychology Press.

Nosek, B. A., Greenwald, A. G., Mahzarin R. Banaji, Lai, C. K., Axt, J., Ratliff, K., Smith, C., Bar-Anan, Y., O'Shea, B., Lofaro, N., Umansky, E., Simon, L., Xu, F. K., & Frost, N. (2015). Project Implicit Demo Website Datasets. https://doi.org/10.17605/OSF.IO/Y9HIQ

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2012). Implicit social cognition. In S. Fiske & C. Macrae (Eds.), Handbook of Social Cognition (pp. 315–353).

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. European Review of Social Psychology, 18(1), 36–88.

O'Boyle, E., & Harter, J. (2013). State of the American workplace: Employee engagement insights for U.S. business leaders. Gallup.

O'Brien, K. S., Latner, J. D., Ebneter, D., & Hunter, J. A. (2013). Obesity discrimination: The role of physical appearance, personal ideology, and anti-fat prejudice. International Journal of Obesity, 37(3), 455.

Omar, M., Soffer, S., Agbareia, R., Bragazzi, N. L., Apakama, D. U., Horowitz, C. R., Charney, A. W., Freeman, R., Kummer, B., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2025). Sociodemographic biases in medical decision making by large language models. Nature Medicine, 1–9. https://doi.org/10.1038/s41591-025-03626-6

O'Neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy (First paperback edition). B/D/W/Y Broadway Books.

Onnis, L., & Lim, A. (2024). Distributed semantic representations of inanimate nouns are gender biased in gendered languages. Proceedings of the Annual Meeting of the Cognitive Science Society, 46. https://escholarship.org/uc/item/50m8883c

Pasad, A., Shi, B., & Livescu, K. (2023). Comparative Layer-Wise Analysis of Self-Supervised Speech Models. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096149

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. Sociological Methodology, 25, 111. https://doi.org/10.2307/271063

ROGERS, R. (2025, APRIL 23). AI IS SPREADING OLD STEREOTYPES TO NEW LANGUAGES AND CULTURES. WIRED. HTTPS://WWW.WIRED.COM/STORY/AI-BIAS-SPREADING-STEREOTYPES-ACROSS-LANGUAGES-AND-CULTURES-MARGARET-MITCHELL/

RÖHNER, J., & LAI, C. K. (2021). A DIFFUSION MODEL APPROACH FOR UNDERSTANDING THE IMPACT OF 17 INTERVENTIONS ON THE RACE IMPLICIT ASSOCIATION TEST. PERSONALITY AND SOCIAL PSYCHOLOGY BULLETIN, 47(9), 1374–1389. HTTPS://DOI.ORG/10.1177/0146167220974489

RUDMAN, L. A. (2004). SOURCES OF IMPLICIT ATTITUDES. CURRENT DIRECTIONS IN PSYCHOLOGICAL SCIENCE, 13(2), 79–82. HTTPS://DOI.ORG/10.1111/J.0963-7214.2004.00279.X

RUDMAN, L. A., PHELAN, J. E., & HEPPEN, J. B. (2007). DEVELOPMENTAL SOURCES OF IMPLICIT ATTITUDES. PERSONALITY AND SOCIAL PSYCHOLOGY BULLETIN, 33(12), 1700–1713. HTTPS://DOI.ORG/10.1177/0146167207307487

SCHIMMACK, U. (2021). THE IMPLICIT ASSOCIATION TEST: A METHOD IN SEARCH OF A CONSTRUCT. PERSPECTIVES ON PSYCHOLOGICAL SCIENCE, 16(2), 396–414. HTTPS://DOI.ORG/10.1177/1745691619863798

STAATS, C. (2016). UNDERSTANDING IMPLICIT BIAS: WHAT EDUCATORS SHOULD KNOW. AMERICAN EDUCATOR, 39(4), 29.

TAKSHI, S. (2020). UNEXPECTED INEQUALITY: DISPARATE-IMPACT FROM ARTIFICIAL INTELLIGENCE IN HEALTHCARE DECISIONS. JOURNAL OF LAW AND HEALTH, 34(2), 215–251.

TOUVRON, H., MARTIN, L., STONE, K., ALBERT, P., ALMAHAIRI, A., BABAEI, Y., & SCIALOM, T. (2023). LLAMA 2: OPEN FOUNDATION AND FINE-TUNED CHAT MODELS. ARXIV PREPRINT.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. Quarterly Journal of Experimental Psychology, 67(6), 1176–1190.

Wachter, S. M., & Megbolugbe, I. F. (1992). Impacts of housing and mortgage market discrimination racial and ethnic disparities in homeownership. Housing Policy Debate, 3(2), 332–370. https://doi.org/10.1080/10511482.1992.9521099

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45(4), 1191–1207.

Wilson, K., & Caliskan, A. (2024). Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7, 1578–1590. https://doi.org/10.1609/aies.v7i1.31748

Wingfield, C., & Connell, L. (2022). Understanding the role of linguistic distributional knowledge in cognition. Language, Cognition and Neuroscience, 37(10), 1220–1270. https://doi.org/10.1080/23273798.2022.2069278

Yasui, M. (2015). A review of the empirical assessment of processes in ethnic–racial socialization: Examining methodological advances and future areas of development. Developmental Review, 37, 1–40. https://doi.org/10.1016/j.dr.2015.03.001

Zhang, H., Sneyd, A., & Stevenson, M. (2020). Robustness and Reliability of Gender Bias Assessment in Word Embeddings: The Role of Base Pairs. In K.-F. Wong, K. Knight, & H. Wu (Eds.), Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (pp. 759–769). Association for Computational Linguistics. https://aclanthology.org/2020.aacl-main.76

Zou, L., & Khern-am-nuai, W. (2023). AI and housing discrimination: The case of mortgage applications. AI and Ethics, 3(4), 1271–1281. https://doi.org/10.1007/s43681-022-00234-9