Modeling Implicit Attitudes with Natural Language Data: A Comparison of Language Models

Alexander Porshnev[1,*], Kevin Dirk Kiy [1], Diarmuid O'Donoghue[1], Manokamna Singh[1], Cai Wingfield[2] and Dermot Lynott[1,*]

[1] *Maynooth University, Maynooth, Co Kildare, Ireland*

[2] *The University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom*

Abstract

Prejudicial attitudes, such as those relating to age, race, or gender, exert a powerful influence on individuals, and are pervasive throughout society. Recent research suggests that the statistical patterns of how words are used in language may capture such biases, with language models providing approximations for people's linguistic experience. 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 (where men are routinely favoured over women in the workplace) 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, and Large Language Models (LLMs), to determine how well they capture people's behaviour in the IAT. We examined data from over 800,000 participants, tested against over 600 language model variants. While we find that LLMs perform well in modelling IAT responses, they are not significantly better than simpler count vector and predict models, with these other models actually providing better fits to the behavioural data using Bayesian estimates. Our findings suggest that societal biases may be encoded in language, but that resource-greedy large language models are not necessary for their detection.

Keywords

Linguistic distributional models, Computational modelling, Large language models, Implicit association test, Bias

1. Introduction

Janet is less likely to be called for an interview than John. Jamal is less likely to get a job than James, and obese applicant Julie is less likely to be shortlisted than her perceived healthy colleague Joan. Implicit biases (ones that we are not consciously aware of) are pervasive in our society [1] and can lead directly to prejudicial decision-making [e.g., 2–4]. Biases linked to gender, race, perceived health status, and many other characteristics are seen in employment, education, criminal justice, politics, and healthcare [5]. 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 [3]. Similarly, while women engineers publish in journals with higher Impact Factors than their male peers, they receive fewer citations from the scientific community [6]. 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 cost of up to $550 billion in the US alone [7]. Yet despite the acknowledged prevalence of such biases, we still do not fully understand where these biases come from or how they are transmitted [8].

However, we know that language is a primary form of cultural transmission [9,10], making it a plausible candidate for the communication and entrenchment of society's implicit biases. Furthermore, recent work in computational modelling suggests that implicit biases may be captured by the latent statistical patterns in language [11–14]. In other words, the way words appear together in language may influence our unconscious, and potentially prejudicial, attitudes towards others.

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09—10, 2024, Dublin, Ireland

[*] Corresponding author.

alexander.porshnev@mu.ie (A. Porshnev); dermot.lynott@mu.ie (D.Lynott);

0000-0002-0075-1061 (A. Porshnev); 0009-0009-1771-533X (K.D.Kiy); 0000-0002-3680-4217 (D.O'Donoghue), 0000-0002-0187-3597 (M.Singh), 0000-0002-0254-199X (C.Wingfield), 0000-0001-7338-0567 (D.Lynott)

However, we still do not have a good handle on whether these patterns can be adequately captured by existing language models.

Recent work has demonstrated that statistical distributional properties of words reflect human biases and prejudicial judgements [11–13]. The way positive and negative terms are distributed in language closely reflects the positive or negative biases people exhibit towards various concepts, as measured by implicit association tests (IAT), for example. The IAT is a computerised task, with strong internal consistency and test–retest reliability [8], and is the most commonly used measure of people's automatic associations between concepts, used in thousands of studies [e.g., 5]. In an IAT, participants classify stimuli into categories as quickly as possible, where faster responses indicate stronger associations between concepts [15]. 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., 8].

Investigating how language models capture human biases measured by Implicit Association Tests (IATs) Lynott et al. [12] used n-gram co-occurrence counts from a corpus of over 1 trillion words to show that stereotypical Black names (e.g., Jamal) co-occur with more negative attributes than White names (e.g., Brad), correlating strongly with implicit biases (r = 0.79) found in IATs. In the work of Caliskan et al. [11] word embeddings like GloVe were used to capture biases such as preferences for flowers over insects and gender imbalances in professions. In a recent paper, Bhatia and Walasek [16] go further, using large samples of human participants, which provides the much-needed statistical power for these kinds of analyses, and directly links human behaviour in IAT tasks with cosine similarity between stimuli words in language models.

While these studies highlight the potential of language models in bias detection, limitations include generally small sample sizes and reliance on a small number of language models. Zhang et al. [17] have raised concerns about the robustness of word embedding-based bias detection, noting that results may depend heavily on model parameters. Thus, it remains unclear whether language models generally can reliably capture biases, or if only a select few perform well under specific conditions.

To address this issue, in the current study we conduct an analysis of a range of language models, parameters and distance measures to determine a) what types of language models best capture behavioural responses to gender–career concepts as measured by the implicit association test, and b) what distance measures show the best relationship between language models and human behaviour.

While much work in AI, often 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 [18,19]. Thus, a priori we might expect LLMs to perform very well, but special efforts to remove biases by the developers can make LLMs less suitable for this task, with greater computational complexity not buying 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 [20], possible exposure of sensitive information [21], 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. Linguistic Distributional Models

We examined four families of Linguistic Distributional Models (LDMs) that ranged considerably in their complexity: n-gram, count vector, predict, and large language models. The first three families have been used extensively in previous research, particularly in cognitive and psycholinguistic work [e.g., 18,22–24], which has found that larger models do not necessarily lead to better performance in modelling cognitive tasks [e.g., 18]. We also included three well-known large language models from Meta and MistralAI [25,26], which were primarily developed with a focus on text generation. A more detailed description for many of these models, including calculations for each measure, can be found in Wingfield and Connell [18], Pennington et al. [24], Touvron et al.[25] and Jiang et al. [26].

Table 1
Summary of all 663 models, including variants by corpus, window radius, distance measures, and embedding size. Custom models are those where various parameters have been manipulated by the researchers, while "open" 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	Number of LDMs
		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
		Open models		
count[b]	GloVE	Global 300		3
llm[b]	llama-2-7b-chat.Q4_K_M	4096	4096	18
llm[b]	cor_llama-2-13b-chat.Q5_K_S	4096	5120	18
llm[b]	mistral-7b-v0.1.Q4_K_M	4096	4096	18

[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*

In this study we examine two classes of models: "custom" and "open". The "custom" trained models – n-gram, count vector (excluding GloVe), and predict – and "open" 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 [27], British National Corpus – BNC [28], and the BBC Subtitle Corpus; details provided below), context window sizes (1, 3, 5, 10), and embedding sizes (for prediction models: 50, 100, 200, 300, 500). For "open" models we have no control over training procedures.

To measure word distances we used Conditional probability, Log n-gram frequency, PMI n-gram, PPMI n-gram, Probability ratio n-gram for n-gram models and vector metrics (Euclidean distance, Cosine distance, Correlation distance) for count, predict and LLM models (for a summary of "custom" and "open" models used in analysis see Table 1).

2.2. Behavioral data and evaluating model performance

We collated human behavioral data from the Project Implicit Gender & Career IAT study, obtained from an Open Science Framework repository [29]. This data includes raw data per trial, including response times (RTs) for stimuli categorization in 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 corpora used to generate the LDMs. We selected participants who were 18 or older and participated in the common version of the IAT (without additional stimuli), those who did not have prior experience with IAT tasks [30,31], for whom raw trial data is available, and where the calculated bias effect size D was found to be equal to the effect size in the preprocessed dataset (to avoid discrepancies between raw and preprocessed data). Summary of data preprocessing presented in Supplementary materials, Table A.

After data cleaning, the final dataset comprised 802,070 participants from the USA and UK, spanning the period from 2005 to 2021. Descriptive statistics presented in Supplementary materials, Table B-E. Next, we randomly split the whole sample into two subsamples (Sample A: 401,025, Sample B: 401,045) balanced by country and year. 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).

2.2.1. Preregistration and analysis

We preregistered our approach to data handling and our planned analyses (https://aspredicted.org/ZWZ_9RV).

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 [32], but which are not of theoretical importance in this case. Subsequently, we add a single LDM (e.g., word distances for n-gram models and vector distances for other families, including LLMs) to determine if this model improves fit over the baseline model, using $p < .05$ as a threshold for significance. For each LDM, 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. Because the sign of correlations indicating a good fit between model values and behavioural data differed between models [18], we report absolute Pearson's correlation values for ease of cross-comparison. BIC is a useful additional measure as it allows us to quantify the strength of evidence for and against each model being considered [33]. For example, values of $BF_{10} > 3$ are seen as providing positive support for the alternative hypothesis (models containing the LDM feature), values between 0.33 and 3 are in the anecdotal range, while values < 0.33 are seen as providing better support for the null hypothesis (i.e., in favour of the baseline model – see Jeffreys, 1961). We completed this analysis for all models. We ran all regression models in R [34] and flexmix [35], as well as multiple helper packages [36–41].

Thus, regression analysis allows us to find language models for which distances for the stimuli words (or word similarity for n-gram models) provide additional information to the baseline model for the RT of participants (for the same stimuli words). This allows us to investigate the link between participants behavior and semantic distances in language model spaces, and to contribute to the discussion how statistical distributional properties of words reflect human biases and prejudicial judgements.

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 correlation strength, 42.5% of all language models lead to significant improvements over the baseline model. Figure 1.a) summarises the model correlations: mean performance was best for LLMs (Mean r = .495, 81.5% significant models, best-model r = .619), followed by count vector (Mean r = .219, 42.5% significant models, best-model r = .535), predict (Mean r = .209, 42.8% significant models, best-model r = .601), and then n-gram models (Mean r = .179, 33.3% significant models, Best-model r = .492), with 9 of the top 10 performing models being LLMs (see Figure 1.b and Supplementary materials, Table F).

When we examined model performance using BIC as a measure of model fit, we saw a slightly different picture (Figure 1.c). Using BIC change from the baseline model, only 2 of the top 10 models were LLMs, with the other 8 coming from the predict family of models, including a mix of CBOW and Skip-gram variants (Figure 1.d and Supplementary materials, Table G). From each model family, we found that the best model fit for predict models had a BIC change = 28.92, followed by LLMs (best model BIC change = 18.8), count-vector (best model BIC change = 18.6) and lastly n-gram (best model BIC change = 9.92). Figures 1.b and 1.d 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 = 0.263, SD = 0.166) is greater than that of cosine (M = 0.226, SD = 0.165) and correlation (M = 0.264, SD = 0.165), which in turn are greater than association strength of n-gram models (M = 0.179, SD = 0.114). If we consider the top performing models (Supplementary materials, Tables F and G), we can see that it is dominated by models using Euclidean distance. The domination of the Euclidean distance was not expected, as in contemporary research this metric sometimes even not regarded (e.g. [16]). Thus, the question of the best distance remains open and need further investigation

For the three different corpora that were used for the customizable models, we found that models using the BBC Subtitle corpus performed best on average (M = 0.314, SD = 0.146), followed by those using the BNC (M = 0.209, SD = 0.163), and then those using UKWAC (M =0.175, SD = 0.146). 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.

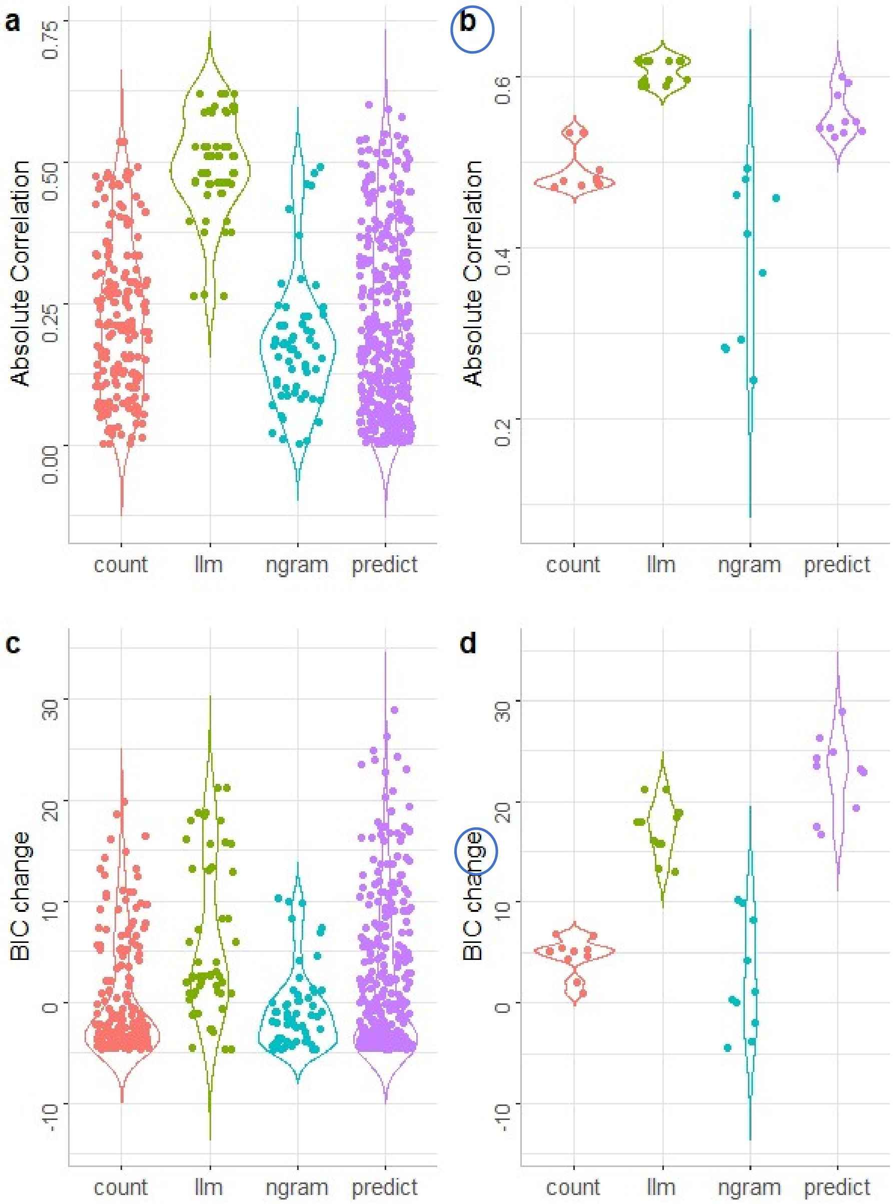
For embedding sizes, there is a notable increase in average performance for larger embeddings, with the strongest performance for LLMs with embedding sizes of 4,096 (M = 0.509, SD = 0.103) and 5120 (M = 0.467, SD = 0.048). However, these mean values mask the fact that there are individual models with extremely good performance at almost all embedding sizes. For example, the best-performing predict models with embedding sizes of only 50 and 100 have model performance of 0.593 and 0.601 respectively.

Similarly, for context window radius size, while larger radiuses perform better on average (e.g., radius 32768, M = 0.516, SD = 0.081; radius 4096 M = 0.485, SD = 0.094), the best-performing predict models with a radius of 5 and 10 had model performances of 0.578 and 0.601 respectively. Even the best performing predict model with a radius of only 1 had a correlation of 0.535 with the behavioural data.

We compared the patterns observed in 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

Figure 1

Distribution of correlations (a,b) and BIC change (c,d) between each LDM and mean RT in the four model families for all models (a,c) and top ten models (b,d). Each dark circle represents an individual model instantiation, while the violin outline areas represent the density of correlations in the overall distribution



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% consistent), 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 43% of models showing consistency across both samples, followed by predict (29%), count vector (26%), and lastly n-gram models trailing with 13%. All results for the models can be found in supplementary materials Table H.

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, higher BIC change are observed for less resource-intensive predict models, such as CBOW and Skip-gram models (in the top 10 models by decrease in BIC on sample A, 3 are CBOW and 5 are Skip-gram models, see supplementary materials Table H). 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 corpora can outperform larger, but noisier corpora. For example, the 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 (6 billion words) and the LLMs (trillions of words). 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. 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 the A and B samples of the dataset, with very high correlations between observed effects, although more conservative measures highlight some differences across model families.

What do these patterns suggest? The patterns certainly go counter to a "bigger is better" intuition people often have about language models, and the seemingly constant drive to produce larger and larger models with higher and higher resource requirements. The findings suggest that appropriately tailored non-LLMs can perform as well as, if not better than, LLMs in certain cases. This pattern reflects recent findings indicating that larger LLMs may actually be less reliable than smaller language models [42]. Given that even the simplest of language models can give rise to good model performance in capturing behaviour in Gender–Career implicit association test, it also suggests that biases are encoded very generally in linguistic information, and therefore do not specifically require LLMs for them to be uncovered. Thus, although LLMs do well in the current study, despite their massively greater complexity and resource requirements [e.g., 20], they do not do significantly 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 general lack of grounding in broader sensorimotor experience, which is also critical to people's acquisition of semantic knowledge [22].

A finding that language models can predict human behaviour in IATs demonstrates that cultural biases are reflected in the statistical properties of language, however it is also the case that biases in language – the medium through which much information about the world is obtained – reinforce psychological biases in humans [13]. With LLMs playing an ever-larger role in mediating information, and in the production of new linguistic material, biases in statistical language models risk entering a vicious, self-reinforcing cycle with potential real-world consequences [43].

4.1.1. 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 minor 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 findings elsewhere [e.g., 11,13], we are hopeful that these findings will extend well to other areas of implicit bias.

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. 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 is unclear exactly why this is the case, but there are a number of possible explanations. First, 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 (RLHF) component that fine-tunes the original language model. This is similar to the approach used with the ChatGPT family of models. However, it remains unclear what the specific impact is of this reinforcement learning stage on the linguistic representations within the 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.

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 [17] 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. 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 [17]. Regarding 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 6b 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 [17].

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 publicly available models like Llama2 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 to fully and fairly assess their performance. In addition, the field of LLMs continues to develop rapidly, and LLMs included in this study are now classified by researchers as S(mall)LLMs. Furthermore, in this study we used heavily quantized (Q4) models potentially reducing precision. Thus, results of this study should be regarded with care and for better generalization of fundings more research with an extended set of models is needed.

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 models and model families. This was primarily because this work was originally 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, and so added the Llama 2, Mistral, and GloVe models to provide a more complete picture of language models in 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 [44,45].

4.1.2. Conclusion

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.

Acknowledgements

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

The authors declare that they have no conflict of interest.

Our data, analysis scripts, supplementary materials and figures have been deposited in the project's page on the Open Science Framework:
https://osf.io/n36gr/?view_only=f385b438c2ac430b8f5e2ddc5fc87c0d

References

[1] C. Staats, Understanding Implicit Bias: What Educators Should Know, American Educator 39 (2016) 29.

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

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

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

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

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

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

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

[9] M.A.K. Halliday, Explorations in the functions of language, Arnold, 1973.

[10] S. Kirby, H. Cornish, K. Smith, Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language, Proceedings of the National Academy of Sciences 105 (2008) 10681–10686.

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

[12] D. Lynott, H. Kansal, L. Connell, K. O'Brien, Modelling the IAT: Implicit Association Test reflects shallow linguistic environment and not deep personal attitudes, in: Proceedings of the Annual Meeting of the Cognitive Science Society, 2012.

[13] D. Lynott, M. Walsh, T. McEnery, L. Connell, L. Cross, K. O'Brien, 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 (2019) 842. https://doi.org/10.3389/fpsyg.2019.00842.

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

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

[17] H. Zhang, A. Sneyd, M. Stevenson, 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, Association for Computational Linguistics, Suzhou, China, 2020: pp. 759–769. https://aclanthology.org/2020.aacl-main.76 (accessed October 2, 2024).

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

[19] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2024: pp. 507–520.

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

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

[22] L. Connell, D. Lynott, What Can Language Models Tell Us About Human Cognition?, Current Directions in Psychological Science 33 (2024) 181–189. https://doi.org/10.1177/09637214241242746.

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

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

[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, arXiv Preprint (2023).

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

[27] A. Ferraresi, E. Zanchetta, M. Baroni, S. Bernardini, Introducing and evaluating UKWaC, a very large web-derived corpus of English, in: Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can We Beat Google, 2008: pp. 47–54.

[28] BNC Consortium, British National Corpus, (2007). http://hdl.handle.net/20.500.12024/2554.

[29] F.K. Xu, N. Lofaro, B.A. Nosek, A.G. Greenwald, J. Axt, L. Simon, N. Frost, Gender-Career IAT 2005-2023, (2024). https://osf.io/abxq7/.

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

[31] J. Röhner, C.K. Lai, A diffusion model approach for understanding the impact of 17 interventions on the race implicit association test, Personality and Social Psychology Bulletin 47 (2021) 1374–1389. https://doi.org/10.1177/0146167220974489.

[32] A. Dymarska, L. Connell, B. Banks, Weaker than you might imagine: Determining imageability effects on word recognition, Journal of Memory and Language 129 (2023) 104398. https://doi.org/10.1016/j.jml.2022.104398.

[33] Z. Dienes, N. Mclatchie, Four reasons to prefer Bayesian analyses over significance testing, Psychonomic Bulletin & Review 25 (2018) 207–218.

[34] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, 2023. https://www.R-project.org/.

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

[36] H. Wickham, R. François, L. Henry, K. Müller, D. Vaughan, dplyr: A grammar of data manipulation (Version 1.1.4), GitHub, 2023. https://github.com/tidyverse/dplyr; https://dplyr.tidyverse.org.

[37] T. Barrett, M. Dowle, A. Srinivasan, J. Gorecki, M. Chirico, T. Hocking, B. Schwendinger, data.table: Extension of data.frame [R package], (2006). https://doi.org/10.32614/CRAN.package.data.table.

[38] D. Lüdecke, M. Ben-Shachar, I. Patil, P. Waggoner, D. Makowski, Performance: An R package for assessment, comparison, and testing of statistical models, Journal of Open Source Software 6 (2021) 3139. https://doi.org/10.21105/joss.03139.

[39] D. Lüdecke, sjPlot: Data visualization for statistics in social science (Version 2.8.16) [R package], (2024). https://CRAN.R-project.org/package=sjPlot.

[40] E.F. Haghish, md.log: Produces markdown log file with a built-in function call [R package], (2017). https://doi.org/10.32614/CRAN.package.md.log.

[41] W. Revelle, psych: Procedures for psychological, psychometric, and personality research (Version 2.4.6), Northwestern University, 2024. https://CRAN.R-project.org/package=psych.

[42] Y. Zhou, P. Xu, X. Liu, B. An, W. Ai, F. Huang, Explore Spurious Correlations at the Concept Level in Language Models for Text Classification, (2024). http://arxiv.org/abs/2311.08648 (accessed September 19, 2024).

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

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

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