# Team #10 - Culture Hood
# Analysis Of Cultural Biases In Language Models

Vaibhav Sharma; 40290623; vaibhav.sharma17649@gmail.com
and
Varshap Walia; 40230028; varshwalia@gmail.com

December 12, 2024

**Abstract.** We investigated cultural biases in Large Language Models (LLM's) using the two methodologies CrowS-Pairs and Hofstede's six cultural dimensions, standardized to the VSM framework. For CrowS-pairs we analyzed biases in different LLM's by comparing the generated likelihood of sentences. We analyzed the sentence likelihoods for stereotyping or anti-stereotyping statements for both advantaged and disadvantaged groups on two languages(English and Chinese). For VSM, we analyzed variations in Hofstede's six cultural dimensions across different LLM's at different temperatures to check the consistency of models. We identified disparities in model behaviors providing valuable insights for mitigating biases in these models.

# Contents

# 1 Goal of the project

The goal of this project is to comprehensively analyze cultural biases and demographic stereotypes in Large Language Models (LLM's) through two distinct methodologies: CrowS-Pairs dataset evaluation and Hofstede's cultural dimensions for VSM. We are interested in understanding what cultural biases in the LLM models lead to and how they move across dimensions and different demographics. The project aims to achieve this by:

• **Cultural Dimension Analysis (VSM):** Taking the six cultural dimensions(Figure 3) of Hofstede's Value Survey Model (VSM) and **standardizing** and **analyzing** to see model trends. Our goal is to conduct the **comparison** of the influence between the different set of **culture factors** such as Age, Sex, Social Class, Region, Education, Country level across different **LLM models** and **languages**.

• **Bias Evaluation with CrowS-Pairs:** We evaluate the tendency of models to favour **stereotypical** sentences over **anti-stereotypical** sentences across **nine bias categories** mainly Race, Gender, Sexual Orientation, Religion, Age, Nationality, Disability, Physical Appearance, Socioeconomic Status using the CrowS-Pairs dataset. By comparing **sentence likelihoods** leveraging autoregressive models, we analyzed how much these models favours **advantaged**(Figure 7) and **disadvantaged groups**(Figure 8).

# 2 Methodology

## 2.1 Value Model Survey methodology

• **Randomized Data Creation:** We started by defining random countries and their regions from all continents except Antarctica and also included other important demographic details such as **age**, **marital status**, **education**, **social class** ensuring they align with Hofstede's cultural dimensions(Figure 4). Once completed, the dataset was saved as an Excel file for further use for both English and Chinese languages.

• **Survey Prompt Creation:** Designed prompts using demographic data and Hofstede's cultural dimensions and also translated prompts into Chinese while retaining cultural contexts. Saved the generated prompts for querying LLM's.

• **Queried LLM Models:** Tested multiple models at varying temperatures (0.7, 1, 1.3) for English and Chinese data: GPT-3.5 Turbo, GPT-4o, Gemini-1.5 Flash/Pro and saved the model responses in JSON format for

both languages.

• **Processed Model Outputs:** Converted JSON responses to CSV format. Extracted question-wise responses and categorized them into Hofstede dimensions.

• **Calculated Hofstede Scores:** Computed dimension scores for each prompt based on the question groups from VSM Questionnaire: PDI: Q1–Q4, IDV: Q5–Q8, MAS: Q9–Q12, UAI: Q13–Q16, LTO: Q17–Q20, IVR: Q21–Q24 and scaled the raw scores using baseline values and a scaling factor.

• **Visualization and Comparison:** Created comparative visualizations (e.g. violin plots) to analyze Hofstede dimensions across: Models (e.g. GPT-3.5, GPT-4o, Gemini-1.5), Languages (English vs. Chinese), Temperatures (0.7, 1, 1.3)

• **Analyzed Trends across various demographics:** Compared trends in Hofstede's dimensions scores across different LLM models, languages and demographics and highlighted deviations across models and their biasis.

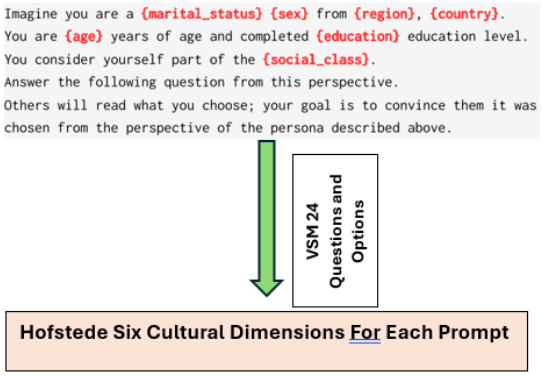


Figure 1: Survey prompt template followed by the VSM24 questionnaire for each prompt

## 2.2 Crows pairs methodology

• **Dataset Preparation:** We began by loading paired sentences that included both stereotypical and anti-stereotypical examples featuring both advantaged(Figure 7) and disadvantaged(Figure 8) groups. These sentences covered **nine** categories: **Race/Color, Gender/Identity, Sexual Orientation, Religion, Age, Nationality, Disability, Physical Appearance, and Socioeconomic Status.**

• **Queried Models for Likelihood scores:** We utilized autoregressive language models, including GPT-4o and Gemini-1.5, at varying temperatures (0.7, 1, 1.5). Paired sentences were inputted along with prepared prompts to measure and record likelihood scores. The results were then saved in JSON/CSV formats for further analysis.
• **Computed Bias Metric:** We computed the bias metric by comparing the likelihood scores of pair of sentences(Stereo and Anti-stereo). This involved determining the model's preference for stereotypical sentences over anti-stereotypical ones. A higher preference for stereotypical sentences indicated a bias in the model.
• **Visualized Results:** We visualized the results by comparing various LLM models across different bias categories at varying temperatures. This highlighted trends and deviations, such as a model's stronger preference for stereotyping or anti-stereotyping in specific languages or bias categories under particular temperature settings.
• **Analyzed Trends:** We analyzed trends by evaluating variations across different LLM models, bias categories, and temperature settings. This included examining how bias differed between languages and across various evaluation setups.
**Note:** We also explored an alternative method for analyzing biases, including generating masked sentences and evaluating them based on predicted likelihoods. However, this approach was discarded as it did not yield meaningful insights. Please refer to the code for implementation details.

# 3 Evaluation

## 3.1 Results

**Hofstede Scores - English**

| Model Name | Version | Temperature | Language | PDI | IDV | MAS | UAI | LTO | IVR |
|---|---|---|---|---|---|---|---|---|---|
| chatgpt | 3.5 | 0.7 | en | 53.62 | 42.56 | 53.19 | 35.19 | 42.69 | 40.88 |
| | | 1 | en | 40.00 | 39.19 | 43.12 | 42.62 | 39.13 | 36.50 |
| | | 1.3 | en | 51.31 | 45.44 | 50.00 | 42.69 | 43.88 | 45.13 |
| | 4.0 | 0.7 | en | 25.25 | 35.25 | 43.38 | 30.69 | 35.69 | 30.25 |
| | | 1 | en | 24.94 | 36.44 | 45.31 | 32.50 | 35.75 | 29.69 |
| | | 1.3 | en | 25.75 | 34.38 | 45.06 | 33.69 | 37.44 | 30.75 |
| gemini | 1.5 | 0.7 | en | 41.50 | 39.31 | 50.12 | 27.12 | 47.63 | 50.31 |
| | | 1 | en | 41.48 | 38.95 | 49.75 | 27.71 | 47.79 | 49.68 |
| | | 1.3 | en | 42.61 | 38.83 | 49.94 | 27.27 | 47.35 | 49.56 |

**Hofstede Scores - Chinese**

| Model Name | Version | Temperature | Language | PDI | IDV | MAS | UAI | LTO | IVR |
|---|---|---|---|---|---|---|---|---|---|
| chatgpt | 3.5 | 0.7 | cn | 56.06 | 37.38 | 68.25 | 25.00 | 56.00 | 56.12 |
| | | 1 | cn | 56.06 | 37.25 | 68.31 | 25.37 | 56.12 | 56.12 |
| | | 1.3 | cn | 56.12 | 38.12 | 67.31 | 26.31 | 55.56 | 55.50 |
| | 4.0 | 0.7 | cn | 28.38 | 43.25 | 47.31 | 37.31 | 39.25 | 35.44 |
| | | 1 | cn | 28.06 | 44.88 | 45.38 | 39.69 | 39.56 | 35.56 |
| | | 1.3 | cn | 28.37 | 41.81 | 46.06 | 40.38 | 38.31 | 35.56 |
| gemini | 1.5 | 0.7 | cn | 50.69 | 54.56 | 51.25 | 26.12 | 50.50 | 46.81 |
| | | 1 | cn | 50.81 | 55.06 | 52.06 | 26.25 | 50.06 | 47.62 |
| | | 1.3 | cn | 50.25 | 54.00 | 53.06 | 25.69 | 50.50 | 45.88 |

Figure 2: Comparison of Hofstede Scores for English and Chinese datasets across models.

Model Summary of Bias Data Metrics (English)

| | mdl | ver | temp | lan | avg_adv_str | avg_dis_str | avg_adv_anti | avg_dis_anti | avg_diff | avg_diff_str | avg_diff_anti |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | chatgpt | 4o | 0.7 | en | 0.50 | 0.47 | 0.69 | 0.65 | 0.18 | 0.18 | 0.19 |
| 2 | chatgpt | 4o | 1.0 | en | 0.56 | 0.45 | 0.65 | 0.70 | 0.19 | 0.19 | 0.19 |
| 3 | chatgpt | 4o | 1.3 | en | 0.60 | 0.44 | 0.66 | 0.66 | 0.19 | 0.20 | 0.19 |
| 4 | gemini | 1.5 | 0.7 | en | 0.52 | 0.42 | 0.55 | 0.61 | 0.16 | 0.13 | 0.19 |
| 5 | gemini | 1.5 | 1.0 | en | 0.48 | 0.42 | 0.60 | 0.57 | 0.16 | 0.14 | 0.18 |
| 6 | gemini | 1.5 | 1.3 | en | 0.49 | 0.42 | 0.57 | 0.60 | 0.16 | 0.13 | 0.19 |

Model Summary of Bias Data Metrics (Chinese)

| | mdl | ver | temp | lan | avg_adv_str | avg_dis_str | avg_adv_anti | avg_dis_anti | avg_diff | avg_diff_str | avg_diff_anti |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | chatgpt | 4o | 0.7 | cn | 0.52 | 0.45 | 0.66 | 0.53 | 0.17 | 0.18 | 0.16 |
| 2 | chatgpt | 4o | 1.0 | cn | 0.52 | 0.47 | 0.64 | 0.57 | 0.19 | 0.20 | 0.17 |
| 3 | chatgpt | 4o | 1.3 | cn | 0.56 | 0.43 | 0.67 | 0.52 | 0.21 | 0.20 | 0.21 |
| 4 | gemini | 1.5 | 0.7 | cn | 0.59 | 0.38 | 0.50 | 0.56 | 0.16 | 0.19 | 0.13 |
| 5 | gemini | 1.5 | 1.0 | cn | 0.59 | 0.38 | 0.47 | 0.58 | 0.16 | 0.18 | 0.15 |
| 6 | gemini | 1.5 | 1.3 | cn | 0.55 | 0.40 | 0.48 | 0.58 | 0.18 | 0.20 | 0.15 |

Figure 3: Comparison of bias metrics in CrowS-Pairs for advantaged and disadvantaged groups for English and Chinese datsets.

### 3.2 Analysis

**VSM Analysis(Figure 2, 5 and 6):**
**Note:-** Please refer to **Figure 4** to understand Hofstede's six cultural dimensions.
• **Power Distance Index (PDI):** The Chinese dataset (**cn**) had higher PDI values, reflecting stronger acceptance of hierarchy, while the English dataset (**en**) demonstrated more compact and egalitarian tendencies. Models like GPT-4o and Gemini showed broader variability for the Chinese dataset.
• **Individualism vs. Collectivism (IDV):** The English dataset leaned toward collectivism, whereas the Chinese dataset exhibited higher individualism. Gemini models showed more pronounced individualistic tendencies for the Chinese dataset.
• **Masculinity vs. Femininity (MAS):** Masculinity scores were consistently high across all datasets, reflecting a competitive culture and traditional gender roles. The Chinese dataset showed slightly more variability compared to the English dataset.
• **Uncertainty Avoidance Index (UAI):** Both datasets displayed moderate values, with the English dataset showing slightly broader variability, indicating a more flexible approach to uncertainty.
• **Long-Term vs. Short-Term Orientation (LTO):** The Chinese dataset favored short-term rewards, while the English dataset focused on long-term goals and planning, as seen in its more compact distribution.
• **Indulgence vs. Restraint (IVR):** The English dataset emphasized indulgence, prioritizing personal happiness, while the Chinese dataset leaned toward restraint, focusing on duty and order.

**CrowS-Pairs Analysis(Figure 2):**

All models consistently displayed **stereotypical bias**, favoring **advantaged groups** (Figure 7) over **disadvantaged groups** (Figure 8).
• **Gemini** was unexpectedly **less biased** compared to **ChatGPT**.
• Both datasets exhibited **similar patterns**, with an average **20% likelihood difference** between **advantaged** and **disadvantaged groups**, indicating **polarized results**.
• The **Chinese** dataset performed better in **anti-stereotypical scenarios** for disadvantaged groups.
• Bias likelihood **increased** with higher **temperatures** (randomness) in the prompts.
• **Older models**, like **ChatGPT-3.5**, often **refused to respond** to prompts containing overt bias related to **gender, race, or religion**. ie. make sense of the results and draw general conclusions

**CrowS-Pairs Category-Wise Bias Analysis(Figures 9-17):** The CrowS-Pairs evaluation highlighted biases in LLM's across nine categories, comparing stereotypical and anti-stereotypical likelihoods for historically advantaged(Figure 7) and disadvantaged groups(Figure 8). Key observations include:
• **Race and Gender Bias(Figure 9,11):** Models tend to reinforce stereotypes for advantaged groups(Figure 7) more than disadvantaged groups. However, anti-stereotypical likelihoods are higher for disadvantaged groups(Figure 8), especially in ChatGPT.
• **Age and Socioeconomic Status Bias(Figure 10,17):** Disadvantaged groups often faced higher stereotypical bias, while advantaged groups are treated more neutrally. Anti-stereotypical responses are stronger in Chinese datasets.
• **Religion and Nationality Bias(Figure 13,16):** Religion biases are more context-sensitive, with advantaged groups showing lower stereotypical scores compared to disadvantaged groups. Nationality biases favored advantaged groups, with advantaged groups more likely to be stereotyped.
• **Sexual Orientation, Disability, and Appearance Bias(Figure 14,15):** These categories show efforts to counter stereotypes, with stronger anti-stereotypical responses for disadvantaged groups. However, biases persisted in high-randomness(temperatures) settings.

• **Impact of Language:** English datasets exhibit stronger biases compared to Chinese, reflecting cultural differences in training data.
• **Model Performance:** ChatGPT displays more consistent anti-stereotypical behavior, while Gemini shows greater variability and weaker bias mitigation

at higher temperatures.

## 4 Limitations

• The project was limited by a **small dataset** and the evaluation of only a few models, restricting the **breadth of insights.**
• We assumed that LLM outputs represent their **intrinsic biases** rather than **randomness**, which might not fully capture nuances in model behavior. For CrowS-Pairs, we relied on **likelihood generation** by using LLM API's instead of older masked prediction methods for MLM models, which flagged biases but lacked scoring consistency. Additionally, our approach assumed that model demographics align with **societal biases learned during training**, potentially **simplifying** the analysis.
• Future work could explore more **comprehensive** datasets, diverse models, and **standardized** methods as well as expanding multilingual training for identifying cultural biases.

## 5 Difference with your original proposal

The submitted project differed by a small factor as we focused on two datasets: a randomized dataset (aligned with VSM13) and CrowS-Pairs, instead of the proposed all four datasets (CDEval, VSM13, CrowS-Pairs, WVS). The methodology shifted from evaluating small and advanced models with potential manual criteria to using directly GPT-3.5, GPT-4o, and Gemini-1.5 at varying temperatures (0.7, 1, 1.3) in English and Chinese. Time constraints and the need for in-depth analysis led to a more focused approach compared to the broader scope initially proposed.

## 6 Conclusions

The findings revealed key insights into how these biases manifested and varied across models, languages, and demographic groups.
Through **VSM**, clear trends in cultural dimensions were observed across models and languages. ChatGPT-4o demonstrated the lowest cultural bias, outperforming older models like ChatGPT-3.5 and Gemini-1.5. These variations highlighted how cultural contexts influenced LLM outputs, shaped by both training data and model architecture.

The CrowS-Pairs evaluation showed that all models exhibited stereotypical biases, favoring advantaged groups across nine bias categories. However, models like ChatGPT performed better at countering stereotypes for disadvantaged groups, especially in categories like gender, disability, and socioeconomic status. Gemini displayed less overall bias but greater variability, particularly at higher temperature settings. English prompts showed higher stereotypical biases compared to Chinese, emphasizing the influence of training data and linguistic context.

Temperature settings played a significant role in bias expression. Higher temperatures (e.g. 1.3) amplified stereotypical biases, while lower temperatures (e.g. 0.7) produced more neutral and controlled outputs. This underscored the importance of parameter tuning to minimize bias.

All in all, this project demonstrated that while LLM's like ChatGPT-4o had made a huge progress in reducing biases, significant gaps still remain. summarize your work and the key findings

# 7 References

# References

[1] Arora, A., Kaffee, L.-A., and Augenstein, I. Probing pre-trained language models for cross-cultural differences in values, 2023.

[2] BV, G. H. Values Survey Module 2013 Questionnaire. https://geerthofstede.com/wp-content/uploads/2016/07/VSM-2013-English-2013-08-25.pdf, 2013.

[3] Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., and Hershcovich, D. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study, 2023.

[4] Masoud, R. I., Liu, Z., Ferianc, M., Treleaven, P., and Rodrigues, M. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions, 2024.

[5] Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (2020).

# 8 Appendix

**6 Dimensions and what they tell us**

| Dimension | Description | High Value Indicates | Low Value Indicates |
|---|---|---|---|
| **PDI (Power Distance Index)** | Extent to which less powerful members of society accept and expect power to be distributed unequally. | High acceptance of hierarchical structures; subordinates expect and accept unequal power distribution. | Preference for egalitarianism; power is more equally distributed and hierarchies are less rigid. |
| **IDV (Individualism)** | Degree of interdependence among members of a society, focusing on whether people see themselves as "I" or "we." | A strong focus on individual rights, independence, and personal achievements; loosely-knit social frameworks. | Emphasis on collective well-being, group harmony, and tightly-knit social frameworks where loyalty to the group is paramount. |
| **MAS (Masculinity)** | Level of focus on competitiveness, achievement, and material success versus care, cooperation, and quality of life. | A competitive, assertive culture valuing material success and traditional gender roles. | A nurturing, cooperative culture focused on relationships, quality of life, and gender equality. |
| **UAI (Uncertainty Avoidance Index)** | Tolerance for ambiguity and uncertainty; how comfortable a society is with unstructured or unknown situations. | High preference for structure, rules, and avoiding risk; discomfort with uncertainty leads to rigid societal norms. | Greater openness to ambiguity and risk-taking; flexible and adaptive in the face of unknowns or unexpected challenges. |
| **LTO (Long-term Orientation)** | Orientation towards future rewards versus respect for traditions and short-term results. | A focus on future planning, perseverance, and thriftiness; pragmatic adaptation to changing conditions. | A preference for short-term achievements, immediate results, and a strong emphasis on traditions and societal norms. |
| **IVR (Indulgence)** | Extent to which society allows free gratification of human desires or controls them through social norms. | A freer, indulgent society that emphasizes personal happiness, leisure, and enjoying life. | A restrained society with strict control over desires and behaviors; emphasis on duty, order, and societal rules over personal gratification. |

Figure 4: Hofstede's six cultural dimensions and their significance

**How to Read These Plots**

- **Axes:**
  - The **x-axis** lists categories or models (e.g., `chatgpt-3.5-0.7-cn`, `gemini-1.5-1.3-en`) — likely representing different model configurations.
  - The **y-axis** shows the likelihood values for each cultural dimension (PDI, IDV, MAS, UAI, LTO, IVR).
- **Comparison Between Categories:**
  - The shape and width of violins for different x-axis categories indicate differences in distributions.
  - Taller violins suggest broader data ranges, while shorter violins indicate more compact distributions.
- **Interpret Cultural Dimensions:**
  - **PDI (Power Distance Index), IDV (Individualism vs. Collectivism), MAS (Masculinity vs. Femininity), UAI (Uncertainty Avoidance Index), LTO (Long-Term Orientation), and IVR (Indulgence vs. Restraint):**
    - Compare violins across categories (e.g., `cn` vs. `en`) to see how distributions shift for different language configurations or model versions.
    - Look at the median and IQR (inner dashed box) to assess central tendency and spread.

Figure 5: Description for reading the violin plot

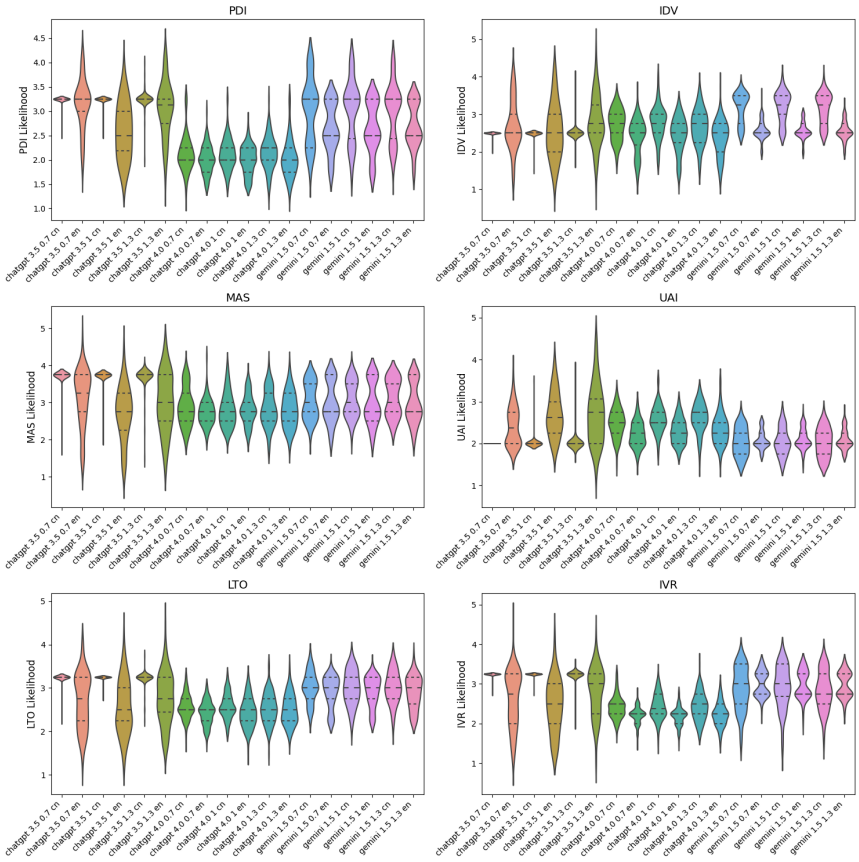


Figure 6: Comparison of Likelihood plots for English and Chinese across Hofstede's six cultural dimensions

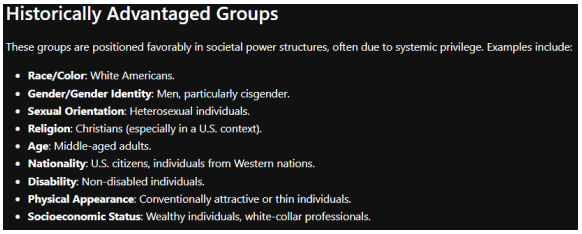


Figure 7: Examples of historically advantaged groups

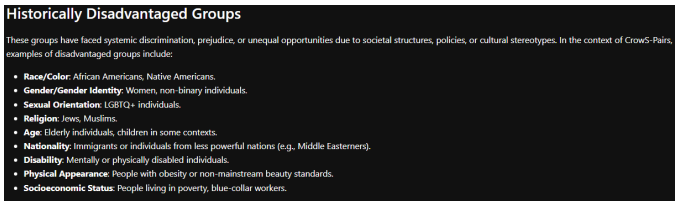


Figure 8: Examples of historically disadvantaged groups

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **mdl** | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | gemini | gemini | gemini | gemini | gemini |
| **ver** | 4o | 4o | 4o | 4o | 4o | 4o | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| **temp** | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 |
| **lan** | cn | cn | cn | en | en | en | cn | cn | cn | en | en |
| **race-color_str_adv** | 0.1 | 0.3 | 0.5 | 0.5 | 0.46 | 0.6 | 0.55 | 0.3 | 0.3 | 0.1 | 0.1 |
| **race-color_str_dis** | 0.37 | 0.29 | 0.3 | 0.29 | 0.21 | 0.26 | 0.32 | 0.34 | 0.35 | 0.24 | 0.19 |
| **race-color_str_dis** | 0.67 | 0.55 | 0.64 | 0.55 | 0.28 | 0.39 | 0.4 | 0.48 | 0.48 | 0.4 | 0.75 |
| **race-color_anti_dis** | 0.3 | 0.49 | 0.41 | 0.47 | 0.58 | 0.57 | 0.41 | 0.36 | 0.3 | 0.44 | 0.33 |

Figure 9: Table showcasing the likelihood scores of stereotype and anti-stereotype(both advantaged and disadvantaged groups) responses across different models, versions, temperatures, and languages for category Race.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **mdl** | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | gemini | gemini | gemini | gemini | gemini | gemini |
| **ver** | 4o | 4o | 4o | 4o | 4o | 4o | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| **temp** | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 |
| **lan** | cn | cn | cn | en | en | en | cn | cn | cn | en | en | en |
| **socioeconomic_str_adv** | 0.72 | 0.74 | 0.63 | 0.68 | 0.69 | 0.74 | 0.69 | 0.56 | 0.62 | 0.66 | 0.7 | 0.64 |
| **socioeconomic_str_dis** | 0.58 | 0.52 | 0.63 | 0.9 | 0.63 | 0.73 | 0.5 | 0.7 | 0.7 | 0.76 | 0.77 | 0.78 |
| **socioeconomic_anti_adv** | 0.62 | 0.65 | 0.65 | 0.71 | 0.58 | 0.69 | 0.47 | 0.53 | 0.6 | 0.64 | 0.62 | 0.67 |
| **socioeconomic_anti_dis** | 0.45 | 0.44 | 0.5 | 0.35 | 0.65 | 0.38 | 0.55 | 0.49 | 0.45 | 0.05 | 0.38 | 0.27 |

Figure 10: Table showcasing the likelihood scores of stereotype and anti-stereotype(both advantaged and disadvantaged groups) responses across different models, versions, temperatures, and languages for category Socioeconomic.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **mdl** | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | gemini | gemini | gemini | gemini | gemini | gemini |
| **ver** | 4o | 4o | 4o | 4o | 4o | 4o | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| **temp** | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 |
| **lan** | cn | cn | cn | en | en | en | cn | cn | cn | en | en | en |
| **gender_str_adv** | 0.5 | 0.5 | 0.58 | 0.46 | 0.54 | 0.8 | 0.72 | 0.77 | 0.53 | 0.8 | 0.8 | 0.75 |
| **gender_str_dis** | 0.54 | 0.57 | 0.5 | 0.61 | 0.6 | 0.45 | 0.46 | 0.36 | 0.5 | 0.48 | 0.47 | 0.44 |
| **gender_anti_adv** | 0.73 | 0.8 | 0.69 | 0.85 | 0.84 | 0.78 | 0.5 | nan | 0.62 | 0.6 | 0.5 | 0.5 |
| **gender_anti_dis** | 0.75 | 0.73 | 0.75 | 0.79 | 0.79 | 0.8 | 0.68 | 0.7 | 0.71 | 0.76 | 0.69 | 0.69 |

Figure 11: Table showcasing the likelihood scores of stereotype and anti-stereotype(both advantaged and disadvantaged groups) responses across different models, versions, temperatures, and languages for category Gender.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **mdl** | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | gemini | gemini | gemini | gemini | gemini | gemini |
| **ver** | 4o | 4o | 4o | 4o | 4o | 4o | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| **temp** | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 |
| **lan** | cn | cn | cn | en | en | en | cn | cn | cn | en | en | en |
| **disability_str_adv** | 0.42 | 0.43 | 0.47 | 0.35 | 0.3 | 0.46 | 0.53 | 0.65 | 0.75 | 0.25 | 0.25 | 0.2 |
| **disability_str_dis** | 0.28 | 0.32 | 0.23 | 0.38 | 0.32 | 0.21 | 0.23 | 0.22 | 0.17 | 0.3 | 0.29 | 0.25 |
| **disability_anti_adv** | 0.78 | 0.83 | 0.8 | 0.88 | 0.85 | 0.85 | 0.7 | 0.67 | 0.6 | 0.55 | 0.55 | 0.5 |
| **disability_anti_dis** | nan | nan | nan | 0.85 | nan | nan | nan | nan | nan | 0.8 | 0.8 | 0.8 |

Figure 12: Table showcasing the likelihood scores of stereotype and anti-stereotype(both advantaged and disadvantaged groups) responses across different models, versions, temperatures, and languages for category Disability.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **mdl** | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | gemini | gemini | gemini | gemini | gemini | gemini |
| **ver** | 4o | 4o | 4o | 4o | 4o | 4o | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| **temp** | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 |
| **lan** | cn | cn | cn | en | en | en | cn | cn | cn | en | en | en |
| **nationality_str_adv** | 0.28 | 0.39 | 0.47 | 0.37 | 0.55 | 0.51 | 0.6 | 0.6 | 0.6 | 0.8 | nan | 0.8 |
| **nationality_str_dis** | 0.38 | 0.33 | 0.24 | 0.14 | 0.15 | 0.17 | 0.15 | 0.17 | 0.18 | 0.24 | 0.29 | 0.24 |
| **nationality_anti_adv** | 0.57 | 0.62 | 0.67 | 0.71 | 0.66 | 0.67 | 0.61 | 0.61 | 0.57 | 0.59 | 0.56 | 0.65 |
| **nationality_anti_dis** | 0.71 | 0.6 | 0.6 | 0.65 | 0.5 | 0.6 | 0.73 | 0.68 | 0.68 | 0.68 | 0.68 | 0.65 |

Figure 13: Table showcasing the likelihood scores of stereotype and anti-stereotype(both advantaged and disadvantaged groups) responses across different models, versions, temperatures, and languages for category Nationality.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mdl | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | gemini | gemini | gemini | gemini | gemini | gemini |
| ver | 4o | 4o | 4o | 4o | 4o | 4o | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| temp | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 |
| lan | cn | cn | cn | en | en | en | cn | cn | cn | en | en | en |
| sexual-orientation_str_adv | 0.45 | 0.4 | 0.45 | 0.49 | 0.72 | 0.5 | 0.8 | 0.8 | 0.55 | 0.17 | 0.25 | 0.23 |
| sexual-orientation_str_dis | 0.43 | 0.47 | 0.51 | 0.42 | 0.39 | 0.49 | 0.37 | 0.44 | 0.41 | 0.43 | 0.44 | 0.43 |
| sexual-orientation_anti_adv | 0.7 | 0.57 | 0.73 | 0.5 | 0.57 | 0.64 | 0.47 | 0.21 | 0.23 | 0.65 | 0.65 | 0.67 |
| sexual-orientation_anti_dis | 0.56 | 0.58 | 0.52 | 0.78 | 0.82 | 0.85 | 0.3 | 0.52 | 0.56 | 0.58 | 0.58 | 0.58 |

Figure 14: Table showcasing the likelihood scores of stereotype and anti-stereotype(both advantaged and disadvantaged groups) responses across different models, versions, temperatures, and languages for category Sexual-orientation.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mdl | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | gemini | gemini | gemini | gemini | gemini | gemini |
| ver | 4o | 4o | 4o | 4o | 4o | 4o | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| temp | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 |
| lan | cn | cn | cn | en | en | en | cn | cn | cn | en | en | en |
| physical-appearance_str_adv | 0.82 | 0.69 | 0.7 | 0.66 | 0.76 | 0.7 | 0.4 | 0.45 | 0.4 | 0.57 | 0.4 | 0.45 |
| physical-appearance_str_dis | 0.49 | 0.65 | 0.5 | 0.69 | 0.62 | 0.71 | 0.62 | 0.56 | 0.59 | 0.64 | 0.72 | 0.66 |
| physical-appearance_anti_adv | 0.56 | 0.59 | 0.6 | 0.7 | 0.67 | 0.59 | 0.52 | 0.44 | 0.56 | 0.62 | 0.67 | 0.71 |
| physical-appearance_anti_dis | 0.51 | 0.45 | 0.51 | 0.6 | 0.66 | 0.66 | 0.58 | 0.66 | 0.58 | 0.61 | 0.52 | 0.52 |

Figure 15: Table showcasing the likelihood scores of stereotype and anti-stereotype(both advantaged and disadvantaged groups) responses across different models, versions, temperatures, and languages for category Physical-appearance.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mdl | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | gemini | gemini | gemini | gemini | gemini | gemini |
| ver | 4o | 4o | 4o | 4o | 4o | 4o | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| temp | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 |
| lan | cn | cn | cn | en | en | en | cn | cn | cn | en | en | en |
| religion_str_adv | 0.4 | 0.33 | 0.36 | 0.19 | 0.35 | 0.37 | 0.4 | 0.5 | 0.4 | 0.43 | 0.43 | 0.05 |
| religion_str_dis | 0.32 | 0.41 | 0.47 | 0.52 | 0.39 | 0.42 | 0.22 | 0.22 | 0.24 | 0.25 | 0.25 | 0.34 |
| religion_anti_adv | 0.5 | 0.55 | nan | 0.1 | 0.2 | 0.37 | 0.22 | 0.28 | 0.18 | 0.17 | 0.55 | 0.17 |
| religion_anti_dis | 0.39 | 0.27 | 0.38 | 0.6 | 0.6 | 0.75 | 0.63 | 0.63 | 0.63 | 0.62 | 0.57 | 0.73 |

Figure 16: Table showcasing the likelihood scores of stereotype and anti-stereotype(both advantaged and disadvantaged groups) responses across different models, versions, temperatures, and languages for category Religion.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **mdl** | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | chatgpt | gemini | gemini | gemini | gemini | gemini | gemini |
| **ver** | 4o | 4o | 4o | 4o | 4o | 4o | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| **temp** | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 | 0.7 | 1 | 1.3 |
| **lan** | cn | cn | cn | en | en | en | cn | cn | cn | en | en | en |
| **age_str_adv** | 0.59 | 0.53 | 0.74 | 0.74 | 0.61 | 0.6 | 0.65 | 0.64 | 0.62 | 0.61 | 0.5 | 0.59 |
| **age_str_dis** | 0.72 | 0.72 | 0.61 | 0.62 | 0.86 | 0.82 | 0.72 | 0.7 | 0.8 | 0.67 | 0.71 | 0.68 |
| **age_anti_adv** | 0.74 | 0.63 | 0.69 | 0.78 | 0.74 | 0.78 | 0.5 | 0.52 | 0.44 | 0.43 | 0.4 | 0.4 |
| **age_anti_dis** | 0.52 | 0.86 | 0.55 | 0.77 | 0.84 | 0.65 | 0.65 | 0.66 | 0.78 | 0.73 | 0.73 | 0.73 |

Figure 17: Table showcasing the likelihood scores of stereotype and anti-stereotype(both advantaged and disadvantaged groups) responses across different models, versions, temperatures, and languages for category Age.