Supplementary Materials  
for the Research Report  
“Bias Toward Chinese in English Natural Language”

Open Practices (All data and code are available at [https://osf.io/9wgkb/](https://osf.io/9wgkb/?view_only=e1ac4ba21d794d7e985e5ba3c1d83f16))

* **“Dict\_Stereotype.csv”**: Dictionary of stereotype content, with 12 pairs of words for each of the four stereotype facets (i.e., a total of 48 word pairs)
* **“FMAT\_China\_1.RData”**: Raw data produced by BERT models with the FMAT package functions for (1) *Stereotypes* of Chinese (and other groups)
* **“FMAT\_China\_2.RData”**: Raw data produced by BERT models with the FMAT package functions for (2) *Prejudice* toward Chinese (and other groups)
* **“FMAT\_China\_3.RData”**: Raw data produced by BERT models with the FMAT package functions for (3) *Discrimination* toward Chinese (and other groups)
* **“Perceiver.RData”**: Raw data produced by BERT models with the FMAT package functions for (4) *Supplementary Analysis* of text producers (see below)
* **“Supp\_FMAT\_China.html”**: An HTML file combining the R code workflow and full results (please open this file with a web browser, e.g., Chrome)

BERT Model Information

**Table S1.** Summary of the 12 BERT Pretrained Language Models Sampled in This Study.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model name on [HuggingFace](https://huggingface.co/models) | Case-sensitive | Vocabulary size | Embedding dimension | Hidden layer | Pretraining corpora |
| bert-base-uncased | No | 30,522 | 768 | 12 | Wiki, Book |
| bert-base-cased | Yes | 28,996 | 768 | 12 | Wiki, Book |
| bert-large-uncased | No | 30,522 | 1,024 | 24 | Wiki, Book |
| bert-large-cased | Yes | 28,996 | 1,024 | 24 | Wiki, Book |
| distilbert-base-uncased | No | 30,522 | 768 | 6 | Wiki, Book |
| distilbert-base-cased | Yes | 28,996 | 768 | 6 | Wiki, Book |
| albert-base-v1 | No | 30,000 | 128 | 12 | Wiki, Book |
| albert-base-v2 | No | 30,000 | 128 | 12 | Wiki, Book |
| roberta-base | Yes | 50,265 | 768 | 12 | Wiki, Book, CC, Open |
| distilroberta-base | Yes | 50,265 | 768 | 6 | Open |
| vinai/bertweet-base | Yes | 64,001 | 768 | 12 | Twitter |
| vinai/bertweet-large | Yes | 50,265 | 1,024 | 24 | Twitter |

*Note*. BERT = Bidirectional Encoder Representations from Transformers; Wiki = Wikipedia (2.5 billion words); Book = BookCorpus (800 million words from 11,038 unpublished books scraped from the Internet); CC = CommonCrawl (63 million English news articles); Open = OpenWebText (8 million documents from Reddit); Twitter = Tweets (850 million English Tweets).

Table S1 presents detailed information of the 12 BERT models sampled in this study, the same sample as used in existing FMAT research (Bao, 2024; Bao & Gries, 2024): BERT (base/large, uncased/cased), ALBERT (v1/v2), DistilBERT (uncased/cased), RoBERTa, DistilRoBERTa, BERTweet (base/large). These BERT models are all openly available from HuggingFace (<https://huggingface.co/models>). Notably, these models have been pretrained on large English-language text corpora through 2019, including Wikipedia (2.5 billion words), BookCorpus (800 million words from 11,038 books), CommonCrawl (63 million news articles), OpenWebText (8 million Reddit documents), and Twitter (850 million Tweets). Therefore, the models have inherited semantic information about social bias embedded in the pretraining corpora. Any downstream tasks no longer require access to the raw text corpora. We directly used such pretrained models as artificial intelligent (AI) respondents to “answer” queries about social bias, without resampling the raw text materials. Indeed, we performed analyses based on BERT models themselves, but we did not access the raw text corpora used to train the BERT models.

Supplementary Analysis #1: Who Produced the Texts?

We conducted a supplemental analysis to better understand who might have produced the texts used to train the 12 English-language BERT models. While the exact demographic characteristics of the text producers are unknown, a convergence was found between (1) relative semantic probabilities of *American*, *British*, *Canadian*, and *Australian* filling in the mask of “I am [MASK].” and (2) relative percentages of populations in these countries (see Bao, 2024). For a more systematic estimate of the text producer nationalities, we used the FMAT to test a wider range of nationality option words.

Method

The FMAT requires a query (a sentence with a masked word) for a BERT model to understand the context and estimate how likely each word in the model’s vocabulary might replace the mask. To detect the likely nationalities of the text producers, we designed two parallel queries: “I am [MASK].” and “We are [MASK].”

Target words for the [MASK] included 28 nationalities or ethnicities that were in the vocabulary of all 12 BERT models: (1) eight major English-speaking nationalities (*American*, *British*, *Canadian*, *Australian*, *English*, *Irish*, *Scottish*, *Welsh*); (2) ten European nationalities where English is not the official language but may be used (*Austrian*, *Belgian*, *Danish*, *Dutch*, *French*, *German*, *Italian*, *Polish*, *Portuguese*, *Spanish*); and (3) ten non-White nationalities or ethnicities (*African*, *Nigerian*, *Asian*, *Chinese*, *Japanese*, *Korean*, *Indian*, *Pakistani*, *Indonesian*, *Malaysian*). The estimated raw probability of a masked word is not its word frequency but how semantically probable it would be to appear in the mask given the query context (Bao, 2024). The raw data contained 672 observations of probabilities (= 28 [MASK] target words × 2 query templates × 12 BERT models).

Following previous FMAT research (Bao, 2024; Bao & Gries, 2024), we first assessed the reliability of FMAT: (1) the interrater agreement among the 12 BERT models in understanding query sentences and estimating semantic probabilities, which was evaluated with the average-score intraclass correlation coefficient (ICCaverage); and (2) the internal consistency between queries, which was evaluated with Cronbach’s α (αquery).

For the main analysis, we computed relative semantic probabilities of target words, for each BERT model and FMAT query, by dividing the raw probability of each word by the sum probability of all the 28 words. Then, we aggregated the results for each word across BERT models and FMAT queries.

Results

The FMAT showed good interrater agreement among the 12 BERT models (ICCaverage = .88) and internal consistency between the two queries (αquery = .86). The relative semantic probabilities of *American* (16.2%), *British* (9.0%), *Canadian* (4.5%), *Australian* (2.5%), *English* (11.8%), *Irish* (5.7%), *Scottish* (2.7%), and *Welsh* (2.0%) added up to 54.5%. These eight major English-speaking nationalities plus the ten European nationalities together had a relative semantic probability of 82.2%, suggesting that text producers were overwhelmingly Westerners. By contrast, the ten non-White nationalities/ethnicities had relative probabilities ranging from just 0.4% (*Pakistani*) to 1.6% (*African*, *Asian*) to 4.1% (*Chinese*). Therefore, most of the text producers for the sampled BERT models were likely Westerners, primarily native English speakers from major Anglophone countries. The eight Anglophone nationalities were therefore used as the baseline control condition for pairwise comparison.

Supplementary Analysis #2: Disaggregating Anglophone Control Groups

In the main analysis, we averaged results across all eight Anglophone nationalities to produce an overall “Anglophone” control group. While this treatment makes full use of the information and avoids cherry-picking any nationality, it is also meaningful to test nuances of bias within each Anglophone nationality separately. In doing so, we can also check the robustness of our main findings across the control nationalities.

Table S2 presents standardized FMAT effect sizes of bias toward Chinese (vs. control group). The first column labelled “Overall” shows results identical to Table 2 in the main text. As the most robust results (for all or most Anglophone control groups), Chinese people were semantically associated with less moral/trustworthy and less assertive/dominant stereotypes, contempt (vs. admiration) prejudice, and both active and passive harm (vs. facilitation) discrimination.

We also found nuances in two facets of stereotypes, “Warmth: sociable/friendly” and “Competence: capable/intelligent”: Chinese were perceived to be (1) as sociable/friendly as American/British/Canadian/Australian/Irish people, but less sociable/friendly than English/Scottish/Welsh; and (2) as capable/intelligent as Canadian/English/Scottish/Welsh people, but more capable/intelligent than American/British/Australian/Irish people.

**Table S2.** FMAT Effect Sizes of Bias Toward *Chinese* (vs. Aggregated and Disaggregated Control Groups).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bias components | Chinese vs. Anglophone nationalities (eight control groups) | | | | | | | | |
| **Overall** | American | British | Canadian | Australian | English | Irish | Scottish | Welsh |
| **Stereotypes** |  |  |  |  |  |  |  |  |  |
| Warmth: sociable/friendly | **–0.10\*** | 0.03 | –0.03 | –0.07 | –0.01 | –0.20\*\*\* | –0.07 | –0.24\*\*\* | –0.20\*\*\* |
| Warmth: moral/trustworthy | **–0.21\*\*\*** | –0.11\* | –0.05 | –0.34\*\*\* | –0.13\*\* | –0.14\*\* | –0.23\*\*\* | –0.34\*\*\* | –0.35\*\*\* |
| Competence: capable/intelligent | **0.07** | 0.21\*\*\* | 0.18\*\*\* | 0.01 | 0.11\* | 0.01 | 0.14\*\* | –0.02 | –0.09 |
| Competence: assertive/dominant | **–0.23\*\*\*** | –0.22\*\*\* | –0.15\*\* | –0.24\*\*\* | –0.11\* | –0.15\*\* | –0.26\*\*\* | –0.41\*\*\* | –0.33\*\*\* |
| **Prejudice** |  |  |  |  |  |  |  |  |  |
| Admiration (vs. contempt) | **–0.12\*\*** | –0.02 | 0.10\* | –0.29\*\*\* | –0.19\*\*\* | 0.08 | –0.15\*\*\* | –0.17\*\*\* | –0.33\*\*\* |
| Envy (vs. pity) | **0.05** | 0.08\* | 0.04 | 0.09\* | 0.06 | 0.01 | 0.01 | 0.04 | 0.05 |
| **Discrimination** |  |  |  |  |  |  |  |  |  |
| Active facilitation (vs. harm) | **–0.27\*\*\*** | –0.26\*\*\* | –0.16\*\* | –0.37\*\*\* | –0.32\*\*\* | –0.19\*\*\* | –0.22\*\*\* | –0.25\*\*\* | –0.39\*\*\* |
| Passive facilitation (vs. harm) | **–0.34\*\*\*** | –0.24\*\*\* | –0.13\*\* | –0.48\*\*\* | –0.40\*\*\* | –0.22\*\*\* | –0.27\*\*\* | –0.42\*\*\* | –0.51\*\*\* |

*Note*. FMAT = Fill-Mask Association Test. Effect sizes are standardized estimates (Cohen’s *d*) from the Linear Mixed Model of each bias component: 0.2 is considered small, 0.5 medium, and 0.8 large.

\* *p* < .05. \*\* *p* < .01. \*\*\* *p* < .001. (Blue/red cells are significantly positive/negative values.)

Supplementary Analysis #3: Biases Toward Other East Asian Subgroups

To explore whether the pattern of our results was unique to Chinese or just common to all East Asian subgroups, we conducted similar tests on bias toward *Japanese* and *Korean* people with all and each of the eight Anglophone nationalities as the control group.

**Bias toward Japanese people** (Table S3): The results showed a similar pattern to bias toward Chinese. Japanese people were semantically associated with less moral/trustworthy and less assertive/dominant stereotypes, contempt (vs. admiration) prejudice, and both active and passive harm (vs. facilitation) discrimination. The effect sizes of discrimination toward Japanese were smaller than (about half as large as) that toward Chinese, however.

**Bias toward Korean people** (Table S4): Biases toward Koreans were substantially different from those toward Chinese and Japanese, with further differences between disparate Anglophone control groups. Korean people were (weakly) semantically associated with more capable/intelligent and less assertive/dominant stereotypes, and were (weakly) more admired than scorned, with no solid association with discrimination.

In sum, biases toward East Asians varied between East Asian target subgroups, with the pattern for Koreans distinct from that for Chinese/Japanese.

**Table S3.** FMAT Effect Sizes of Bias Toward *Japanese* (vs. Aggregated and Disaggregated Control Groups).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bias components | Japanese vs. Anglophone nationality (eight control groups) | | | | | | | | |
| **Overall** | American | British | Canadian | Australian | English | Irish | Scottish | Welsh |
| **Stereotypes** |  |  |  |  |  |  |  |  |  |
| Warmth: sociable/friendly | **–0.09** | 0.04 | –0.02 | –0.06 | –0.00 | –0.18\*\*\* | –0.06 | –0.22\*\*\* | –0.19\*\*\* |
| Warmth: moral/trustworthy | **–0.15\*\*** | –0.05 | 0.02 | –0.27\*\*\* | –0.06 | –0.08 | –0.17\*\*\* | –0.28\*\*\* | –0.29\*\*\* |
| Competence: capable/intelligent | **–0.02** | 0.12\* | 0.10\* | –0.08 | 0.02 | –0.08 | 0.06 | –0.11\* | –0.17\*\*\* |
| Competence: assertive/dominant | **–0.30\*\*\*** | –0.29\*\*\* | –0.22\*\*\* | –0.31\*\*\* | –0.18\*\*\* | –0.21\*\*\* | –0.33\*\*\* | –0.48\*\*\* | –0.40\*\*\* |
| **Prejudice** |  |  |  |  |  |  |  |  |  |
| Admiration (vs. contempt) | **–0.10\*** | –0.00 | 0.12\*\* | –0.27\*\*\* | –0.17\*\*\* | 0.10\* | –0.13\*\* | –0.15\*\*\* | –0.31\*\*\* |
| Envy (vs. pity) | **0.02** | 0.05 | 0.01 | 0.06 | 0.03 | –0.02 | –0.02 | 0.01 | 0.02 |
| **Discrimination** |  |  |  |  |  |  |  |  |  |
| Active facilitation (vs. harm) | **–0.14\*\*** | –0.14\* | –0.03 | –0.24\*\*\* | –0.19\*\*\* | –0.06 | –0.09 | –0.13\* | –0.26\*\*\* |
| Passive facilitation (vs. harm) | **–0.22\*\*\*** | –0.12\* | –0.02 | –0.37\*\*\* | –0.28\*\*\* | –0.10\* | –0.15\*\* | –0.31\*\*\* | –0.40\*\*\* |

*Note*. FMAT = Fill-Mask Association Test. Effect sizes are standardized estimates (Cohen’s *d*) from the Linear Mixed Model of each bias component: 0.2 is considered small, 0.5 medium, and 0.8 large.

\* *p* < .05. \*\* *p* < .01. \*\*\* *p* < .001. (Blue/red cells are significantly positive/negative values.)

**Table S4.** FMAT Effect Sizes of Bias Toward *Koreans* (vs. Aggregated and Disaggregated Control Groups).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bias components | Korean vs. Anglophone nationality (eight control groups) | | | | | | | | |
| **Overall** | American | British | Canadian | Australian | English | Irish | Scottish | Welsh |
| **Stereotypes** |  |  |  |  |  |  |  |  |  |
| Warmth: sociable/friendly | **0.09** | 0.22\*\*\* | 0.16\*\* | 0.12\* | 0.18\*\*\* | –0.00 | 0.12\* | –0.05 | –0.01 |
| Warmth: moral/trustworthy | **0.08** | 0.18\*\*\* | 0.25\*\*\* | –0.04 | 0.17\*\*\* | 0.15\*\* | 0.06 | –0.04 | –0.06 |
| Competence: capable/intelligent | **0.13\*\*** | 0.27\*\*\* | 0.25\*\*\* | 0.07 | 0.17\*\*\* | 0.07 | 0.20\*\*\* | 0.04 | –0.02 |
| Competence: assertive/dominant | **–0.10\*** | –0.08 | –0.02 | –0.10\* | 0.03 | –0.01 | –0.13\*\* | –0.28\*\*\* | –0.20\*\*\* |
| **Prejudice** |  |  |  |  |  |  |  |  |  |
| Admiration (vs. contempt) | **0.09\*** | 0.19\*\*\* | 0.31\*\*\* | –0.08 | 0.02 | 0.28\*\*\* | 0.06 | 0.03 | –0.13\* |
| Envy (vs. pity) | **0.01** | 0.05 | 0.00 | 0.05 | 0.02 | –0.03 | –0.02 | 0.00 | 0.01 |
| **Discrimination** |  |  |  |  |  |  |  |  |  |
| Active facilitation (vs. harm) | **0.01** | 0.02 | 0.12\* | –0.09 | –0.04 | 0.09 | 0.06 | 0.03 | –0.11\* |
| Passive facilitation (vs. harm) | **–0.08** | 0.02 | 0.12\* | –0.23\*\*\* | –0.14\*\* | 0.04 | –0.01 | –0.17\*\* | –0.26\*\*\* |

*Note*. FMAT = Fill-Mask Association Test. Effect sizes are standardized estimates (Cohen’s *d*) from the Linear Mixed Model of each bias component: 0.2 is considered small, 0.5 medium, and 0.8 large.

\* *p* < .05. \*\* *p* < .01. \*\*\* *p* < .001. (Blue/red cells are significantly positive/negative values.)

References

Bao, H.-W.-S. (2024). The Fill-Mask Association Test (FMAT): Measuring propositions in natural language. *Journal of Personality and Social Psychology, 127*(3), 537–561.

Bao, H.-W.-S., & Gries, P. (2024). Intersectional race–gender stereotypes in natural language. *British Journal of Social Psychology, 63*(4), 1771–1786.