

Profiling Bias in LLMs: Stereotype Dimensions in Contextual Word Embeddings

Carolin M. Schuster, Maria-Alexandra Dinisor, Shashwat Ghatiwala and Georg Groh

TUM School of Computation, Information and Technology

Technical University of Munich

{carolin.schuster, alexandra.dinisor, shashwat.ghatiwala}@tum.de
grohg@in.tum.de

Abstract

Large language models (LLMs) are the foundation of the current successes of artificial intelligence (AI), however, they are unavoidably biased. To effectively communicate the risks and encourage mitigation efforts these models need adequate and intuitive descriptions of their discriminatory properties, appropriate for all audiences of AI. We suggest bias profiles with respect to stereotype dimensions based on dictionaries from social psychology research. Along these dimensions we investigate gender bias in contextual embeddings, across contexts and layers, and generate stereotype profiles for twelve different LLMs, demonstrating their intuition and use case for exposing and visualizing bias.

1 Introduction

Amongst many other semantic concepts, large language models (LLMs) pick up stereotypes from the data they are trained on. Unbiased data is hard to come by, especially in the amounts needed for the ever-larger models, which are the foundation of the current successes of AI and the respective hype. Thus bias in these models is basically unavoidable, making it necessary to understand its characteristics and extents to communicate the risks and find ways to mitigate adverse discriminatory effects on affected populations.

Past research on bias has often involved word embedding association tests (Caliskan et al., 2017), inspired by the implicit association tests (IAT) (Greenwald et al., 1998) of social psychology. By another inspiration from the social sciences, a newer direction of natural language processing (NLP) research transforms opaque embeddings into a space of meaningful dimensions

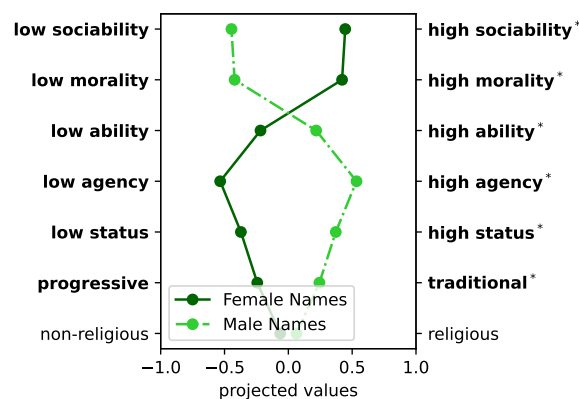


Figure 1: 7D stereotype profile for Llama-3-8B, revealing differences in embeddings of 100 female and 100 male-associated names.

*Statistically significant differences ($p < 0.05$).

(Mathew et al., 2020; Kwak et al., 2021; Şenel et al., 2022; Engler et al., 2022), enabling new ways to study concepts. Similar to semantic differentials (Osgood et al., 1957), this methodology relies on antonyms (e.g. fast vs. slow) or opposing concepts described by lexicons.

In this work we study bias in LLMs by transforming their embeddings based on the stereotype content model (SCM) by Fiske et al. (2002), enabling the study along theoretically and empirically grounded stereotype dimensions. The SCM entails two primary dimensions originating from interactions, where people seek to understand the other party's intent (dimension of warmth) and their capabilities (dimension of competence). Stereotypically, women are thereby associated with higher warmth, and men with higher competence. Furthermore we employ the extended model (Abele et al., 2016; Ellemers, 2017; Goodwin, 2015; Koch et al., 2016), allowing us to provide detailed 7D bias profiles as shown in Figure 1.

dimension	direction	n	terms	$n_{add.}$	additional terms
sociability	high	43	nice, friendliness, warmth	199	accomodating, witty
	low	42	unfriendly, unsociability, distant	162	acid, withdrawn
morality	high	51	humane, morality, benevolent	205	allegiance, true
	low	69	untrustworthiness, evil, insincere	635	abandon, wrongful
ability	high	40	intelligence, capable, graceful	302	accomplished, ace
	low	39	ignorant, stupid, inefficient	160	awkward, unadvised
agency	high	42	motivated, autonomous, resolute	256	action, worker
	low	39	vulnerable, submission, helpless	113	bowing, unsure
status	high	21	superior, wealth, important	187	advantage, win
	low	13	poor, insignificant, unsuccessful	117	bankrupt, welfare
politics	traditional	12	conventional, conservative	34	classical, capitalist
	progressive	16	modern, liberal, democrat	45	contemporary, feminist
religion	religious	18	believer, church, god-fearing	146	spirit, testament
	non-religious	10	atheist, skeptical, secular	6	unholy, impious

Table 1: Examples of terms and their directions on stereotype dimensions from the theoretically grounded dictionary by Nicolas et al. (2021). The additional terms were collected from their extended dictionary created by a semi-automated method. High-level stereotype dimensions are constructed as follows: warmth = sociability + morality, competence = ability + agency.

Contributions. In summary, we (i) show how the stereotype content model can be employed to expose and visualize bias in contextual embeddings¹, (ii) generate bias profiles for twelve LLMs for gender-associated names and gendered terms, displaying overall stereotypical associations of warmth and competence, (iii) provide insights on stereotype dimensions and gender bias across context examples and network layers.

2 Related Work

Inspired by the human implicit association test (Greenwald et al., 1998), Caliskan et al. (2017) developed the first Word Embedding Association Test (WEAT) to assess the association between two target concepts (e.g., scientist vs. librarian) and two attributes (e.g., male vs. female) in static word embeddings by cosine similarity and a permutation test. Later Tan and Celis (2019) built a first approach to measure bias for LLMs using the contextual embeddings of the words within examples. The Contextual Embedding Association Test (CEAT) (Guo and Caliskan, 2021) employs a random effects model to quantify bias with sampled contexts from a corpus.

A newer approach to interpreting the high-dimensional embedding spaces is again inspired by a concept from the social sciences; seman-

¹Code available at <https://github.com/carolinmschuster/profiling-bias-in-llms>

tic differentials (Osgood et al., 1957). Mathew et al. (2020) introduced POLAR, a transformation of static word embeddings to a new polar, interpretable space. The polar opposites are antonyms such as hot-cold or soft-hard, and their word vectors are employed to define the new dimensions, which were shown to align with human judgment in an evaluation study. Similar frameworks are SemAxis (An et al., 2018), FrameAxis (Kwak et al., 2021) and BiImp (Şenel et al., 2022).

More recently, the SensePolar framework was introduced by Engler et al. (2022), extending the POLAR approach to contextual word embeddings. The poles are hereby defined not by the word alone but by their embedding within sense-specific example sentences from a dictionary. The authors showed that these more interpretable embeddings can achieve similar performance to regular ones on natural language understanding (NLU) tasks, and furthermore confirmed the approach by a human evaluation study.

Most similar to our work Fraser et al. (2021) analyzed stereotype dimensions in static embeddings, combining the POLAR framework by Mathew et al. (2020) with the warmth and competence dimensions of the stereotype content model (Fiske et al., 2002). They demonstrated that static word embeddings can recreate the stereotype dimensions from literature by predicting the cold-warm and competent-incompetent associations for

additional known words, and by further comparing the results to psychological surveys.

For contextual embeddings Ungless et al. (2022) measured bias with CEAT (Guo and Caliskan, 2021) based on the warmth and competence dimensions and in a generation-based approach Jeoung et al. (2023) elicited evaluation of different social groups on these dimensions, with multiple prompting strategies. The stereotype content model has also been used for de-biasing methods (Ungless et al., 2022; Omrani et al., 2023). The researchers suggest that this theory-driven approach has an advantage because it is social-group-agnostic and thus does not require iteration over discriminated groups or previous knowledge of specific bias.

In another projection approach, Omrani Sabbaghi et al. (2023) used a maximum margin support vector classifier to learn the valence subspace (pleasantness vs. unpleasantness) and projected the word ‘person’ to this dimension, placing different words in its context. The bias between words is measured by their effect on the contextualized representation of ‘person’.

This work furthermore relates with a broader range of studies trying to understand the contents of contextual representations, most notably by knowledge probing (e.g. Tenney et al. (2019b); Schuster and Hegelich (2022)). See Cao et al. (2024) for a recent survey.

3 Experimental Setup

3.1 Stereotype Dimensions & Dictionaries

Our analysis of stereotype dimensions and bias in LLMs is grounded in the stereotype content model by Fiske et al. (2002), who showed that there are two major dimensions of warmth and competence and that many stereotypes are mixed along these two. Following Nicolas et al. (2021) we also study the more fine-grained dimensions of sociability and morality for warmth (Abele et al., 2016) and ability and agency for competence (Ellemers, 2017; Goodwin, 2015). Expanding the set of concepts by the Agency-Beliefs-Communion model (Koch et al., 2016), we further include the dimensions of status, politics, and religion. This allows us to also provide a more detailed and extended stereotype profile for the LLMs.

Akin to previous work (Fraser et al., 2021; Omrani et al., 2023) we use the dictionaries in the sup-

plementary data from Nicolas et al. (2021)², which were validated by human evaluation studies, for the construction of the stereotype space.

The ‘seed dictionary’ is theory-driven, using terms from literature, while the ‘full dictionary’ contains additional terms collected by a semi-automated method, identifying synonyms using the English lexical database WordNet (Miller et al., 1990)³. Both dictionaries distinguish seven stereotype dimensions, as shown in Table 1 with examples of terms per dimension and direction. While most dimensions are coded low–high, the politics dimension is coded progressive–traditional and the dimension of religion is coded non-religious–religious.

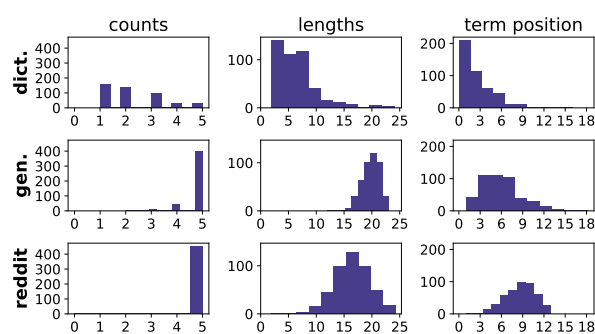


Figure 2: Properties of context examples: Histograms of example counts, numbers of words and positions of dictionary terms within the examples.

3.2 Context Examples

As we are working with contextual embeddings the context of the terms becomes a crucial design choice for the study of stereotype dimensions and bias (see also Engler et al. (2022)).

Generated Examples: For our main experiments, we generate gender non-specific contexts with Llama-3-8B-instruct (AI@Meta, 2024a; Dubey et al., 2024) by the instruction to avoid names and gender-specific pronouns. Thus, no additional gender bias is introduced. As synset information is available for the terms in the seed dictionary, the prompts for these terms additionally include the term definition from WordNet (Miller et al., 1990), allowing a more precise generation for the specific word meaning.

Dictionary Examples: As in the original SensePolar paper (Engler et al., 2022), we also

²<https://osf.io/yx45f/>

³<https://wordnet.princeton.edu/>

retrieve context examples from WordNet (Miller et al., 1990), and only for the seed stereotype dictionary do we manually add examples from other dictionaries where WordNet does not provide any.

Reddit Examples: We include one setup with natural data, where we sample term examples from a Reddit Corpus⁴, similar as done in the Contextual Word Embedding Association Test (Guo and Caliskan, 2021).

No Context: In this setting only the terms are passed through the models, preventing contextualization beyond term subwords.

Example properties are shown in Figure 2. We set the number of examples to five for comparison and to limit computational time, but there often are fewer available for the dictionary examples. Dictionary examples are also the shortest, as they are often short phrases, e.g. “friendly advice”. The generated examples are the longest with an average of 20 words. Reddit examples are truncated on both sides to include context around the term, which may explain later term positions.

3.3 Polar Projection

For the computation of stereotype dimensions, we follow the SensePolar framework by Engler et al. (2022), which is an extension of the POLAR framework (Mathew et al., 2020) for contextual embeddings. Hereby, the embeddings are transformed into an interpretable space based on polar dimensions, which, in our case, are defined by the stereotype content dictionary. We transform the embeddings at two levels: **(i)** Warmth + competence, and **(ii)** seven granular dimensions of the extended stereotype content model.

Similar to Fraser et al. (2021), we take the words for each stereotype dimension from the theory-driven ‘seed dictionary’ (Nicolas et al., 2021), and we average individually the word embeddings for the high and for the low classified words, for which the numbers are shown in Table 1. Word lists for warmth (sociability + morality) and competence (ability + agency) are compiled of the words of their subordinate dimensions.

As a first step, we calculate the sense embedding \mathbf{s} for a word with its specific sense and m sense-specific context examples, as shown in Equation 1. We hereby average the contextual embeddings \mathbf{w} across the different context examples,

also averaging across subwords when words are split due to subword tokenization.

$$\mathbf{s} = \frac{1}{m} \sum_{j=1}^m \mathbf{w}_{c_j}^s \quad (1)$$

For n words belonging to the pole of a stereotype dimension, e.g. “friendliness” and “sociability” for the pole “high sociability”, we average their sense embeddings to obtain the average pole embedding \mathbf{p} . Next, we stack the vectors and subtract the low-dimension embeddings from the high-dimension embeddings to obtain the change of basis matrix \mathbf{a} , describing the newly defined space with \mathbf{h} stereotype dimensions:

$$\mathbf{p} = \frac{1}{n} \sum_{i=1}^n \mathbf{s}_i \quad (2)$$

$$\mathbf{a}_h = \mathbf{p}_{high} - \mathbf{p}_{low} \quad (3)$$

Regarding the warmth and competence transformation, there are only two direction vectors. If, for example, the original contextual embedding has 768 dimensions, the change of basis matrix \mathbf{a} has a shape of (2, 768).

Before projecting a word of interest to the new dimensions, we compute its embedding \mathbf{x} by again averaging across its context examples as shown in Equation 4. Following prior work (Engler et al., 2022) we then project the embedding to the new interpretable space by the inverted change of basis matrix as shown in Equation 5.

$$\mathbf{x} = \frac{1}{k} \sum_{i=1}^k \mathbf{x}_{c_i} \quad (4)$$

$$\mathbf{d} = (\mathbf{a}^T)^{-1} \mathbf{x} \quad (5)$$

The new embedding \mathbf{d} in the 2D or 7D stereotype space can be interpreted as follows: Similar to the semantic differential technique, a higher value signifies a higher association with the high pole, for example, “high morality”, and a lower value signifies a more significant association with the low pole, e.g., “low morality”. By projecting multiple terms we can compare their differences on these dimensions.

⁴<https://www.kaggle.com/datasets/kaggle/reddit-comments-may-2015>

	Warmth	Competence	Sociability	Morality	Ability	Agency	Status	Politics	Religion
Llama-3-8B (AI@Meta, 2024a)	0.79	0.81	0.66	0.87	0.72	0.77	0.8	0.73	0.74
Llama-3-8B-Instruct (AI@Meta, 2024a)	0.79	0.82	0.65	0.88	0.75	0.77	0.83	0.83	0.81
Llama-3.2-3B (AI@Meta, 2024b)	0.81	0.8	0.63	0.87	0.66	0.76	0.79	0.77	0.62
Llama-3.2-3B-Instruct (AI@Meta, 2024b)	0.78	0.84	0.62	0.88	0.75	0.78	0.81	0.86	0.62
Gemma-2B (TeamGemma et al., 2024a)	0.66	0.67	0.58	0.75	0.66	0.69	0.78	0.58	0.96
Gemma-2-2B (TeamGemma et al., 2024b)	0.7	0.67	0.64	0.8	0.74	0.7	0.75	0.63	0.95
OLMo-1B-hf (Groeneveld et al., 2024)	0.83	0.84	0.76	0.86	0.83	0.77	0.78	0.87	0.6
Bloom-1B7 (Le Scao et al., 2022)	0.79	0.76	0.68	0.87	0.85	0.77	0.62	0.71	0.84
GPT-Neo-125M (Black et al., 2022)	0.66	0.33	0.55	0.75	0.34	0.69	0.41	0.58	0.05
GPT2 (Radford et al., 2019)	0.66	0.67	0.57	0.76	0.53	0.72	0.77	0.6	0.93
ALBERT-base-v2 (Lan et al., 2019)	0.7	0.68	0.62	0.77	0.7	0.69	0.71	0.69	0.65
BERT-base-uncased (Devlin et al., 2019)	0.79	0.83	0.7	0.83	0.8	0.72	0.78	0.72	0.53

Table 2: Accuracy for the direction prediction task. Additional terms in the extended stereotype dictionary (Nicolas et al., 2021) are embedded and projected to the stereotype dimensions. Projected positive/negative values predict high/low direction, e.g. a value of -0.3 for warmth is registered as low warmth. The highest accuracy for each dimension is shown in bold. Please refer to Table 1 for examples of terms with high and low labels for each dimension.

Projection of Additional Terms To evaluate the consistency of the stereotype dimensions we follow the approach by Fraser et al. (2021) and project additional terms from the extended dictionary by Nicolas et al. (2021) to the stereotype space. Hereby, we use the same types of context examples as for the polar space creation. For each term, we predict its direction on its assigned dimension by the sign of its polar value, e.g., a value of -0.5 for sociability is registered as low sociability. To calculate the accuracy, we compare these predictions against the labels in the dictionary.

Projection of Gender-Associated Names & Gendered Terms For the analysis of gender bias, we project gender-associated words to our stereotype dimensions, utilizing two larger binary ‘vocabulary populations’ and individual terms for transgender and nonbinary gender (see Figure 4).

The largest populations are 100 historically female-associated names (e.g., Mary, Patricia) and 100 male-associated names (e.g., James, Michael), taken from the most popular given names of the last century in the United States⁵.

Second, we employ binary gendered terms by definition as utilized in experiments of WEAT (Math vs. Arts and Science vs. Art) (Caliskan et al., 2017). For each gender, we project nine terms:

- Female terms: female, woman, girl, sister, she, daughter, mother, aunt, grandmother
- Male terms: male, man, boy, brother, he, son, father, uncle, grandfather

As examples for our gendered terms and names, we use neutral templates, such as “This is [NAME]” or “This is [TERM]” to provide context without unnecessarily introducing additional bias (compare May et al. (2019); Tan and Celis (2019)). We average across the different templates for a more robust contextual representation of names and terms.

For easier interpretation and comparison between models, we standardize the projected polar values separately for names and terms, as preliminary work showed that named entities and pronouns can show different average tendencies on stereotype dimensions. To assess the significance of the observed differences, we employ t-tests.

3.4 Models

For our evaluation, we project open source models of multiple generations available in the Huggingface Library⁶ onto stereotype dimensions. Model names and references are shown in Table 2. Except for the layer-wise analysis, we extract their average contextual representations across all layers, including the first embedding layer.

⁵<https://www.ssa.gov/oact/babynames/decades/century.html>

⁶<https://huggingface.co/>

4 Results

4.1 Prediction of Direction for Additional Terms

Predicting the direction on the stereotype dimensions for new terms from the extended stereotype dictionary (Nicolas et al., 2021), we find most studied models can perform this task well, with different strengths. Table 2 shows the results when embedding and projecting with the generated examples. OLMo-1B-hf (warmth and competence) and Llama3.2-3B-Instruct (only competence) achieve the performance closest to that of static embeddings by Fraser et al. (2021), where the FastText-based model scored respectively 0.85 for warmth and 0.86 for competence. OIMO and the Llama models also perform very well for the granular dimensions. Predicting morality is the overall easiest task for the models, while the other subdimension of warmth, sociability, poses the most difficult task. Accuracy varies greatly for religion, where there are 142 high, but only 6 low-labeled additional terms.

Only GPT-Neo-125M performs worse than chance for some dimensions, however, this pertains only to raw polar values. When we use a different cut-off than zero for predicting low/high directions by mean-centering the projected values, the model achieves much better results, e.g., 0.76 accuracy for warmth and 0.71 for competence. Similarly, GPT2 and the Gemma models benefit from a mean-based cut-off value, gaining up to 10 percentage points per dimension. Thus projections are spread on different ranges of values, but all models can reasonably discriminate between low and high-labeled terms on the stereotype dimensions.

4.2 Gender Stereotype Profiles

For all twelve studied LLMs, we find statistically significant bias for gender-associated names, as evident in the 2D profiles for warmth and competence in Figure 3. In line with human bias found in studies of the stereotype content model (Fiske et al., 2002), the models highly agree on the relative associations of female names with warmth and male names with competence when using the gender non-specific generated contexts. GPTNeo poses an exception, where both dimensions are associated with female names, and for OLMo, the difference in competence is insignificant.

For the much smaller ‘vocabulary populations’

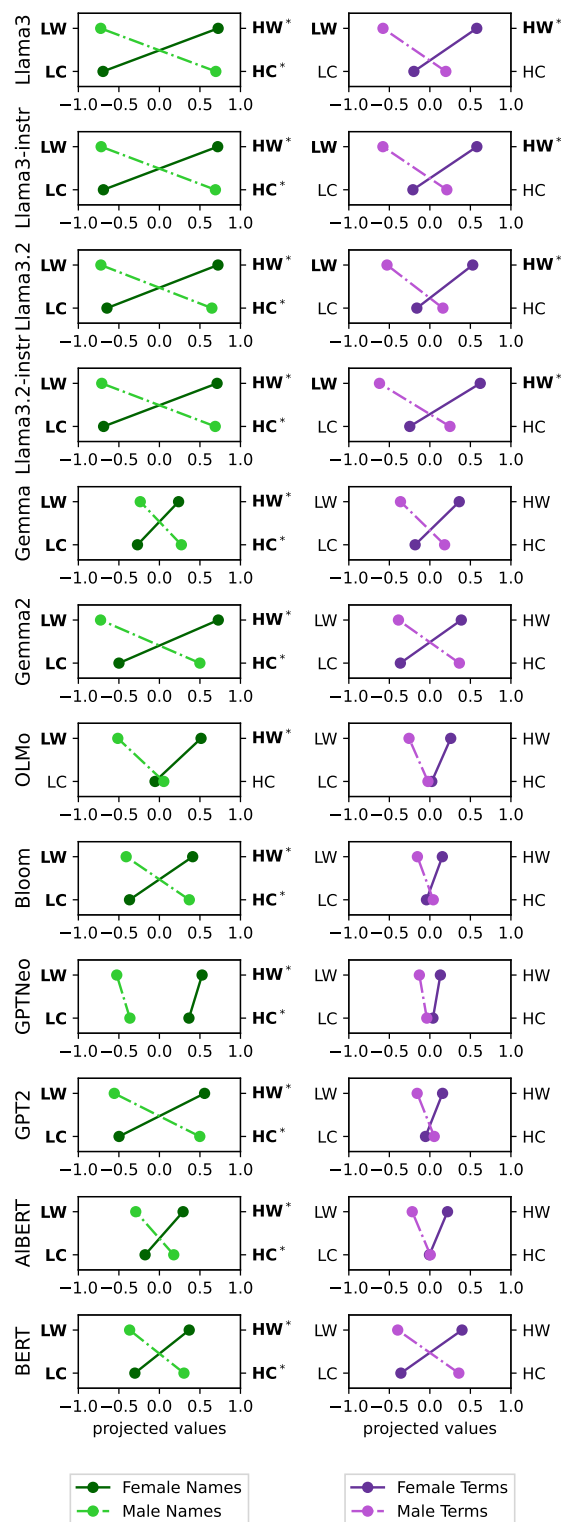


Figure 3: 2D stereotype profiles for 100 female/male-associated names (left) and 9 female/male gendered terms (right). LW/HW = Low/High Warmth. LC/HC = Low/High Competence. *Dimensions with statistically significant differences ($p < 0.05$).

of gendered terms (nine terms per gender), warmth is the more relevant dimension than competence, with only the former being significantly biased for all four Llama-3 models. For some models, e.g. GPT2, the gendered term differences are small, but the bias direction is very consistent when comparing name and term stereotype profiles. In Figure 4, we additionally see the projections of five individual terms beyond binary gender. Nonbinary and transgender-related terms are associated with lower warmth than binary term means, which was found for all newer models (studied variants of Llama3, Gemma, OLMo). For competence, there was no observable trend. For further and statistical analyses of individual terms and small vocabulary populations, future work needs to extend the context examples (as discussed in section 5).

Zooming into the 7-dimensional stereotype profiles for Llama-3-8B (see Figure 1) and Llama-3.2-3B-instruct in Figure 5, we can perceive the previously shown gender associations in more detail. The profiles between Llama-3 and the newer and instruction-tuned 3.2 version are quite similar: Sociability and morality (warmth) are linked with female names, which is true for 10 and 11 of all studied models. Ability and agency (competence) are significantly related with male names, which is true for 9 and 7 of the studied models. Also for status, six models show a significant association with male names. Furthermore six models find male names to be more traditional on the political dimension, while there is no perceivable trend for religion. For the small populations of binary gendered terms, there is only one clearly biased dimension, where seven models agree on a stereotypical association of sociability with female terms.

4.3 Context Examples and Bias across Layers

Comparing the performance for generated, dictionary, Reddit examples, and no context across layers in Figure 6a, we find that the influence of context on the new term prediction task depends on the model. For Llama-3-8B, the accuracy seems quite stable compared to the smaller Gemma-2B variant, where we see high variation across layers and context types. However, as elaborated in subsection 4.1, while the cut-off value of zero works well for most, including all Llama models, Gemma could better discriminate between low/high labeled terms by a different cut-off value.

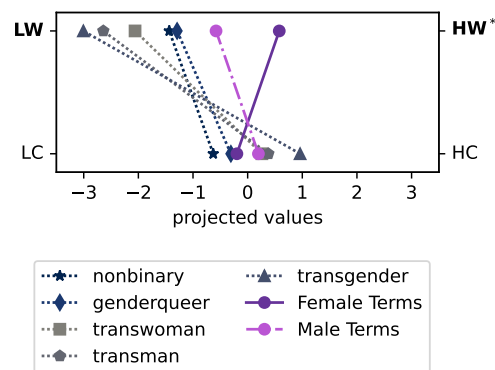


Figure 4: 2D Stereotype profile for Llama-3-8B (see Figure 3) with additional projections of individual nonbinary terms. LW/HW = Low/High Warmth. LC/HC = Low/High Competence.

For GPT2, the Reddit examples lead to much lower accuracy, while for BERT, the no context condition performs markedly worse. Overall the concepts of warmth and competence behave similarly throughout the layers.

On the right in Figure 6b, we see that bias across layers is rather consistent for all models, where higher values signify bias towards female-associated names/terms and lower values signify bias towards male-associated names/terms. Shown by the example of the generated contexts, stereotypical associations permeate throughout the networks. In some cases, the first and last layers behave differently, with a reversed bias direction compared to the overall model.

5 Discussion

Our results provide substantial evidence of stereotype dimensions in the embedding space of LLMs and a gender bias that predominantly corresponds to the human bias found in studies of the stereotype content model (Fiske et al., 2002). For all studied models, female names are relatively associated with higher warmth, and for most models, male names are associated with higher competence. There is less evidence of bias for the studied gendered terms, which in part is likely due to the small groups of only nine terms per gender. The direction of gender differences is, however, overwhelmingly consistent. We furthermore find stereotype dimensions and bias across layers, in line with prior work that semantics are spread throughout the network (Tenney et al., 2019a).

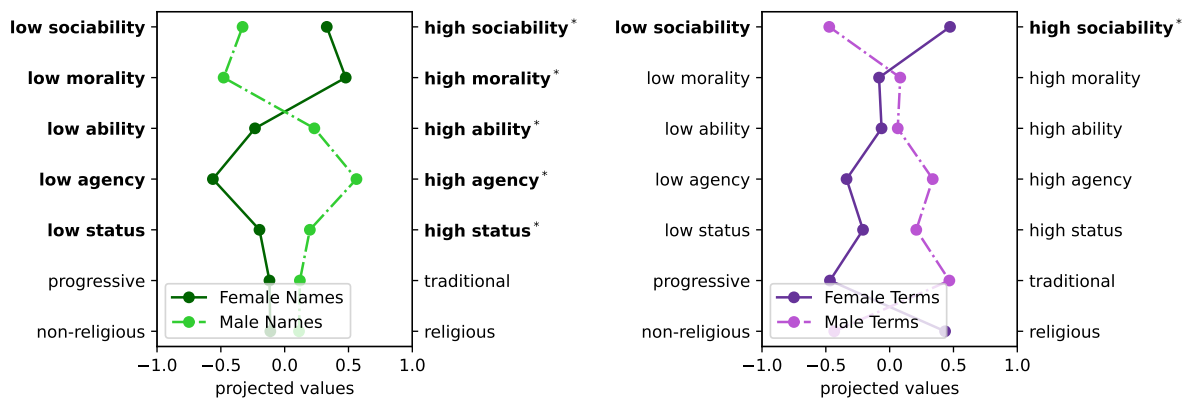


Figure 5: 7D stereotype profiles for 100 female/male-associated names (left) and 9 female/male gendered terms (right) for Llama-3.2-3B-instruct. *Statistically significant differences ($p < 0.05$).

The projection of contextual embeddings based on the stereotype content model can deliver robust insights when analyzing large vocabulary groups. As the magnitude of the values depends on the properties of the original embedding space, statistical analysis is employed to assert the significance of gender differences. This is viable with the larger collection of gender-associated names, also providing context for the differences between binary gendered terms. For the analysis of small vocabulary groups or individual terms, e.g. for comparing terms of binary and nonbinary gender, increasing the number of context examples offers potential for statistical tests.

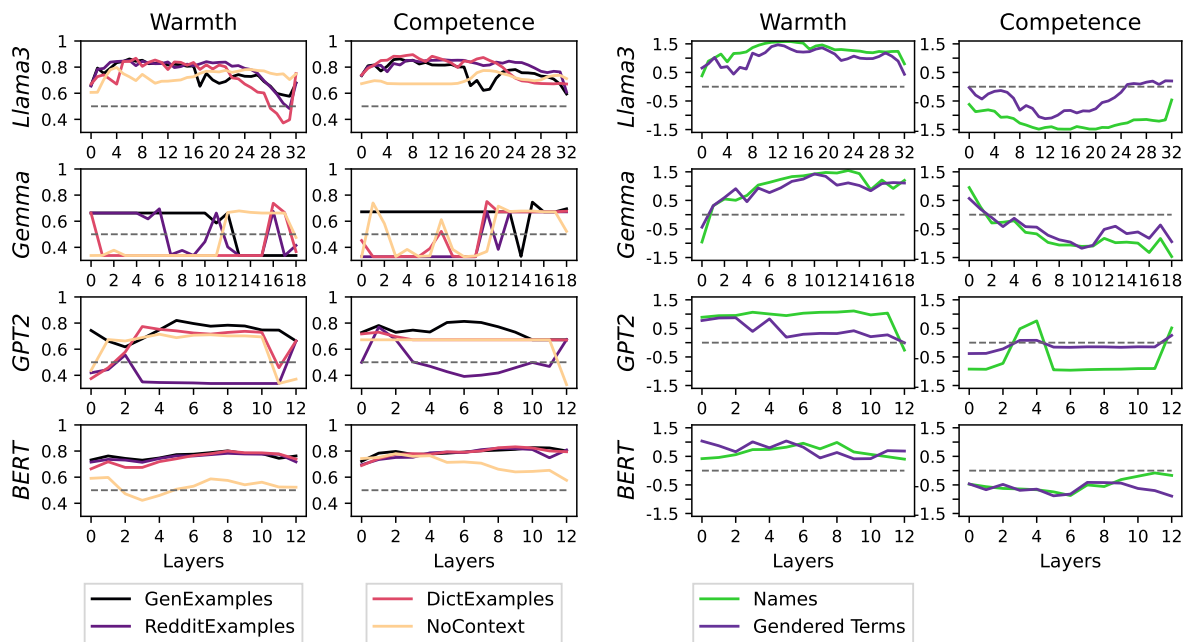
While both 2D and 7D stereotype dimensions provide interesting results, a significant gender bias is most evident in the warmth and competence dimensions. These benefit from the larger numbers of low and high-rated words in the dictionaries, increasing the robustness of the concept representations. Likely associations are also more stable when relating to broader concepts. Therefore, the high-level projection is a suitable first level of analysis and starting point for bias mitigation.

Significant gender bias, however, may also occur on a more granular level. Different dimensions can be relevant depending on domains and tasks, such as progressive-traditional in the realm of politics, and the mode of projection can be easily adapted with the presented methodology. A combined projection with other dimensions such as valence (unpleasantness vs. pleasantness) (see e.g. Omrani Sabbaghi et al. (2023)), could provide even further insights.

As we have shown, the term context can have

a considerable influence on the behavior of the stereotype dimensions. Thus, examples for pole and projected terms should be chosen deliberately. Gender non-specific context is our default choice because no additional bias is introduced through the examples and we get a clearer picture of the bias already present within the pre-trained embeddings. Even smaller open-source LLMs are now able to provide examples with this property at scale. However, measurement is certainly best conducted with domain-specific data when a specific use case exists. While we use simple templates as contexts for the gendered terms and names, these could as well be sampled from a target domain or be generated to test specific scenarios. For example, similar to May et al. (2019), this could involve introducing success in a historically male-dominated field to the term/name context, to test if a penalty exists for females, as found in psychological studies (Heilman et al., 2004).

While embedding-based methods for bias measurement have been critiqued for their remoteness from downstream applications (Gallegos et al., 2024), and are certainly no substitute for task-specific investigations, they have multiple advantages. First, the methodology does not depend on natural language datasets that can be leaked into training data and are therefore applicable to older and newer models alike. Second, the same stereotype dimensions can easily be used for bias mitigation (Ungless et al., 2022; Omrani et al., 2023), alleviating representational harm and the risk that it influences downstream behavior. Finally, our paper shows they can be exploited for intuitive visualizations exposing gender bias.



(a) Accuracy for the direction prediction task across layers, with different context examples. Additional terms are projected to the stereotype dimensions; positive/negative values predict high/low direction.

(b) Gender bias across layers with generated gender non-specific examples. Higher values signify bias towards female-associated names/terms; lower values signify bias towards male-associated names/terms.

Figure 6: Layerwise visualization of prediction accuracy and gender bias for selected models.

6 Conclusion

Large pre-trained language models reflect the biases in their training data, which in turn reflect the biases of their creators. As the foundation for AI applications, their biases are further propagated, warranting their study to uncover the risks and promote mitigation efforts.

In this work, we profile gender stereotypes in twelve LLMs by means of the stereotype content model of social psychology (Fiske et al., 2002), thereby theoretically grounding the analysis, which has in the past been described as the missing link for bias measurements (Blodgett et al., 2020). By a matrix transformation, the opaque contextual embeddings of the models reveal interpretable stereotype dimensions. Along the two major dimensions of warmth and competence, we find significant bias for gender-associated names and some evidence of bias for gendered terms, widely aligned with stereotypes found in human studies.

The shown presence of stereotype dimensions in LLMs is a comprehensible replication of semantics in human language, however, the differential associations of social groups along these dimensions constitute a representational harm.

Equal treatment starts with equal representation; stereotypes already statistically significant in embedding space come with the risk of being exploited in downstream tasks, which could lead to different and unfair treatment of social groups. While the first access point would be the training data itself, the embedding space allows a quantification of patterns that is useful for bias assessment and mitigation. The analysis of embedding spaces by interpretable dimensions provides a means to evaluate both functional semantics and harmful associations that should be ‘unlearned’ to prevent their propagation.

Awareness of bias in LLMs needs to be increased beyond the expert audience, as biased models are already deployed, and completely debiased models may hardly be attainable, as they are trained on vast amounts of biased human-created data. The here presented bias profiles based on the stereotype content model employ an intuitive scoring along meaningful scales of opposing concepts (e.g. low vs. high warmth), as proven effective by semantic differentials in human surveys (Osgood et al., 1957). The result is a highly visual solution for communicating bias to wider audiences and users of artificial intelligence.

7 Limitations

The bias profiles presented in this paper concern only gender, but there is a whole range of biases to be profiled in LLMs to evaluate and communicate representational harms. The scope of analysis was also constrained to English contextual embedding spaces and needs to be extended to a multi-lingual setting in the future.

Furthermore, the focus of this paper was binary gender, with historically gender-associated names and gendered terms. While we projected a few terms for transgender and nonbinary gender to the stereotype dimensions, future analysis needs to extend the methodology for these smaller and diverse ‘vocabulary populations’. Increasing the number of context samples for terms offers potential for greater robustness and applicability of statistical tests.

Finally, no general bias measurement benchmark or method, including the one presented in this paper, precludes the absolute necessity of task-specific bias measurements. However, they can be a piece of the puzzle by revealing learned general bias tendencies and providing a means to mitigate and communicate these effectively.

References

- Andrea E. Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality. *Frontiers in psychology*, 7:1810.
- AI@Meta. 2024a. Llama 3 model card. URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- AI@Meta. 2024b. Llama 3.2 model card. URL: https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.
- Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2022. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow, 2021. URL: <https://doi.org/10.5281/zenodo.5297715>.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Boxi Cao, Hongyu Lin, Xianpei Han, and Le Sun. 2024. The life cycle of knowledge in big language models: A survey. *Machine Intelligence Research*, 21(2):217–238.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Naomi Ellemers. 2017. *Morality and the regulation of social behavior: Groups as moral anchors*. Psychology Press.
- Jan Engler, Sandipan Sikdar, Marlene Lutz, and Markus Strohmaier. 2022. SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4607–4619, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902.
- Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.
- Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K

- Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.
- Geoffrey P. Goodwin. 2015. Moral Character in Person Perception. *Current Directions in Psychological Science*, 24(1):38–44.
- Anthony G. Greenwald, Debbie E. McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.
- Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. *Preprint*.
- Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.
- Madeline E. Heilman, Aaron S. Wallen, Daniella Fuchs, and Melinda M. Tamkins. 2004. Penalties for success: reactions to women who succeed at male gender-typed tasks. *Journal of applied psychology*, 89(3):416.
- Sullam Jeoung, Yubin Ge, and Jana Diesner. 2023. StereoMap: Quantifying the awareness of human-like stereotypes in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12236–12256, Singapore. Association for Computational Linguistics.
- Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of personality and social psychology*, 110(5):675.
- Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7:e644.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2020. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. In *Proceedings of The Web Conference 2020*, pages 1548–1558.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.
- Gandalf Nicolas, Xuechunzi Bai, and Susan T. Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.
- Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Deghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.
- Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. 2023. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 542–553.
- Charles Egerton Osgood, George J. Suci, and Percy H. Tannenbaum. 1957. *The measurement of meaning*. 47. University of Illinois press.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Carolyn M. Schuster and Simon Hegelich. 2022. From BERT’s Point of View: Revealing the Prevailing

- Contextual Differences. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1120–1138, Dublin, Ireland. Association for Computational Linguistics.
- Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32.
- TeamGemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- TeamGemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovered the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. Association for Computational Linguistics.
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.
- Eddie Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 207–217, Abu Dhabi, UAE. Association for Computational Linguistics.
- Lütfi Kerem Şenel, Furkan Şahinuç, Veysel Yücesoy, Hinrich Schütze, Tolga Çukur, and Aykut Koç. 2022. Learning interpretable word embeddings via bidirectional alignment of dimensions with semantic concepts. *Information Processing & Management*, 59(3):102925.