Towards Region-aware Bias Evaluation Metrics

Anonymous ACL submission

Abstract

When exposed to human-generated data, language models are known to learn and amplify societal biases. While previous works introduced benchmarks that can be used to assess the bias in these models, they rely on assumptions that may not be universally true. For instance, a gender bias dimension commonly used by these metrics is that of *family–career*, but this may not be the only common bias in certain regions of the world. In this paper, we identify topical differences in gender bias across different regions and propose a region-aware bottom-up approach for bias assessment. Our proposed approach uses gender-aligned topics for a given region and identifies gender bias dimensions in the form of topic pairs that are likely to capture gender societal biases. Several of our proposed bias topic pairs are on par with human perception of gender biases in these regions in comparison to the existing ones, and we also identify new pairs that are more aligned than the existing ones. In addition, we use our region-aware bias topic pairs in a Word Embedding Association Test (WEAT)-based evaluation metric to test for gender biases across different regions in different data domains. We also find that LLMs have a higher alignment to bias pairs for highly-represented regions showing the importance of region-aware bias evaluation metric.

1 Introduction

Human bias refers to the tendency of prejudice or preference towards a certain group or an individual and can reflect social stereotypes with respect to gender, age, race, religion, and so on.

Bias in machine learning refers to prior information which is a necessary prerequisite for intelligence (Bishop, 2006). However, biases can be problematic when prior information is derived from *harmful precedents* like prejudices and social stereotypes. Early work in detecting biases includes the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) and the Sentence Encoder Association Test (SEAT) (May et al., 2019). WEAT is inspired by the Implicit Association Test (IAT) (Greenwald et al., 1998) in psychology, which gauges people's propensity to unconsciously link particular characteristics—like *family* versus *career*—with specific target groups—like female (F) versus male (M). WEAT measures the distances between target and attribute word sets in word embeddings using dimensions[1] similar to those used in IAT.

Biases toward or against a group can vary across different regions due to the influence of an individual's culture and demographics (Grimm and Church, 1999; Kiritchenko and Mohammad, 2018a; Garimella et al., 2022). Psychological studies and experiments that demonstrate human stereotypes vary by continental regions (Damann et al., 2023; Blog, 2017) and even larger concepts like western and eastern worlds (Markus and Kitayama, 2003; Jiang et al., 2019) serve as an inspiration for the use of regions to determine differences across cultures. However, existing bias evaluation metrics like WEAT and SEAT follow a "one-size-fits-all" approach to detect biases across different regions. As biases can be very diverse depending on the demographic lens, a fixed or a small set of dimensions (such as family–career, math–arts) may not be able to cover all the possible biases in the society. In this paper, we address two main research questions about gender bias: (1) Is it possible to use current NLP techniques to automatically identify gender bias characteristics (such as family, career) specific to various regions? (2) How do these gender dimensions compare to the current generic dimensions included in WEAT/SEAT?

Our paper makes four main contributions:

1. An automatic method to uncover gender bias dimensions in various regions that uses (a) topic modeling to identify dominant topics aligning with the F/M groups for different re-

[1] We use 'topic pairs' and 'topic dimensions' interchangeably.

gions, and (b) an embedding-based approach to identify F-M topic pairs for different regions that can be viewed as gender bias dimensions in those regions.

2. An IAT-style test to assess our predicted gender bias dimensions with human subjects. To the best of our knowledge, this is the first study to use a data-driven, bottom-up method to evaluate bias dimensions across regional boundaries.

3. A WEAT-based evaluation setup using our region-aware topic pairs to evaluate gender biases in different data domains (Reddit and UN General Debates) across regions.

4. An analysis of how well our predicted bias dimensions align with those of custom LLMs. We consider several LLMs that include open-source models like `Llama-3-8b` and `Mistral-7b-Instruct`; as well as closed-source models such as `GPT-4`, `Gemini-Pro` and `Claude-3-Sonnet`.

2 Data

We use GeoWAC (Dunn and Adams, 2020a), a geographically balanced corpus that consists of web pages from Common Crawl. Language samples are geo-located using country-specific domains, such as an $.in$ domain suggesting Indian origin (Dunn and Adams, 2020b). The GeoWAC's English corpus spans 150 countries. We select the top three countries with the most examples per region: Asia, Africa, Europe, North America, and Oceania as in (Garimella et al., 2022). We randomly choose 282,000 examples (after pre-processing) for each region, with 94,000 examples belonging to each country within the regions. Dataset details are included in Appendix A.

3 Variations in Gender Bias Tests Across Regions

We start by investigating the differences in existing gender bias tests across different regions using WEAT. WEAT takes in *target words* such as male names and female names, to indicate a specific group, and *attribute words* that can be associated with the *target words*, such as *math* and *art*. It computes bias by finding the cosine distance between the embeddings of the target and attribute words. We compute WEAT scores using word2vec embeddings (Mikolov et al., 2013) trained on the five regions separately. Table 1 shows the region-wise scores for the three gender tests in WEAT.

TARGET WORDS - ATTRIBUTE WORDS	REGION	WEAT
career vs family - Male names vs Female names	Africa	1.798
	Asia	1.508
	North America	1.885
	Europe	1.610
	Oceania	1.727
Math vs Arts - Male terms vs Female terms	Africa	1.429
	Asia	1.187
	North America	0.703
	Europe	0.334
	Oceania	1.158
Science vs Arts - Male terms vs Female terms	Africa	1.247
	Asia	0.330
	North America	0.036
	Europe	-0.655
	Oceania	0.725

Table 1: Region-wise WEAT scores using word2vec.

Although we see a positive bias for most gender bias dimensions, the scores vary across regions. For example, the highest scoring regions vary for the target words-attribute words groups. For *family–career* dimension, North America shows the highest bias, however for *math–arts* and *math–science* dimensions, Africa shows the highest bias. Europe has a negative bias on *science–arts* (indicating a stronger F-science and M-arts association).

These results provide preliminary support to our hypothesis that gender bias dimensions vary across regions, thus propelling a need to come up with further bias measurement dimensions to better capture gender biases in these regions in addition to the existing generic ones in WEAT.

4 A Method to Automatically Detect Bias Dimensions Across Regions

Building upon our WEAT findings, we propose a two-stage approach to automatically detect region-aware bias dimensions that likely capture the biases in specific regions in a bottom-up manner. In the first stage, we utilize topic modeling to identify prominent topics in each region. In the second stage, we use an embedding-based approach to find pairs of topics among those identified in the first stage that are likely to represent prominent gender bias dimensions in each region. Fig 1 shows the pipeline of our methodology.

4.1 Identifying Region-wise Bias Topics

We use topic modeling to identify dominant topics in the male and female examples in each region.

We first build F(emale)- and M(ale)-aligned datasets using the examples from GeoWAC for

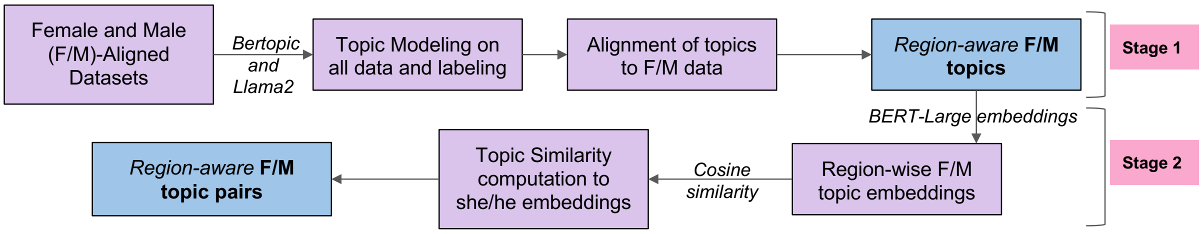


Figure 1: Methodology Pipeline: Stage 1 refers to the extraction of region-aware gender topics using topic modeling, Stage 2 refers to extraction of region-aware gender topic pairs using an embedding based approach

each region. We use the 52 pairs of gender-defined words that are non-stereotypically F/M (e.g., wife, brother, see Appendix G) from (Bolukbasi et al., 2016), and find examples that contain these words. These datasets are used to find gender-aligned topics from GeoWAC. The dataset statistics are specified in Table 6 in Appendix B.

We then use topic modeling to identify dominant topics in the male and female examples in each region. We use `Bertopic` (Grootendorst, 2022), which identifies an optimal number of topics n for a given dataset (see Appendix L.1 for implementation details). We further refine the resulting topics using `Llama2` (Touvron et al., 2023) to label and better understand the topic clusters identified by `Bertopic`. The prompting mechanism for `Llama2` is provided in Appendix H.

We next compute the alignment of the topics to either of the F/M groups. We first compute the topic distribution of a data point, which gives the probability p_{it} of an example i belonging to each topic t. For a topic t, we take n examples that dominantly belong to topic t: $i_1, i_2, ..., i_n$. If m out of n data points belong to the F group in the F-M dataset, and the other (n - m) belongs to the M group, we compute the average of topic probabilities for both groups separately: $p_{Ft} = \frac{(p_{i_1 t}+p_{i_2 t}+......+p_{i_m t})}{m}$ and $p_{Mt} = \frac{(p_{i_{m+1} t}+p_{i_{m+2} t}+......+p_{i_n t})}{(n-m)}$, where p_{Ft} and p_{Mt} refer to the average probability by which a topic dominantly belongs to the F and M groups respectively. If $p_{Ft} > p_{Mt}$, we say the topic is a *bias topic* that aligns with the F group and vice-versa.

4.2 Finding Topic Pairs as Region-wise Bias Dimension Indicators

We use an embedding-based approach to identify F-M topic pairs from the pool of topics identified in the previous stage, to generate topic pairs (bias dimensions) that are comparable to IAT/WEAT pairs.

We use BERT-large (`stsb-bert-large`) from SpaCy's (Honnibal and Montani, 2017) `sentencebert` library to extract contextual embeddings for topic words extracted from the GeoWAC dataset for each region. For a topic t consisting of topic words $w_1, ..w_n$, the topic embedding is given by the average of embeddings of the top ten topic words in that topic.

We identify topic pairs from the embeddings taking inspiration from (Bolukbasi et al., 2016): let the embeddings of the words she and he be E_{she} and E_{he} respectively. The embedding of a topic t_i be E_{t_i}. A female topic F_{t_i} and a male topic M_{t_j} are a topic pair if: $cos(E_{F_{t_i}}, E_{she}) \sim cos(E_{M_{t_j}}, E_{he})$ and/or $cos(E_{F_{t_i}}, E_{he}) \sim cos(E_{M_{t_j}}, E_{she})$, where $cos(i, j)$ refers to the cosine similarity between embeddings i and j, given by $cos(i, j) = \frac{i.j}{||i||||j||}$. The threshold for the difference between the cosine similarities we consider for two topics to be a pair is 0.01, i.e., two topics $(t1, t2)$ are considered a pair if the difference of cosine similarities cos($t1$, she)/cos($t1$, he) and cos($t2$, he)/cos($t2$, she) respectively is < 0.01. We manually choose 0.01 since differences close to 0.01 are almost $= 0$.

4.3 Human Validation Setup

We design an IAT-style test to validate our topic pairs with annotators from different regions. We recruit six annotators from each region controlled by gender (three female and three male). In addition to our topics, we also test for existing WEAT dimensions relating to gender, namely *family–career, math–arts, and science–arts*. For each region, we validate all the region-aware topic pairs using the assistance of our annotators.

As done in IAT, to verify a topic pair, we show the topic names and male/female faces to our annotators along with a set of guidelines. As shown in Fig 2, each topic pair test form contains two tasks. First, the annotators have to press one key for a female face f and a female topic T_f and another key for a male face m with a male topic T_m, timing responses as r_1 and r_2. In the reverse task, they pair T_m with f and T_f with m, timing these as r_3 and r_4. We average r_1 and r_2 for the 'un-reversed' case and r_3 and r_4 for the 'reversed' case. The annotators' implicit association of a gender to a

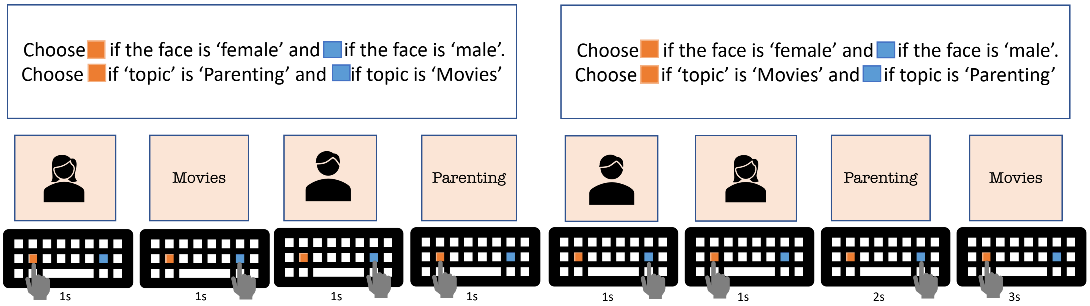


Figure 2: IAT-style test with region-aware topic pairs for human validation. The above example shows the user implicitly associates female to *parenting* and male to *movies*: When guidelines are reversed, they take longer time. Note that we randomize the order of tests for participants to ensure initial pairing bias is accounted for. Also, we have several pages showing faces and topics for each guideline.

topic may influence their response time. A lower response time suggests easier recollection of the guidelines and potential implicit gender-topic associations, and thus lower bias with respect to these topics. We also varied the test order for different annotators to avoid initial pairing bias. We conduct the survey with six annotators each from Africa, Asia, Europe, and North America, also including a *family-career* topic pair, a standard WEAT bias dimension. We provide screenshots of our annotation framework in Appendix M.

4.4 Results: Bias Dimensions across Regions

4.4.1 Region-wise Bias Topics

Table 2 displays the top topics based on u_{mass} (Mimno et al., 2011) coherence for each region.

Several topics are exclusive to certain regions. Some topics like *family* and *parenting*; *cooking*; *pets* and *animal care* are common across some regions for F. Similarly we have *movies*; *politics* and *government*; *sports*; and *music* for M. Finally, there are differences between regions in terms of *education*, *reading*, and *research* (F-Europe, NA, and M-Africa), and *fashion* and *lifestyle* (F-Europe, NA, and M-Africa). Some other popular topics across regions are *religion and spirituality*, *Christian theology* in M; *obituaries and genealogy*, *online dating*, *travel*, and *sailing* in F (see Appendix D for a comprehensive list of topics). We provide an example of topic clusters in Appendix J.

REGION	FEMALE	MALE
Africa	Credit cards and finances, Royalty and Media, Trading strategies and market analysis, Dating and relationships guides, Parenting and family relationships	Fashion and Lifestyle, Male enhancement and sexual health, Nollywood actresses and movies, Nigerian politics and government, Essay writing and research
Asia	Hobbies and Interests, Healthy eating habits for children, Social media platforms, Royal wedding plans, Online Dating and Chatting	DC comic characters, Mobile Application, Phillippine Politics and Government, Sports and Soccer, Career
Europe	Pets and animal care, Fashion and Style, Education, Obituaries and Genealogy, Luxury sailing	Political developments in Northern Ireland, Christian Theology and Practice, Crime and murder investigation, EU Referendum and Ministerial Positions, Criminal Justice System
North America	Pets, Cooking: culinary delights and chef recipes, Fashion and style, Family dynamics and relationships, Reading and fiction	Civil War and history, Middle East conflict and political tensions, Movies and filmmaking, Political leadership and party dynamics in Bermuda, Rock Music and songwriting
Oceania	Cooking and culinary delights, Romance, Weight loss and nutrition for women, Water travel experience, Woodworking plans and projects	Harry Potter adventures, Art and Photography, Superheroes and their Universes, Music recording and Artists, Football in Vanuatu

Table 2: Top five topics for F and M for each region, extracted using Bertopic and Llama2.

4.4.2 Region-wise Bias Dimensions

Table 3 shows the top five topic pairs per region, chosen based on the u_{mass} score from the top 10 topics each for F and M from the topic modeling scheme. As expected, topic pairs differ by region, and we also note new topic pairs that do not appear in the WEAT tests. Among the top ones, there are recurring topics in F such as *dating and marriage*, *family and relationships*, *luxury sailing*, and *education*, whereas in M, we have *politics*, *religion*, *sports*, and *movies*. These region-specific pairs may supplement generic tests to detect regional biases.

4.4.3 Unigram/Bigram Analysis

We find several topics that are common across regions. However, they may differ across cultures and may reveal varied perceptions of biases. Several topics also change associations to genders based on regions. For example, *'fashion and lifestyle'* in Africa is associated with *males*, however, it is associated with *females* in Europe and

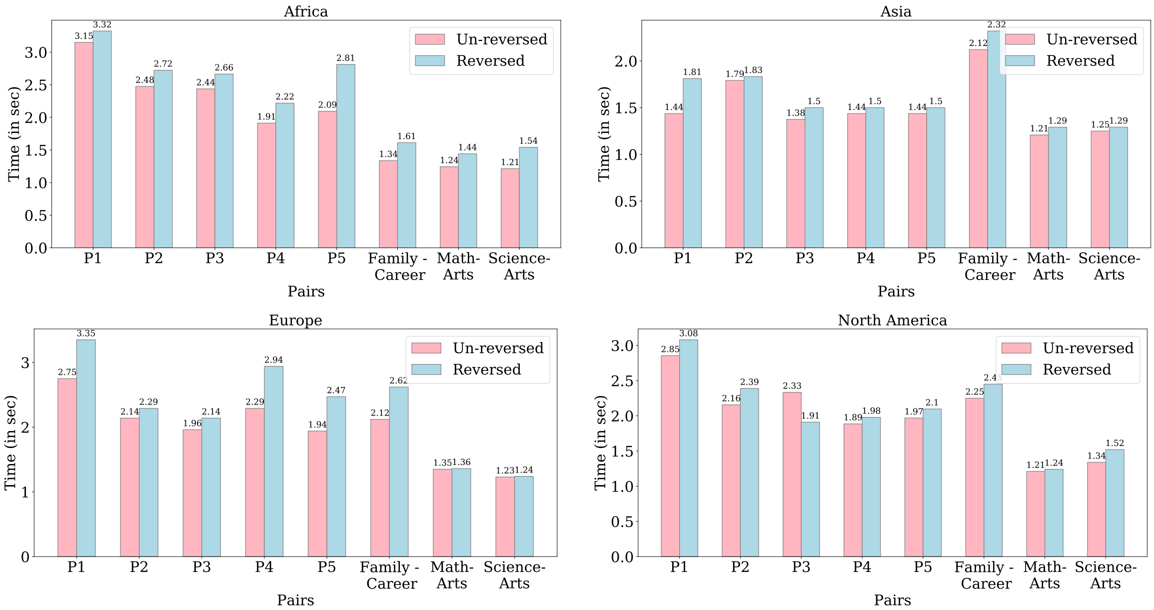


Figure 3: Human validation results across regions. 'Unreversed' refers to bias dimensions with the same gender associations as our topic pairs, 'Reversed' refers to bias dimensions with the opposite gender associations.

REGION	F-M TOPIC PAIR
Africa	Parenting and family relationships-Nollywood Actress and Movies (P1) Marriage and relationships - Sports and Football (P2) Womens' lives and successes - Fashion and Lifestyle (P3) Music - Social Media (P4) Dating and relationships advice - Religious and Spiritual growth (P5)
Asia	Hotel royalty - Political leadership in India (P1) Healthy eating habits for children - Sports and Soccer (P2) Royal wedding plans - Social Media platforms for video sharing (P3) Royal wedding plans - Religious devotion and spirituality (P4) Marriage - Bollywood actors and films (P5)
Europe	Education - Music (P1) Comfortable hotels - Political decision and impact on society (P2) Luxury sailing - UK Government Taxation policies (P3) Obituaries and Genealogy - Christian Theology and Practice (P4) Fashion and style - Christian theology and practice (P5)
North America	Online Dating for Singles - Religion and Spirituality (P1) Fashion and Style - Reproductive Health (P2) Education and achievements - Reinsurance and capital markets (P3) Family dynamics and relationships - Nike shoes and fashion (P4) Reading and fiction - Cape Cod news (P5)
Oceania	Family relationships - Religious beliefs and figures (P1) Woodworking plans and projects - Music record and Artists (P2) Weight loss and nutrition for women - Building and designing boats (P3) Exercises for hormone development - Superheroes and their Universes (P4) Kids' furniture and decor - Building and designing boats (P5)

Table 3: Top five region-aware topic pairs for F and M for each region using en embedding-based approach.

North America. Several topics like *'family and parenting'* are commonly associated with females across different regions while *'politics'* is associated with males. To this end, we compute the top uni-grams and bi-grams for topic pairs that are common across regions in Appendix E.

4.4.4 Human Validation Results

Fig 3 shows response times for top five topic pairs in each region for un-reversed and reversed scenarios. Larger time differences indicate more bias, suggesting that the pair could be a potential gender bias dimension for that region. If un-reversed time is lower, it suggests a stronger association of T_f with the F group and T_m with the M group. The family-career pair was also surveyed as a standard WEAT bias dimension. Please refer to Table 3 for topic pair numbers (P1...P5) of each topic pair.

As expected, the *family–career* pair shows the highest bias across all three general IAT topic pairs. There are smaller differences among the other *math–arts* and *science–arts*. We also note that some pairs, such as *dating and relationships advice–religious and spiritual growth* (P5) for Africa, *hotel royalty–political leadership in India* (P1) for Asia, *obituaries and geneology–Christian theology* (P4), *education–music* (P1), and *fashion and style–Christian theology* (P5) for Europe, and *online dating–religion and spirituality* (P1), *fashion and style–reproductive health* (P2) for North America have differences higher than those for *family–career* in the respective regions, indicating that the participants associated more biases on our uncovered bias dimensions than the existing one in WEAT. These findings support our hypothesis that gender bias dimensions vary across regions and also bring preliminary evidence that the region-aware bias dimensions we uncover are in line with

Region	F-M topic pair	Reddit	UN General Debates
Africa	Parenting and family relationships-Nollywood Actress and Movies	0.500	0.979
	Marriage and relationships - Sports and Football	-0.051	0.224
	Womens' lives and successes - Fashion and Lifestyle	0.480	0.493
	Music - Social Media	1.894	1.721
	Dating and relationships advice - Religious and Spiritual growth	1.475	1.061
Asia	Hotel royalty - Political leadership in India	1.365	1.768
	Healthy eating habits for children - Sports and Soccer	0.006	-0.068
	Royal wedding plans - Social Media platforms for video sharing	1.05	1.393
	Royal wedding plans - Religious devotion and spirituality	1.183	1.335
	Marriage - Bollywood actors and films	1.543	0.918
Europe	Education - Music	1.261	1.920
	Comfortable hotels - Political decision and impact on society	0.324	0.485
	Luxury sailing - UK Government Taxation policies	1.232	1.558
	Obituaries and Genealogy - Christian Theology and Practice	0.001	-0.405
	Fashion and style - Christian theology and practice	1.730	1.028
North America	Online Dating for Singles - Religion and Spirituality	1.728	1.830
	Fashion and Style - Reproductive Health	1.723	1.095
	Education and achievements - Reinsurance and capital markets	-0.148	-0.364
	Family dynamics and relationships - Nike shoes and fashion	0.109	0.691
	Reading and fiction - Cape Cod news	0.251	0.506
Oceania	Family relationships - Religious beliefs and figures	0.305	0.267
	Woodworking plans and projects - Music record and Artists	0.056	-0.258
	Weight loss and nutrition for women - Building and designing boats	0.336	0.582
	Exercises for hormone development - Superheroes and their Universes	-0.05	-0.07
	Kids' furniture and decor - Building and designing boats	0.612	0.524

Table 4: Region-aware WEAT-based evaluation on Reddit and UNGDC. Highest scores are highlighted for each dataset across regions.

the human perception of bias in those regions. We also find that all the regions have biases that conform to our topic pairs gender association except P3: *education–reinsurance and capital markets* in North America, where the associated bias is negative. These findings confirm that topic pairs indeed differ across regions and that these differences must be taken into consideration when identifying and evaluating biases.

5 WEAT-based Evaluation Using Region-aware Topic Pairs

To measure biases in different data domains and regions, we extract region-aware topics using the GeoWAC dataset which spans Common Crawl separated by regions, and create a WEAT-style evaluation setup using these topics.

Data. We consider two datasets: (i) Reddit data and (ii) UN General Debates (Baturo et al., 2017). The Reddit data consists of data from subreddits corresponding to specific regions: `r/asia`, `r/africa`, `r/europe`, `r/northamerica`, and `r/oceania`. We use the official Reddit API to extract data, consisting of 500 top posts[2] from each subreddit. The posts are pre-processed to remove URLs and signs, and each post contains at least 30 words. The UN General Debate Corpus (UNGDC) includes texts of General Debate statements from 1970 to 2016. These statements, similar to annual legislative state-of-the-union addresses, are delivered by leaders and senior officials to present their government's perspective on global issues. We filter the countries for each region and extract 500 data points per region, maintaining equal representation across region.[3]

Method. WEAT tests consist of keywords corresponding to each attribute and topic word sets like family-career and male-female terms. To create a similar setup, we utilize KeyBERT (Grootendorst, 2020) to gather top topic representations corresponding to each topic extracted from GeoWAC. For male/female terms, we use the same representative words from WEAT. To further make it specific to a particular region, we employ GPT-4 (OpenAI et al., 2024) to generate common male/female names used in the regions and add them to the list. We provide the list of words in Table 12 of Appendix F. We use fastText (Bojanowski et al.,

[2]The Official Reddit API has rate limits, therefore 500 top posts from each subreddit ensures an equal number of examples for each region.

[3]Oceania has limited available countries in UNGDC, hence the adherence to 500 data points for each region.

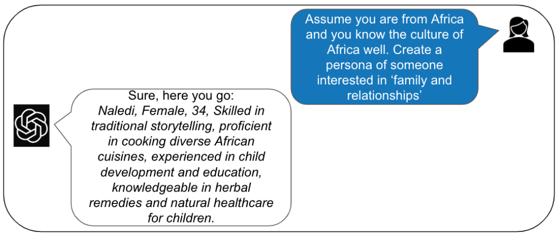


Figure 4: Example Prompt for Persona Generation

2017)[4] embedding algorithm to generate embeddings of the lists and compute the distances between the topic words and male/female terms (like WEAT).

Results. In Table 4, a high number of positive scores indicates the presence of a positive bias for our region-aware topic pairs. This means a presence of bias with the same gender association as our topic pairs, for example, if *'music-social media'* is F-M topic pair in Africa according to our study, a positive score on the Reddit dataset means that bias is in the same direction. The few negative scores in the table indicate that these topic pairs do not conform to the same gender bias associations. However, a higher negative magnitude also shows the presence of bias, therefore, these topic pairs are still important.

Additionally, magnitudes of many scores are high (> 0.5) which shows a high presence of bias (positive/negative) corresponding to the topic pairs. We highlight the top-scoring bias topic pairs for each region in Table 4. High-bias topics vary for each region based on the dataset. For example, *'music-social media'* has the highest bias in Africa for both datasets, however for Asia, we find that *'marriage - Bollywood actors and films'* and *'Hotel royalty - Political leadership in India'* are the topic pairs with the highest biases in Reddit and UN General Debates respectively, indicating that biased topic pairs may be domain-dependent.

Using our topic pairs in this WEAT-style evaluation setup provides an illustration of how our automatically curated region-aware bias dimensions can be used in designing a region-aware bias evaluation test. It also shows the effectiveness of our region-aware bias topic pairs in capturing the dimensions that are likely to contain gender biases across regions.[5]

[4]We choose fastText because it allows to extract embeddings of words that are not present in the target text (as our topics are derived from GeoWAC).

[5]Note that our topic pairs although extracted from GeoWAC are somewhat generalizable to other datasets like Reddit and UNGDC, we do not claim that these are best topic pairs achievable as topic pairs are also data dependent, but we can use our methodology to extract bias topic pairs that may exist in specific datasets.

6 Alignment of Region-Aware Bias Dimensions with LLM outputs

To understand if LLMs generate similar biases as our region-aware bias topic pairs, we devise a persona generation task by LLMs. We prompt the LLM to output personas interested in different 'topics' from the topic pairs that we extract. Fig 4 shows an example of the prompt given to an LLM to generate personas. We experiment with different LLMs: `GPT-3.5` (Brown et al., 2020), `GPT-4`, `Mistral-7b-Instruct` (Jiang et al., 2023), `Claude-3 Sonnet`,[6] and `Gemini-Pro` (Team et al., 2024). Many studies use LLM-generated personas for multi-agent interactions in different settings in societies (Park et al., 2023; Zhou et al., 2024). But, if an LLM generates biased personas, for example, a female persona takes care of children, and a male persona is strong and takes care of emergencies, this would lead to further biases in consequent tasks. Therefore, we employ persona generation to check for the presence of any biases in the personas created by LLMs. To measure biases, we find the number of matched LLM output persona genders to the genders of our topic pairs. We average our results over seven runs.

Results. We plot the results of persona gender mismatched by LLMs in Fig 5. The y-axis shows % mismatch between the LLM generated persona gender and the gender of the topic in our topic pair. For example, a mismatch is when LLM outputs a persona with 'female' for *Politics* in Asia, which is a 'male' topic according to our findings. Regions with high representation: North America, Europe and Asia have fewer mismatches, with North America having the lowest mismatch. Conversely, less represented regions like Africa and Oceania show higher mismatch rates. Among models, `Mistral-7b` (7B) has the highest mismatch rate while `Gemini-Pro` (50T) has the least, which may stem from varying model sizes. Overall, all the models exhibit similar mismatch trends for both highly and less represented regions. Fewer mismatches in highly-represented regions show the importance of evaluation using region-specific topic pairs. Higher mismatches in underrepresented regions like Africa and Oceania suggest LLMs don't mimic these areas' biases, which can be beneficial. However, due to growing research on LLMs' cultural alignment, a more precise, region-specific bias evaluation metric becomes essential.

[6]https://claude.ai/

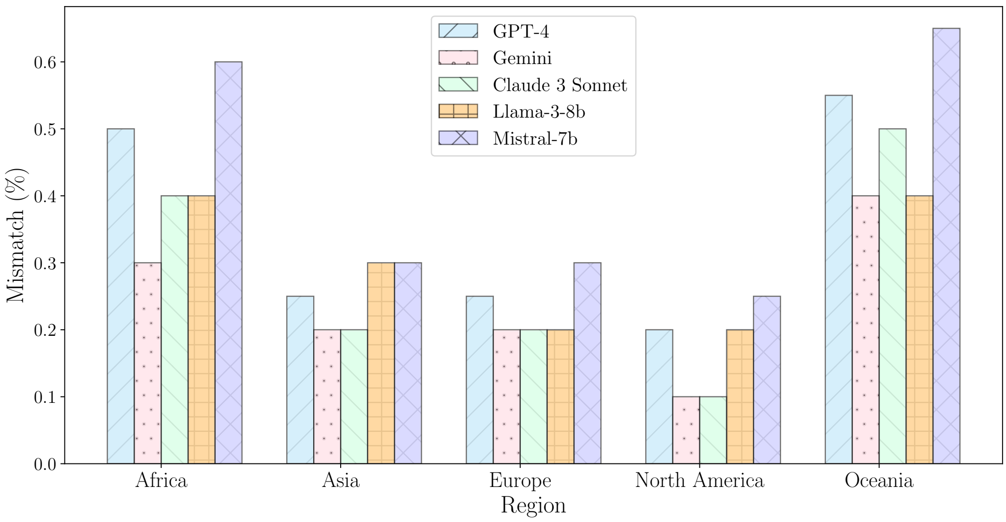


Figure 5: Bias Evaluation of LLM outputs using region-aware bias topic pairs through 'persona generation'.

7 Related Work

IAT is one of the earliest and well-known method for measuring implicit social biases in humans (Greenwald et al., 1998). Inspired by the IAT, WEAT uses word embeddings to measure biases in text (Caliskan et al., 2017). Another extension of WEAT is the Sentence Embedding Association Test (SEAT), which measures biases at the sentence level (May et al., 2019). Additionally, various bias detection measures in NLP focus on post-training model predictions, such as gender swapping (Stanovsky et al., 2019). Moreover, there are specific gender bias evaluation test sets in tasks like coreference resolution (Rudinger et al., 2018; Zhao et al., 2018; Webster et al., 2018) and sentiment analysis (Kiritchenko and Mohammad, 2018b).

Several studies have emphasized the significance of considering cultural awareness in the study of social phenomena. The demographics of individuals can shape their worldviews and thoughts (Garimella et al., 2016), potentially influencing their language preferences and biases in daily life. Notably, some studies have observed a bias towards Western nations in current LLMs (Dwivedi et al., 2023). Recent research has focused on cross-cultural aspects of LLMs, including aligning them with human values from different cultures (Glaese et al., 2022; Sun et al., 2023) and exploring them as personas representing diverse cultures (Gupta et al., 2024). To the best of our knowledge, no previous work has proposed a data-dependent approach to extract region-aware bias topics. Given the known biases in LLMs, a region-specific metric could greatly lead to an accurate evaluation of biases. This research holds significant importance in addressing cross-cultural biases effectively.

8 Conclusion

In this paper, we proposed a bottom-up approach using data to identify region-aware topic pairs that capture gender biases across different regions. Our human evaluation results demonstrated the validity of our proposed region-aware dimensions.

We employed a region-aware WEAT-based evaluation setup to assess biases in two additional datasets: Reddit and the UN General Debate Corpus. The presence of region-specific biases in these datasets underscores the importance of a region-aware bias evaluation metric. Additionally, when examining LLM outputs against the gender associations in our region-aware bias topic pairs, we found that biases align closely for three highly represented regions: North America, Europe, and Asia. This emphasizes the value of region-aware topic pairs in bias evaluation of LLMs. Conversely, biases do not align well for Africa and Oceania, indicating that LLMs do not adopt these regions' specific biases–a potential benefit. Yet, it also highlights the 'cultural alignment' issue in LLMs. More research on the cultural alignment of LLMs underlines the need to consider region-specific bias topic pairs for all regions in future studies.

Future work includes incorporating testing different model/dataset combinations and topic-pair dependency on data. We aim to study biases in different languages and explore region-aware bias mitigation techniques.

9 Limitations

We utilized the GeoWAC corpus as our sole data source for extracting topic pairs from various regions. However, we acknowledge the importance of incorporating additional datasets in our future work. Additionally, our WEAT-based evaluation was conducted on relatively smaller datasets. So, we intend to conduct further analysis on a larger dataset to ensure a comprehensive evaluation based on WEAT.

Our study did not account for different languages due to the diverse linguistic landscape of the regions (continents) included in our study. However, the significance of conducting a more detailed analysis to examine variations among different countries would be interesting.

Unfortunately, we encountered difficulties in finding participants from Oceania for human validation. Moving forward, we plan to include insights and findings from Oceania and incorporate a larger population to ensure a more comprehensive human validation.

10 Ethical Considerations

When developing our region-aware topic pairs, it is essential to consider the ethical implications. Since we utilize a much broader aspect of culture, i.e. continents to distinguish among cultures, the region-aware topic pairs we extract may not translate to cultures of communities that are not well-represented in models. Hence, it is important that we utilize topic pairs carefully.

It has been found that AI models often tend to output responses that are Western, educated, industrialized, rich, and democratic (Henrich et al., 2010). In our experiments, we see LLMs also generate biases having the highest alignment with the West. Therefore, LLM experiments also need to be utilized carefully.

Our Reddit data for the region-aware evaluation metric may contain offensive content. However, we have anonymized the data (removed the usernames).

References

Alexander Baturo, Niheer Dasandi, and Slava J. Mikhaylov. 2017. Understanding state preferences with text as data: Introducing the un general debate corpus. *Research & Politics*, 4(2):2053168017712821.

Christopher Bishop. 2006. Pattern recognition and machine learning. *Springer google schola*, 2:5–43.

National Geographic Education Blog. 2017. What continent do you think they are from? drawing humans to reveal internalized bias.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

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

Taylor J Damann, Jeremy Siow, and Margit Tavits. 2023. Persistence of gender biases in europe. *Proceedings of the National Academy of Sciences*, 120(12):e2213266120.

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.

Jonathan Dunn and Ben Adams. 2020a. Geographically-balanced gigaword corpora for 50 language varieties. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2528–2536.

Jonathan Dunn and Ben Adams. 2020b. Mapping languages and demographics with georeferenced corpora. *arXiv preprint arXiv:2004.00809*.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 311–319, Online only. Association for Computational Linguistics.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements.

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.

Stephanie D Grimm and A Timothy Church. 1999. A cross-cultural study of response biases in personality measures. *Journal of Research in Personality*, 33(4):415–441.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of llm personality.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Mengyin Jiang, Shirley KM Wong, Harry KS Chung, Yang Sun, Janet H Hsiao, Jie Sui, and Glyn W Humphreys. 2019. Cultural orientation of self-bias in perceptual matching. *Frontiers in Psychology*, 10:1469.

Svetlana Kiritchenko and Saif Mohammad. 2018a. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018b. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Claudia Malzer and Marcus Baum. 2020. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE.

Hazel Rose Markus and Shinobu Kitayama. 2003. Models of agency: sociocultural diversity in the construction of action.

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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Dugan Nichols. 2011. Men and must-have shoes. *Retrieved from.*

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Rebecca Ross Russell. 2010. *Gender and jewelry: A feminist analysis*. Rebecca Ross Russell.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gemini Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontanon, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Shane Gu, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Kiran Vodrahalli, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjösund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Zeyncep Cankara, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Lora Aroyo, Zhufeng Pan, Zachary Nado, Jakub Sygnowski, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Yamini Bansal, Xavier Garcia, Mehran Kazemi, Piyush Patil, Ishita Dasgupta, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Qingze Wang, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Raphaël Lopez Kaufman, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Chris Welty, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Renshen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Adam Iwanicki, Alejandro Lince, Alexander Chen, Christina Lyu, Carl Lebsack, Jordan Griffith, Meenu Gaba, Paramjit Sandhu, Phil Chen, Anna Koop, Ravi Rajwar, Soheil Hassas Yeganeh, Solomon Chang, Rui Zhu, Soroush Radpour, Elnaz Davoodi, Ving Ian Lei, Yang Xu, Daniel Toyama, Constant Segal, Martin Wicke, Hanzhao Lin, Anna Bulanova, Adrià Puigdomènech Badia, Nemanja Rakićević, Pablo Sprechmann, Angelos Filos, Shaobo Hou, Víctor Campos, Nora Kassner, Devendra Sachan, Meire Fortunato, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Ying Xu, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Ananth Agarwal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Alanna Walton, Alicia Parrish, Mark Epstein, Sara McCarthy, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

REGION	COUNTRY	#EXAMPLES
Africa	Nigeria	3,153,761
	Mali	660,916
	Gabon	645,769
Asia	India	12,327,494
	Singapore	6,130,047
	Philippines	3,166,971
Europe	Ireland	8,689,752
	United Kingdom	7,044,434
	Spain	465,780
North America	Canada	7.965,736
	United States	8,521,094
	Bermuda	244,500
Oceania	New Zealand	94,476
	Palau	486,437
	Vanuatu	165,355

Table 5: Region-specific details in GeoWAC

A GeoWAC dataset details

Table 5 contain the total number of examples per country in a region. We consider the top three countries with the highest number of examples per region.

B F-M Dataset statistics

Table 6 displays the total number of examples from female and male groups per region for the region-specific F-M dataset.

C Cultural differences in biases using WEAT

Table 7 shows the WEAT scores for all WEAT dimensions defined in (Caliskan et al., 2017). We find that scores and p-values differ across regions for different dimensions. High bias dimensions differ across regions, hence it is important to consider region-specific topic pairs.

REGION	TOTAL	#FEMALE	#MALE
Africa	57895	20153	37742
Asia	56877	21400	35477
Europe	59121	21049	38072
North America	70665	27627	43038
Oceania	62101	25951	36150

Table 6: F-M dataset statistics for regions (Total refers to the total number of examples in each region, therefore, total = #female + #male)

D Region-wise topic lists in GeoWAC

Table 8 displays a comprehensive list of topics for female and male groups across all regions.

E Unigram/Bigram Analysis

Table 10 shows the unigrams and bigrams of common topics with different gender associations. We find that 'fashion' is highly associated with shoes when it is a male topic in Africa, whereas in Europe and North America, it is mostly associated with accessories like sunglasses, rings, etc. This shows the typical association of women with jewelry and men with shoes (Russell, 2010; Nichols, 2011). In the case of 'Music', we see that unigrams and bigrams pertaining to Africa contain words related to hip-hop music and artists. For Europe, we find location references and metal music. And finally, Oceania shows references of jazz and rock. We do not find any obvious gender associations in the analysis of the music topic. Table 11 provides a unigram/bigram analysis of topics that are commonly associated with a specific gender across regions. For *parenting and family relationships*, Africa has mentions of children, while Asia and Oceania contain mentions of family events, etc. In North America, we mostly find text about maintaining health in families. For *religion and spirituality*, the unigrams/bigrams are mostly about Jesus and Christianity across regions. For *politics*, we find mentions of specific regions, as expected. *Education* topic is more about being successful in Europe, where it is about degrees in North America. Finally, 'social media' trends are mostly similar. Overall for topics with same gender associations across regions, do not have stark differences.

F WEAT-based evaluation setup details

For male/female terms, we use the same representative words from WEAT: *brother, father, uncle, grandfather, son, he, his, him, man, boy, male* for male and *sister, mother, aunt, grandmother, daughter, she, hers, her, woman, girl, female for female*.

Target words - Attribute words	Region	Region-specific p-value	Region-specific WEAT score	Original WEAT score, p-value
Male names vs Female names - career vs family	Africa	0.016	1.798	1.81, 0.001
	Asia	0.007	1.508	
	North America	0.04	1.885	
	Europe	$6 \cdot 10^{-4}$	1.610	
	Oceania	0.03	1.727	
Math vs Arts - Male vs Female terms	Africa	0.003	1.429	1.06, 0.018
	Asia	0.045	1.187	
	North America	0.007	0.703	
	Europe	0.005	0.334	
	Oceania	0.03	1.158	
Science vs Arts - Male vs Female terms	Africa	0.048	1.247	1.24, 0.01
	Asia	0.004	0.330	
	North America	$1 \cdot 10^{-5}$	0.036	
	Europe	$1 \cdot 10^{-7}$	-0.655	
	Oceania	$2 \cdot 10^{-4}$	0.725	
Young people names vs old people names - pleasant vs unpleasant	Africa	$3 \cdot 10^{-5}$	0.855	1.21, 0.01
	Asia	$4 \cdot 10^{-4}$	0.917	
	North America	0.032	1.325	
	Europe	0.009	0.917	
	Oceania	0.014	0.947	
European American names vs African American names - pleasant vs unpleasant	Africa	$1 \cdot 10^{-5}$	0.008	1.28, 0.001
	Asia	$1 \cdot 10^{-6}$	-0.453	
	North America	0.009	1.29	
	Europe	0.001	0.617	
	Oceania	$1 \cdot 10^{-4}$	0.492	
Instruments vs Weapons - pleasant vs unpleasant	Africa	0.03	1.443	1.53, $< 10^{-7}$
	Asia	0.009	1.001	
	North America	0.01	1.202	
	Europe	0.02	1.21	
	Oceania	0.001	0.951	
Flowers vs Insects - pleasant vs unpleasant	Africa	0.002	0.312	1.5, $< 10^{-7}$
	Asia	0.009	0.869	
	North America	0.003	0.382	
	Europe	0.001	0.332	
	Oceania	0.009	0.660	
Mental disease vs Physical disease - temporary vs permanent	Africa	0.008	0.835	1.38, 0.01
	Asia	0.02	1.201	
	North America	0.008	0.692	
	Europe	0.04	1.382	
	Oceania	0.009	1.620	

Table 7: Region-wise WEAT scores and p-values across all dimensions specific in WEAT using word2vec. Negative scores are highlighted. We compare our region specific scores and p-values with the scores and p-values of the Original paper by (Caliskan et al., 2017)

```
[ ] # System prompt describes information given to all conversations
    system_prompt = """
    <s>[INST] <<SYS>>
    You are a helpful, respectful and honest assistant for labeling topics.
    <</SYS>>
    """

[ ] # Example prompt demonstrating the output we are looking for
    example_prompt = """
    I have a topic that contains the following documents:
    - Traditional diets in most cultures were primarily plant-based with a little meat on top, but with the rise of industrial style meat
    production and factory farming, meat has become a staple food.
    - Meat, but especially beef, is the word food in terms of emissions.
    - Eating meat doesn't make you a bad person, not eating meat doesn't make you a good one.

    The topic is described by the following keywords: 'meat, beef, eat, eating, emissions, steak, food, health, processed, chicken'.

    Based on the information about the topic above, please create a short label of this topic. Make sure you to only return the label and nothing more.

    [/INST] Environmental impacts of eating meat
    """

[ ] # Our main prompt with documents ([DOCUMENTS]) and keywords ([KEYWORDS]) tags
    main_prompt = """
    [INST]
    I have a topic that contains the following documents:
    [DOCUMENTS]

    The topic is described by the following keywords: '[KEYWORDS]'.

    Based on the information about the topic above, please create a short label of this topic. Make sure you to only return the label and nothing more.
    [/INST]
    """
```

Figure 6: `Llama2` prompt

Region	Female	Male
Africa	Credit cards and finances, Royalty and Media, Trading strategies and market analysis, Dating and relationships guides, Parenting and family relationships, Fashionable Ankara Styles, women's lives and successes, online dating	Fashion and Lifestyle, Male enhancement and sexual health, Nollywood actresses and movies, Nigerian politics and government, Essay writing and research, Medical care for children and adults, Journalism and Media Conference, Music industry news and releases, Football league standing and player performances, Academic success and secondary school education, Religious inspiration and spiritual growth, Economic diversification and Socio-economic development
Asia	Hobbies and Interests, Healthy eating habits for children, Social media platforms, Royal wedding plans, Online Dating and Chatting, Adult Services, Gift ideas for Valentine's Day	DC comic characters, Mobile Application, Philippine Politics and Government, Sports and Soccer, Career, Bike enthusiasts, Artists and their work, Youth Soccer Teams, Career in film industry, Political leadership in India, Bollywood actors and films, Religious devotion and spirituality, Phone accessories
Europe	Pets and animal care, Fashion and Style, Education, Obituaries and Genealogy, Luxury sailing, Traveling, Energy and climate change, Family and relationships, Pension and costs, Tech and business operations, Dating, Comfortable hotels, Government transportation policies	Political developments in Northern Ireland, Christian Theology and Practice, Crime and murder investigation, EU Referendum and Ministerial Positions, Criminal Justice System, Israeli politics and International relations, Cancer and medications, UK Government Taxation policies, Art Exhibitions, Political decision and impact on society, Music Gendres and artists, Medical specialties and university training, Political discourse and parliamentary debates
North America	Pets, Cooking: culinary delights and chef recipes, Fashion and style, Family dynamics and relationships, Reading and fiction, Scheduling and dates, Life and legacy of Adolf Hitler, Gender roles and inequality, Education and achievements, Online dating for singles, Luxury handbags, Footwear and Apparel brands, Essay writing and literature	Civil War and history, Middle East conflict and political tensions, Movies and filmmaking, Political leadership and party dynamics in Bermuda, Rock Music and songwriting, Wartime aviation adventures, Religion and Spirituality, Reproductive health, Reinsurance and Capital markets, Nike shoes and fashion, Cape Cod news, NHL players
Oceania	Cooking and culinary delights, Romance, Weight loss and nutrition for women, Water travel experience, Woodworking plans and projects, Time management and productivity, Inspiring stories and books for alleges, Sexual violence and abuse, Car insurance, Exercises for hormone development, kid's furniture and decor	Harry Potter adventures, Art and Photography, Superheroes and their Universes, Music recording and Artists, Football in Vanuatu, Pet care and veterinary services, Building and designing boats, Religious beliefs and figures, Fashion, Classic movie stars, Men's hairstyle and fashion, Male sexual health and supplements

Table 8: Region-wise topics for female and male.

We also utilize GPT-4 to output the ten most common male/female names specific to each region. We provide the lists of word belonging to each topic in Table 12.

G Paired-list for F-M datasets

Here is the list of the 52 pairs used to create the F-M datasets per region:
[monastery, convent], [spokesman, spokeswoman], [Catholic priest, nun], [Dad, Mom], [Men, Women], [councilman, councilwoman], [grandpa, grandma], [grandsons, granddaughters], [prostate cancer, ovarian cancer], [testosterone, estrogen], [uncle, aunt], [wives, husbands], [Father, Mother], [Grandpa, Grandma], [He, She], [boy, girl], [boys, girls], [brother, sister], [brothers, sisters], [businessman, businesswoman], [chairman, chairwoman], [colt, filly], [congressman, congresswoman], [dad, mom], [dads, moms], [dudes, gals], [ex girlfriend, ex boyfriend], [father, mother], [fatherhood, motherhood], [fathers, mothers], [fella, granny], [fraternity, sorority], [gelding, mare], [gentleman, lady], [gentlemen, ladies], [grandfather, grandmother], [grandson, granddaughter], [he, she], [himself, herself], [his, her], [king, queen], [kings, queens], [male, female], [males, females], [man, woman], [men, women], [nephew, niece], [prince, princess], [schoolboy, schoolgirl], [son, daughter], [sons, daughters], [twin brother, twin sister].
Each pair in the above is denoted as a [male, female] pair.

H Llama 2 prompt for topic modeling

The prompt scheme for `Llama2` consists of three prompts: (1) System Prompt: a general prompt that describes information given to all conversations, (2) Example Prompt: an example that demonstrates the output we are looking for, and (3) Main Prompt: describes the structure of the main question, that is with a given set of documents and keywords, we ask the model to create a short label for the topic. Fig 6 displays the three prompts as used in the code.

REGION	FEMALE TOPICS	MALE TOPICS
Africa	Credit card-based financial services Royalty and femininity Financial trading Dating guides Motherhood and parenting	Fashion - footwear and celebrities Male enhancement and sexual health Nollywood Nigerian politics Academic writing
Asia	Hobbies Food and nutrition Social media platforms and content creation Royal weddings Online social interaction and dating	Superhero comic books Mobile applications Philippines politics and people Sports Career
Europe	Pets Fashion Education Deaths and funerals Luxury yachting and sailing	Irish politics Christianity Law enforcement and crime EU and Brexit Criminal justice system
North America	Pets Cooking and Food Fashion Family and relationships Reading novels	Civil War Military Middle Eastern politics and conflicts Movies and direction Bermuda politics Rock music
Oceania	Food and eating habits Romance and emotions Weight loss and nutrition Boat and sailing experience Woodworking and carpentry	Harry Potter Artistic expressions Superheroes of Marvel and DC Albums, songs and artists Vanuatu Football

Table 9: Topic labels by `gpt-4`, see Table 2 for comparison with `Llama2` topic labels

I Topic Cluster Labels using other LLMs

We use `Llama2` to fine-tune our topics to label them for better coherence in our paper. However, we also experiment with `GPT-4` and arrive at similar topics in Table 9. (see Table 2 for comparison with `Llama2` topic labels).

J Topic Word Clusters Example - Africa

Here, we provide an example of how topics look in our data. In Fig 7, we provide word clusters of topics from Africa. The word clusters contain the top 10 words from each topic in Africa. We find that topic labels by `Llama2` are coherent in terms of top topic words.

K Region specific BERTs to identify top words in F/M direction

To motivate our case to investigate differences in biases across regions, we use BERT to compute the top words corresponding to the *she-he* axis in the embedding space. BERT is a pre-trained transformer-based language model that consists of a set of encoders. As a motivation experiment to identify differences in the contextual embedding space for different regions, we fine-tune BERT with the masked language modeling task (no labels) for each region separately. For a given word, we compute its embeddings by averaging out all sentence embeddings where it occurs across the dataset.Similarly, we compute embeddings for all words in the dataset. The tokenized input goes through the BERT model and we take the hidden states at the end of the last encoder layer (in our case, BERT-base, i.e. 12 encoder layers) as sentence embeddings. We identify the top words with the highest projection across the *she-he* axis in the region-specific datasets. If we find differences in the top words across regions, it is possible that dominating bias topics vary by region as well. Fig 8 shows the top words closest to 'she' and 'he' contextual embeddings in our data for each region. We find that top words differ quite a bit across different regions. We find many differences in the top F (close to she) and M (close to he) words across regions. Some top F words are soprano, archaeological (Africa); graduate, secretary (Asia); innovative, graphics (Europe); poets, sentiments (NA); and arts, sleep (Oceania). Some top M words are history, leading (Africa); astronomer, commissioners (Asia); honorary, songwriters (Europe); owner, hospital (NA); and wrestlemania, orbits (Oceania). Gender-neutral words such as poets, secretaries, astronomers, commissioners, songwriters, owners, and so on are closer to either the she or he axes. Although comparable to the findings of (Bolukbasi et al., 2016), the variances among regions inspire us to look deeper into the data to arrive at culture-specific bias themes.

L Implementations details

For training our `Bertopic` model, we use `Google Colab's Tesla T4 GPU`, and it takes 15 min to run topic modeling for a region-specific F-M dataset. Region-specific BERTs are run on NVIDIA RTX2080 GPUs. Each BERT training experiment takes 1 GPU hour. For our LLM experiment, we used NVIDIA-A40 for `Mistral-7b-Instruct` and `Llama-3-8b` for an hour. We do not use any GPUs for `GPT-4`, `Claude-3-Sonnet` and `Gemini-Pro`.

L.1 Bertopic

We use Bertopic's default models: `SBERT` (Reimers and Gurevych, 2019) to contextually embed the dataset, `UMAP` (McInnes et al., 2018) to perform

Topic	Region	Unigrams	Bigrams
Fashion and lifestyle	Africa (male)	march, outlet, air, max, tods, man, said, pas, cher, people	air max, pas cher, princess j, roshe run, nike air, tods outlet, j march, roger vivier, posts email, notify new
	Europe (female)	one, women, fashion, like, new, look, make, hair, girl, dress	oakley sunglasses, louis vuitton, red carpet, new york, fashion model, engagement rings, per cent, year old, christian louboutin, diamond ring
	North America (female)	one, love, like, little, new, made, time, get, make, women	s cooper, cooper main, t shirt, new york, little girl, men women, look good, main store, years ago, check out
Music	Africa (female)	music, song, album, new, video, single, one, singer, also, songs	music industry, hip hop, record label, single titled, new single, chris brown, tiwa savage, ice prince, kanye west, niegrian music
	Europe (male)	man, single, stage, years, world, many, metal, guitar, solo, irish	year shelfmark, black metal, time exercise, musical content, dundee repertory, singer songwriter, edinburgh year, zumba days, male vocalists, millions men
	Oceania (male)	music, album, new, songs, band, first, time, jazz, released, rock	new york, elizabth ii, debut album, years later, big band, rock roll, first time, studio album, los angeles, solo artist

Table 10: Common topics with different gender associations across regions

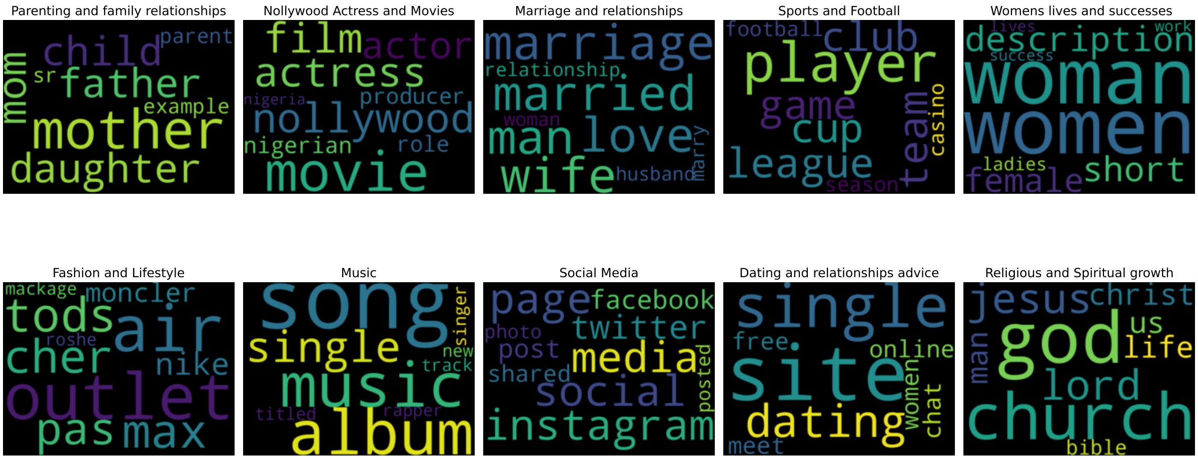


Figure 7: Topic Word Clusters - Africa

dimensionality reduction, `HDBSCAN` (Malzer and Baum, 2020) for clustering to perform topic modeling. We choose the embedding model `BAAI/bge-small-en` from *Huggingface* (Wolf et al., 2019). We set `top_n_words` to 10 and `verbose` as True and set the `min_topic_size` to 100 for the `Bertopic` model. Finally, we use Bertopic's official library to implement the model.

L.2 Llama2

We use `Llama2` to finetune the topics to give shorter labels for each topic. We set the `temperature` to 0.1, `max_new_tokens` to 500 and `repetition_penalty` to 1.1. We utilize Bertopic's built-in representation models to use `Llama2` in our topic model.

L.3 LLM experiment

For `GPT-4`, and `Mistral-7b-Instruct` and `Llama-3-8b`, we utilize the Microsoft Azure API[7], huggingface[8], and huggingface[9] for inference respectively. We use a temperature 0.8 for all models. For `Gemini-Pro` and `Claude-3-Sonnet`, we use the available chat interface.

L.4 Region-specific BERT

We use the uncased version BERT (Devlin et al., 2019) for our region-specific BERT model trained for the MLM objective. We use a batch size of 8, a learning rate of $1 \cdot 10^{-4}$, and an AdamW optimizer

[7] https://learn.microsoft.com/en-us/rest/api/azure/

[8] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[9] https://huggingface.co/meta-llama/Meta-Llama-3-8B

Topic	Region	Unigrams	Bigrams
Parenting and family relationships	Africa (female)	child, registration, form, information, sent, women, foster, best, catholic, women	registration form, form information, child assigned, surgery doctors, new catholic, catholic women, contemporary challenge, best everything, foster short, doctors clinic
	Asia (female)	year, old, weekly, fortnightly, clicking, create, alert, state, 1, terms	year old, weekly fortnightly, create alert, stated agree, conditions acknowledge, finals appearances, together playing, dial guarded, came work, outlet jackets
	North America (female)	women, healthday, loss, three, worked, closely, together, she, elegant, dignified	three women, women worked, closely together, elegant dignified, very pleasant, soft spoken, women men, healthday reporter, tuesday march, participate more
	Oceania (female)	laurel, school, moved, one, day, royal, wedding, house, sister, hopefully	moved one, royal wedding, laurel school, 1 california, weeks dad, high school, one hopefully, nobody knew, sister means, fu school
Religion and Spirituality	Africa (male)	god, man, church, one, life, people, jesus, us, lord, christ,	short description, jesus christ, man god, holy spirit, god said, thank god, bible says, catholic church, today god, every man
	Asia (male)	life, jesus, us, church, one, man ,lord, said, father, christ	holu spirit, jesus christ, pope francis, brothers sisters, son god, men women, holy father, opus dei, eternal life, paul ii
	Europe (male)	god, one, jesus, church, life, people, father, man , said, christ	jesus christ, son man, catholic church, holy spirit, men women, said him, holy father, john paul, jesus said, word god
	North America (male)	god, jesus, one, man, us, life, would, christ, lord, people	recognizable cheering, section league, jesus christ, exact synonyms, past years, god said, years before, thanks mostly, mostly steph, father dell
	Oceania (male)	also, said, best, love, new, come, good, like, men, made	god said, jesus christ, holy spirit, lord krishna, temple god, father devil, eternal life, son god, son man, god father
Politics	Asia (male)	said, one, India, time, people, minister, government, years, state, police, court	indian congress, government plans, modi ministry, human rights, foreign politics, armed forces, international warfare, foreign ministry, middle east, united nations
	Europe (male)	government, said, minister, people, international, country, one, foreign, president, state	make statement, prime minister, human rights, armed forces, secretary state, middle east, united nations, hon friend, foreign secretary, united states
Education	Europe (female)	school, primary, teacher, founder, CEO, judgment, group, named, ranking, prestigious	as founder, founder CEO, judgment group, named fortune, ranking prestigious, world scientist, scientist women, students comprehend, program support, support students
	North America (female)	bachelor, years, student, leader, degree, animal, veterinary, music, taught, communication	bachelors degree, animal veterinary, bachelor music, alison taught, privately years, students ranging, development programmes, including leader, art communication, recent years
Social Media	Africa (male)	onigbinder, aura, pictures, first, gained, popularity, match, beaut, designed, music	aura pictures, gained popularity, match beaut, designed wonder, attending music, music festival, schomburg library, Instagram account, sugar coating, schedule tomorrow
	Asia (male)	time, later, latest, tracks, speedy, Zulfiqar, nasty, children, tweeted, guys	gets later, latest tracks, speedy zulfiqar, children pti, pti tweeted, taking long, long time, hosted pageant, time vincent, love fleeting

Table 11: Common topics with same-gender associations across regions

to train our BERT models for 3 epochs.

M Human Validation

Students and staff from a college campus were recruited as annotators in the study. Screenshots of the form are displayed in Fig 9. We have 6 annotators per region (3 male and 3 female).

N Reproducibility

We open-source our codes, which are uploaded to the submission system. We include commands with hyperparameters in our codes. This would help future work to reproduce our results.

REGION	TOPICS: WORD LISTS
AFRICA	**Nollywood Actress and Movies**: nollywood, actress, actors, drama, celebrity, movie, acting, movies, producer, tv **Parenting and family relationships**: mother, mom, mothers, mum, moms, parent, her, child, momodu, parents **Sports and Football**: players, sports, fifa, team, player, football, mourinho, scored, league, champions **Marriage and relationships**: wives, marriage, husbands, marriages, married, wife, relationships, husband, marry, relationship **Fashion and lifestyle**: cher, nike, max, air, looked, face, love, tods, soldes, scarpe **Womens lives and successes**: women, ladies, woman, female, girls, men, gender, ones, employees, male **Social Media**: instagram, facebook, social, twitter, tweet, snapchat, tweets, tweeted, hashtag, followers **Music**: song, songs, album, hits, music, released, rap, singer, tracks, rapper **Religious and Spiritual Growth**: god, almighty, bible, christ, faith, believers, christian, jesus, prayer, religion **Dating and relationships advice**: dating, women, relationships, ladies, sites, singles, online, single, escorts, websites **Male terms**: male, man, boy, brother, he, him, his, son, Kwame, Mandela, Moyo, Jelani, Tariq, Keita, Obi, Simba, Ayo, Kofi, Jabari, Tunde, Mekonnen, Anwar, Chukwuemeka **Female terms**: sister, mother, aunt, grandmother, daughter, she, hers, her, Aisha, Zahara, Nia, Sade, Amara, Chinelo, Layla, Ayana, Nala, Zuri, Imani, Lola, Kamaria, Nyala, Kaya
ASIA	**Political Leardership in India**: modi, political, said, bjp, told, says, leader, congress, minister, public **Hotel Royalty**: visited, places, stayed, hotels, adventure, pictures, favourite, guest, hiking, hemingway **Sports and Soccer**: sports, team, basketball, players, nba, league, championship, coach, rebounds, finals **Healthy eating habits for children**: food, foods, eating, meals, nutrition, cuisine, diet, dishes, cooking, eat **Social Media platforms for video sharing**: instagram, video, videos, twitter, tweet, facebook, gifs, vlog, youtube, followers **Royal wedding plans**: meghan, duchess, engagement, england, royal, royalty, prince, kate, london, married **Religious devotion and spirituality**: god, bible, holy, faith, prayer, believe, christian, blessed, christ, spiritual **Royal wedding plans**: meghan, duchess, engagement, england, royal, royalty, prince, kate, london, married **Bollywood actors and films**: bollywood, bachchan, kapoor, actors, acting, kareena, actor, film, shahrukh, hindi **Marriage**: married, marriage, marriages, couple, couples, wife, marry, wedding, husband, divorced **Male terms**: male, man, boy, brother, he, him, his, son, Hiroshi, Ravi, Kazuki, Jin, Satoshi, Rohan, Haruki, Dai, Akira, Yuan **Female terms**: sister, mother, aunt, grandmother, daughter, she, hers, her, Sakura, Mei, Aiko, Yuna, Lina, Ji-hye, Mika, Nami, Anika, Rina
EUROPE	**Music**: music, songs, vocalists, album, albums, singing, vocals, singles, rock, song **Education**: school, schools, classroom, students, education, educational, pupils, boys, academy, college **Political decisions and impact on society**: government, public, minister, said, hon, people, first, the, column, committee **Comfortable hotels**: guests, staying, rooms, friendly, welcoming, stayed, hotel, beds, stay, comfortable **UK Government Taxation Policies**: corbyn, taxation, fiscal, tax, taxes, exchequer, labour, governments, government, deficit **Luxury Sailing**: yachts, yacht, boat, sailing, sails, cruising, sail, berths, cruiser, cabin **Christian Theology and Practice**: god, bible, christ, jesus, faith, christian, religious, religion, holy, gave **Obituaries and Genealogy**: died, edward, relatives, anne, lived, elizabeth, funeral, irish, mrs, galway **Christian Theology and Practice**: god, bible, christ, jesus, faith, christian, religious, religion, holy, gave **Fashion and style**: fashion, shoes, style, clothes, clothing, shoe, wear, nike, dress, stylish **Male terms**: male, man, boy, brother, he, him, his, son, Lukas, Matteo, Sebastian, Alexander, Gabriel, Nikolai, Maximilian, Leonardo, Daniel, Adrian **Female terms**: sister, mother, aunt, grandmother, daughter, she, hers, her, Emma, Sophia, Olivia, Isabella, Ava, Mia, Charlotte, Amelia, Lily, Emily
NORTH AMERICA	**Religion and Spirituality**: god, christ, jesus, bible, christian, holy, christians, scripture, faith, heaven **Online Dating for Singles**: dating, singles, hookup, single, relationships, dates, flirting, personals, date, mingle **Reproductive Health**: download, available, pdf, online, edition, manual, free, reprint, kindle, file **Fashion and style**: fashion, dresses, dress, wardrobe, clothes, clothing, style, outfit, vintage, wear **Reinsurance and capital markets**: reinsurance, reinsurers, insurers, insurance, securities, investors, investment, finance, trading, pension **Education and achievements**: school, schools, graduated, college, students, undergraduate, graduation, graduate, attended, education **Nike shoes and fashion**: nike, shoes, sneakers, jordans, jeans, tops, black, boys, men, casual **Family dynamics and relationships**: family, families, children, kids, grandchildren, relatives, grandparents, parents, child, parent **Cape Cod news**: lifeguard, drowned, drowns, newstweet, hospitalized, snorkeling, cape, reported, reuterstweet, pulled **Reading and fiction**: books, book, reading, novels, series, enjoyed, novel, romance, katniss, readers **Male terms**: male, man, boy, brother, he, him, his, son, Liam, Noah, Ethan, Jacob, William, Michael, James, Alexander, Benjamin, Matthew **Female terms**: sister, mother, aunt, grandmother, daughter, she, hers, her, Emma, Olivia, Ava, Sophia, Isabella, Mia, Charlotte, Amelia, Harper, Evelyn
OCEANIA	**Religious beliefs and figures**: god, gods, bible, mankind, faith, christ, spiritual, christian, religion, jesus **Family relationships**: mum, mother, mom, mums, parent, family, parents, baby, dad, father **Music record and Artists**: music, album, albums, jazz, songs, hits, musicians, artists, recordings, blues **Woordworking plans and projects**: plans, furniture, woodwork, wood, woodcraft, woodworking, plywood, carpentry, cabinets, wooden **Building and designing boats**: boatbuilder, boatbuilding, boats, plans, boat, sauceboat, sailboat, build, catamaran, kits **Weight loss and nutrition for women**: diet, workout, exercise, foods, weight, food, eating, healthy, pounds, fat **Superheroes and their Universes**: superhero, superheroes, avengers, marvel, comics, superman, aquaman, heroes, comic, hero **Exercises for hormone development**: hormones, weightlifting, workouts, deadlifts, hormonal, exercises, lifting, testosterone, fitness, squats **Building and designing boats**: boatbuilder, boatbuilding, boats, plans, boat, sauceboat, sailboat, build, catamaran, kits **Kids furniture and decor**: furniture, chairs, sofas, ikea, sofa, cushions, sectional, upholstered, couch, childrens **Male terms**: male, man, boy, brother, he, him, his, son, Manaia, Tane, Kai, Ariki, Mika, Koa, Rangi, Kane, Tama, Hemi **Female terms**: sister, mother, aunt, grandmother, daughter, she, hers, her, Aroha, Moana, Tui, Lani, Kahurangi, Ariana, Malie, Marama, Ava, Kaia

Table 12: Word lists corresponding to each topic for computing region-aware WEAT metric

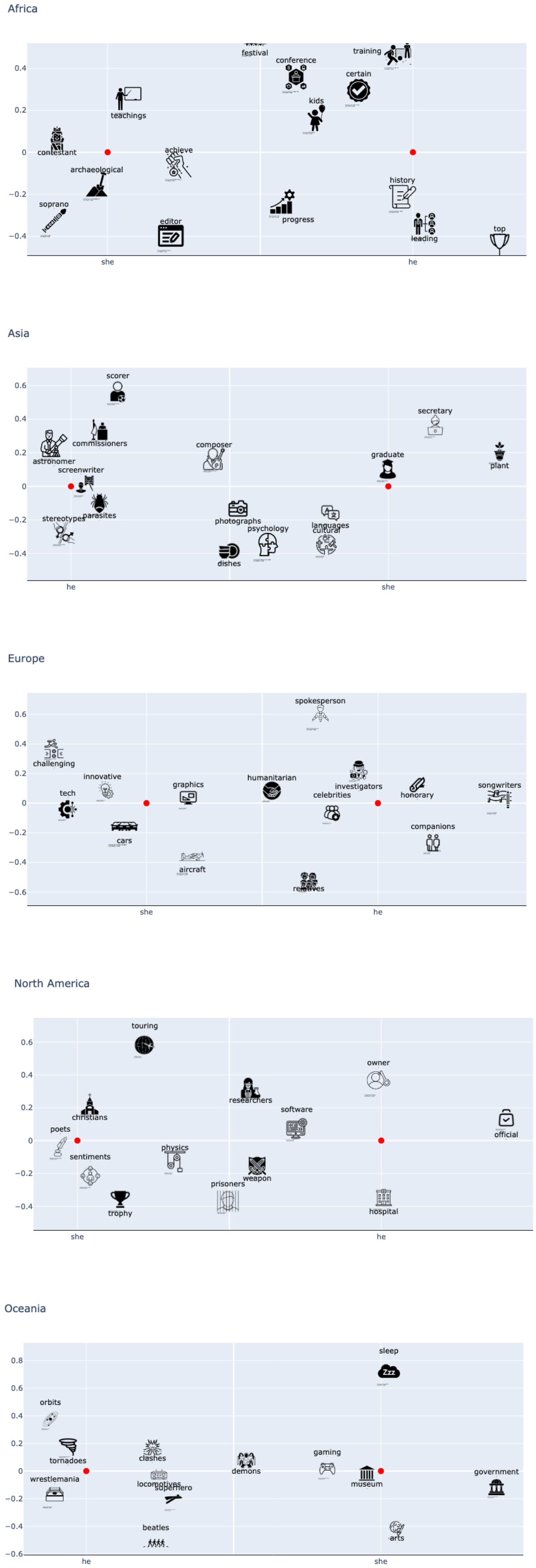
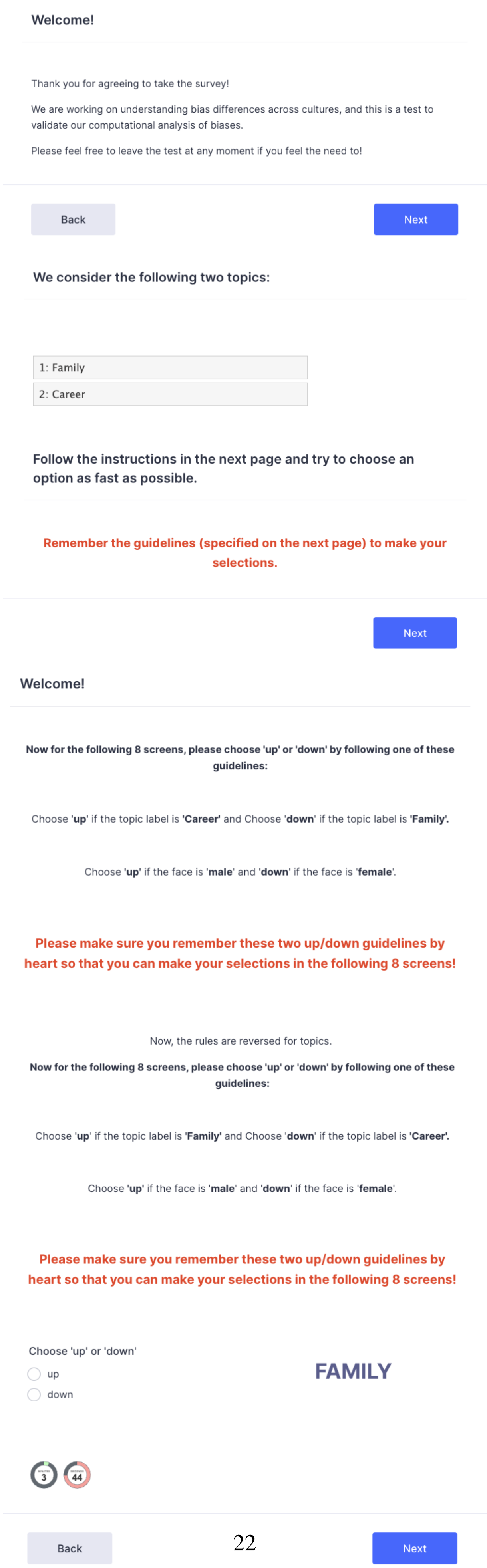


Figure 8: Top words for each region(Africa, Asia, Europe, North America and Oceania) using region-specific BERTs


Welcome!

Thank you for agreeing to take the survey!

We are working on understanding bias differences across cultures, and this is a test to validate our computational analysis of biases.

Please feel free to leave the test at any moment if you feel the need to!

Back Next

We consider the following two topics:

1: Family

2: Career

Follow the instructions in the next page and try to choose an option as fast as possible.

Remember the guidelines (specified on the next page) to make your selections.

Next

Welcome!

Now for the following 8 screens, please choose 'up' or 'down' by following one of these guidelines:

Choose '**up**' if the topic label is '**Career**' and Choose '**down**' if the topic label is '**Family**'.

Choose '**up**' if the face is '**male**' and '**down**' if the face is '**female**'.

Please make sure you remember these two up/down guidelines by heart so that you can make your selections in the following 8 screens!

Now, the rules are reversed for topics.

Now for the following 8 screens, please choose 'up' or 'down' by following one of these guidelines:

Choose '**up**' if the topic label is '**Family**' and Choose '**down**' if the topic label is '**Career**'.

Choose '**up**' if the face is '**male**' and '**down**' if the face is '**female**'.

Please make sure you remember these two up/down guidelines by heart so that you can make your selections in the following 8 screens!

Choose 'up' or 'down'

up

down

FAMILY

3 44

Back Next

Figure 9: Annotation Form Screenshots (We do not include screenshots with faces to protect privacy)