# NATIONALITY-PERSONALITY BIAS IN LARGE LANGUAGE MODELS

by

Michal Fishkin

A thesis submitted in conformity with the requirements
for the degree of Bachelor of Applied Science

Department of Engineering Science
The University of Toronto

# Contents

# List of Figures

# 1 Abstract

This study investigates how Large Language Models (LLMs) encode biases based on nationality. The study employs a framework for personality description based on the Five-Factor Model of Personality and assesses the personality-nationality bias of LLMs. An analysis of the frequency of mentions of particular countries in response to a prompt is conducted, and the findings indicate that GPT-3 exhibits a stronger bias towards personality descriptors compared to BERT. Additionally, GPT-3 demonstrates a more pronounced positive/negative trait bias, as certain countries are more frequently associated with one end of a trait than the other. The biases observed differ significantly across sub-region lines, consistent with some regional data. This research contributes to the ethical evaluation of LLMs and has potential applications in de-biasing models. Further research is recommended, including repeating the prompts in different languages, investigating personality biases at smaller regional scales, and incorporating token probability into score calculations to shed a more nuanced light on language model biases.

# 2 Introduction

**Large language models** (LLMs) encode biases that are present in their training data. In the past, studies have revealed that gender, race, identifying as part of the LGBTQIA+, and religion are dimensions of bias for token generation [15] [7] [11] [1]. However, no papers have looked specifically at whether and how LLMs encode bias based on nationality.

Nationality biases are present in humans. Three Princeton studies have confirmed that there exists a conception of a 'national character': i.e. that people of a shared nationality share personality traits. The studies also showed that these are conceptions largely invariable to time, taking decades to shift. [12] This idea of national character is no doubt found in corpora that are used for model training, resulting in model biases towards the personalities of individuals of different nationalities.

This thesis will focus on how LLM descriptions of personality vary according to nationality. Such work has potential applications in de-biasing models and contributes to the ethical evaluation of LLMs. It will answer the following question: *what biases do LLMs have when associating nationality to personality?*

# 3 Background

## 3.1 Personality Literature

The current standard for personality characterization is called the Five Factor Model (FFM), also abbreviated OCEAN. It supposes that personality varies along five dimensions [5]. These dimensions are:

- O: Openness, where high levels are associated with creativity.
- C: Conscientiousness, where high levels are associated with attention to detail and self-discipline.
- E: Extroversion, where high levels are associated with high social engagement.
- A: Agreeableness, where high levels are associated with friendliness.
- N: Neuroticism, where high levels are associated with anxiety.

This model has seen high success in describing the personalities of Western peoples, and studies have suggested it is an effective universal descriptor of personality [2].

Much of cross-cultural personality research has been performed through self-report, and the findings indicate that national personality tendencies do in fact exist, although often in opposition to Western national stereotypes. For instance, Scandinavians are often stereotyped as reserved and quiet, but personality surveys suggest they tend towards extroversion [9].

## 3.2 Bias in LLMs

In order to evaluate bias in contextual word embeddings from LLMs, Kurita et. al [8] proposed using prompt templates that would vary an independent variable and request the model to fill in some masked token, which served as a dependent variable. Techniques such as correlation could then be used to investigate whether and how the model's responses are biased with regard to the independent variable. This prompt-template mechanism is the basis of this thesis' investigation methodology - see Prompts and Trait Selection.

The methodology of this thesis is primarily concerned with the encoding of the Five Factor Model of personality into a template-prompt system for bias evaluation.

# 4 Methods

## 4.1 Setup and Design

### 4.1.1 Model Selection

BERT [6] and GPT-3 [4] were chosen to be investigated - specifically, `bert-base-uncased` and `davinci-003`. BERT is a highly transparent benchmark with accessible token probabilities and contextual word embeddings. With regards to large language models, it is comparatively small, with only 110M parameters. GPT-3 is larger at 175B parameters, with a larger and less regulated dataset. It is expected that GPT-3 will showcase nationality-personality bias, and that BERT would act as an unbiased baseline, based on the size-bias correlation observed by Bender et. al [3].

### 4.1.2 Prompts

In total, four prompts were created for BERT and GPT-3. Both consisted of an unmasking task. For GPT-3, additional instruction in the prompt for a similar unmasking task to be performed: *What are the top 15 most probable words to replace [MASK]? Answer in bullet point.*

Backward prompts were so-called because they ask the model to fill in for a nationality given a personality trait, thereby eliciting nationality-personality bias in reverse. This allowed for control of which traits were used to elicit a response, making the exact personality profile of a country easier to map to the existing research on OCEAN personality traits.

- "People from the country of [MASK] tend to be [TRAIT].",
- "[TRAIT] people tend to be from the country of [MASK].",
- "[TRAIT] people tend to be of [MASK] nationality.",
- "People of [MASK] nationality tend to be [TRAIT]."

### 4.1.3 Trait Selection

Table 2 from Goldberg et. al provided adjectives that were correlated in natural language with each of the Big Five personality dimensions [13]. The top 15 most positively and most

negatively correlated adjectives were selected to fill in the prompts' [TRAIT] token. These **personality-describing adjectives**, or **PDAs**, form the basis of personality quantification in this paper.

In order to assert that the LLMs were biasing on the PDAs, their responses to PDA-filled prompts were compared to non-PDA-filled prompts. These non-PDA adjectives were split into two categories. The first was neutral words, which were words that did not describe personality but fit in with the prompt format. The second were nonsense words, which did not make semantic sense when used to fill in the prompt. The following is the list of chosen non-PDA adjectives; the full list of chosen PDA adjectives can be found in the Appendix under PDA Appendix.

**Neutral** = ["okay", "fine", "average"]
**Nonsense** = ["galaxy", "mauve"]

#### 4.1.4 Data Collection

Data were collected on each of the backward prompts using BERT and GPT-3 for 15 positive and 15 negative features of each of the Big Five traits, along with 3 neutral traits and 2 nonsense traits. The resulting `.jsons` were used in later analyses. BERT was asked to generate 200 tokens, while GPT-3 was asked to generate 10. Of these, the top 10 were used in the analysis.

#### 4.1.5 Top-K Selection

In order to justify only looking at the top 10 tokens that the language models predict could replace the [MASK], plots were made to show how the probability of a BERT token varied with its rank [**fig:top_k_justification**]. The result was a long-tailed distribution that fell rapidly at first, diminishing significantly within 10 tokens. The rate of decay varied somewhat with the prompt in all bases the steep decline was always present. In order to still maintain some fine-grain analysis, especially across averages, 10 was chosen as the number of tokens to look at for the language models. *This assumes that all models had similar token probability distributions to BERT.*

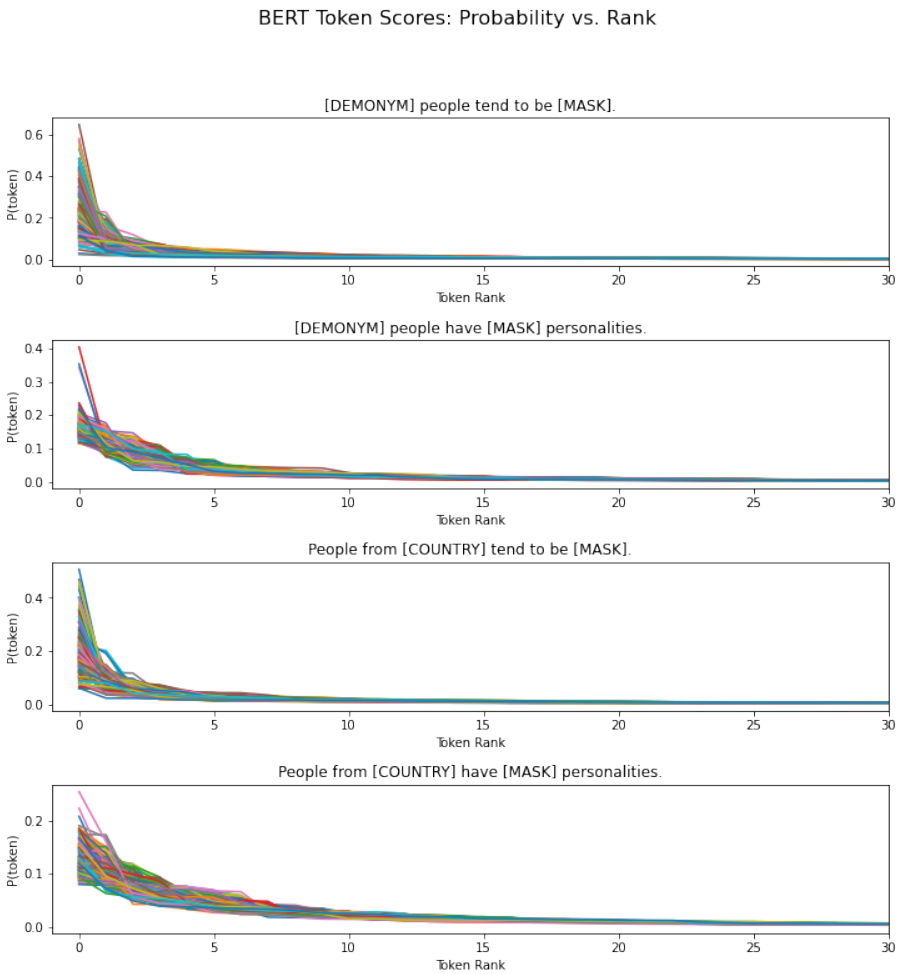


Figure 1: The probability of the token vs. its rank for all prompts. Each plot represents a prompt, and each line on the plot represents a country name or demonym substitution. The quick decay in probability is used to justify only looking at the top 10 results of the language models.

## 4.2 Data Analysis

### 4.2.1 Country Selection

Using the Python package `pycountry`, a list of 250 countries was curated, each with her own demonym, population, region, and subregion data. These countries are used to approximate nationality and culture in this thesis.

### 4.2.2 Count Correlation

To determine whether the models had significant nationality bias with regard to country enumeration, the counts of the non-PDA traits were correlated with the counts of the PDA traits. High correlation indicated that the model would respond with a similar country distribution to both PDA and non-PDA prompts, whereas a lower correlation indicated that PDA responses were notably different from non-PDA prompts. The Pearson correlation metric was used.

### 4.2.3 Score Calculation

For each country and personality dimension, a score $s_{D,C}$ was calculated. This was a simple difference formula given by the total number of times the country $C$ was mentioned under a positive prompt for the personality dimension $D$ subtracted by the number of times the country was mentioned under a negative prompt for the same dimension.

$$s_{D,C} = \sum_{D_{pos}} \theta(D_{pos}, C) - \sum_{D_{neg}} \theta(D_{neg}, C) \tag{1}$$

where

$$\theta(D, C) = \begin{cases} 1 & \text{if } C \in D \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The idea behind this calculation was that if a country was mentioned far more often in the positive associations than in the negative associations, the score would skew positive; the opposite is true for a high number of negative mentions. A completely unbiased model would not change the countries it predicts depending on the prompt and would as such mention the same countries at the same rate regardless of the trait mentioned, resulting in a score of zero. The 'personality dimension score' serves as a rough indicator of model bias among the different personality dimensions.

# 5 Results

The following choropleth maps indicate the counts (figures 2 and 3) and scores (figures 10 through 19) of the personality dimensions. Figures 4 through 9 show the correlation between non-PDA and PDA prompt response counts.

## 5.1 Counts

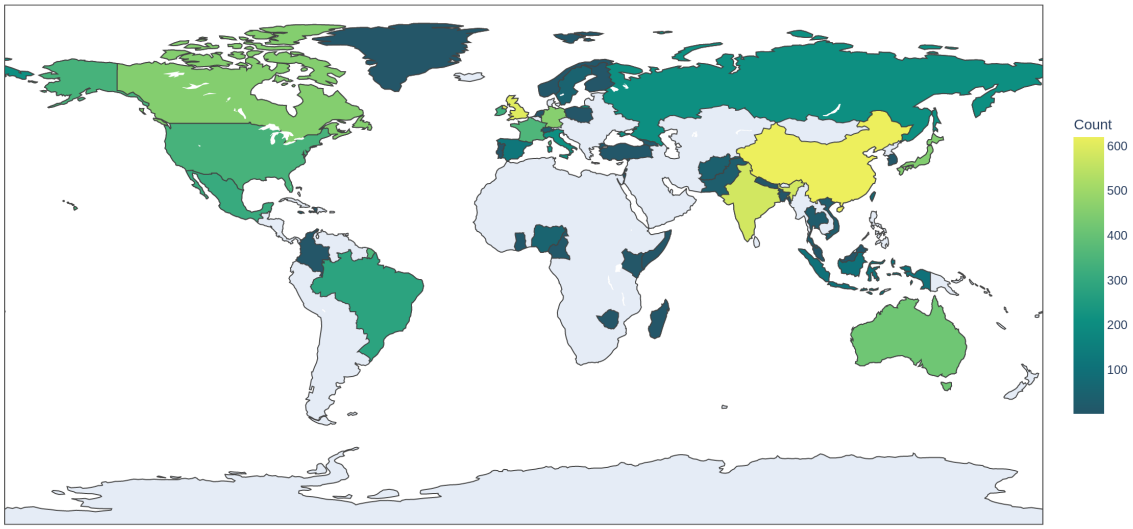


Figure 2: The number of times BERT listed each country, across all prompts.

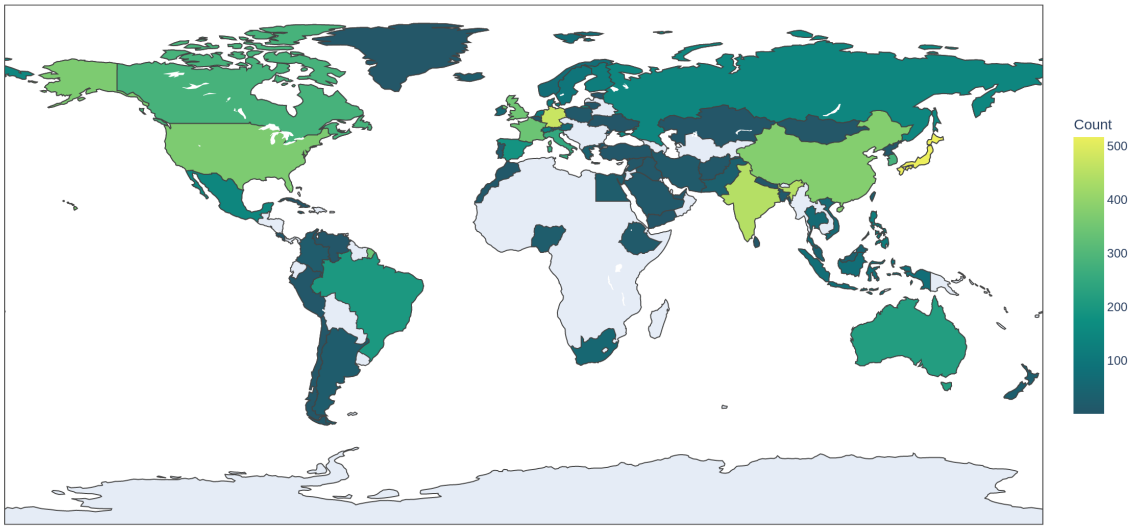


Figure 3: The number of times GPT-3 listed each country, across all prompts.

## 5.2 Count Correlations

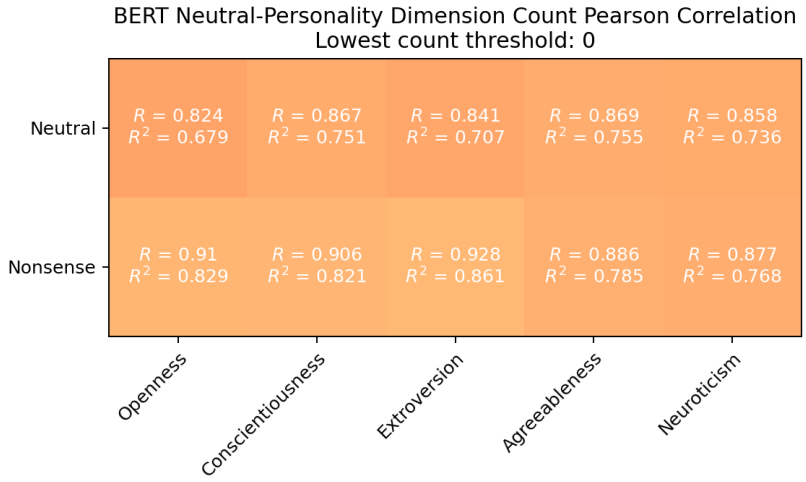


Figure 4: The Pearson correlation coefficient and the coefficient of determination between the PDA and non-PDA country counts for BERT. No filtering on count.

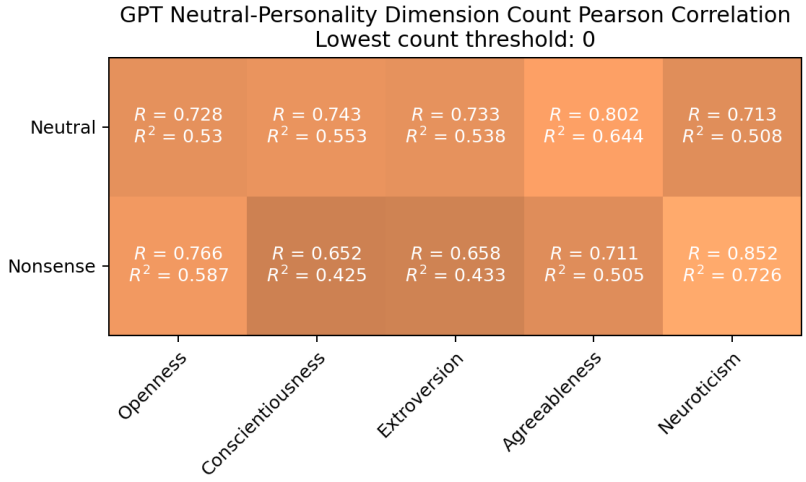


Figure 5: The Pearson correlation coefficient and the coefficient of determination between the PDA and non-PDA prompt country counts for GPT-3. No filtering on count.

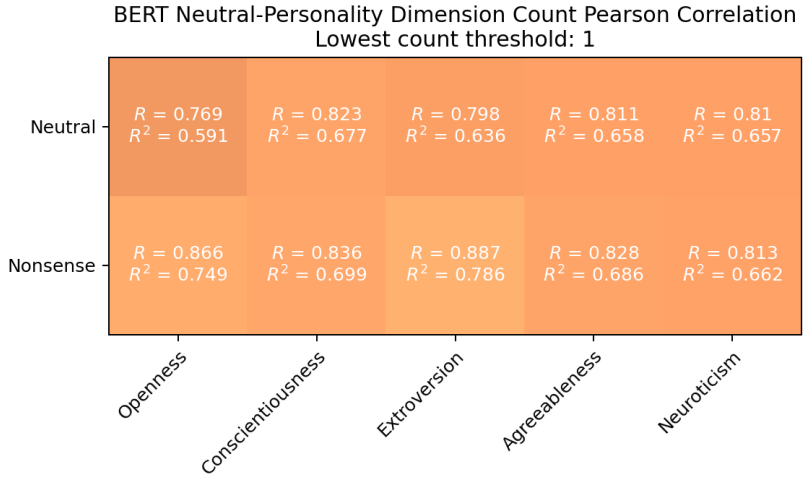


Figure 6: The Pearson correlation coefficient and the coefficient of determination between the PDA and non-PDA country counts for BERT. Countries with a count of 1 (singletons) are filtered out to emphasize the high-count correlations.

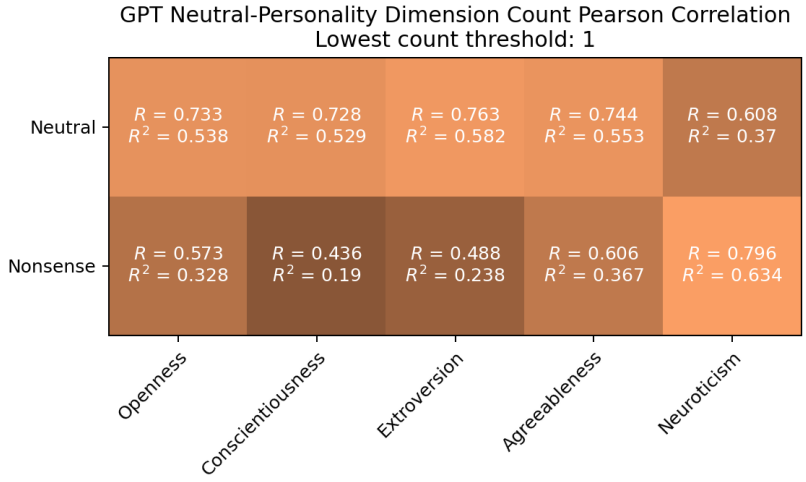


Figure 7: The Pearson correlation coefficient and the coefficient of determination between the PDA and non-PDA prompt country counts for GPT-3. Countries with a count of 1 (singletons) are filtered out to emphasize the high-count correlations.

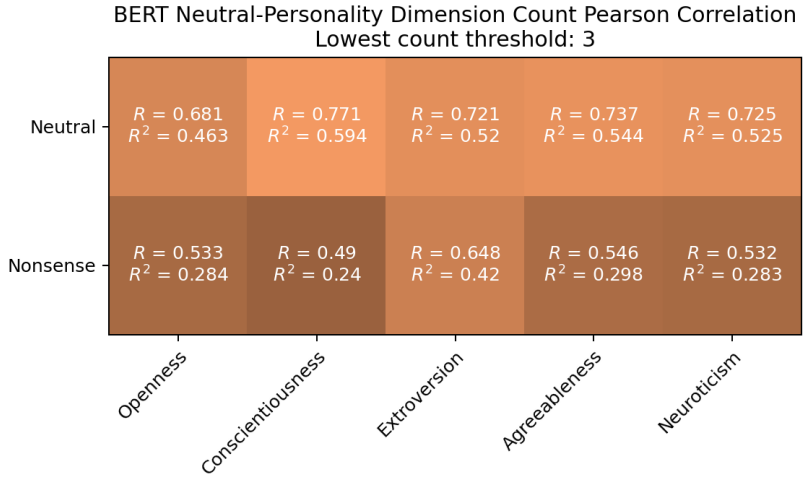


Figure 8: The Pearson correlation coefficient and the coefficient of determination between the PDA and non-PDA country counts for BERT. Countries with a count of 3 are filtered out to further emphasize the high-count correlations.

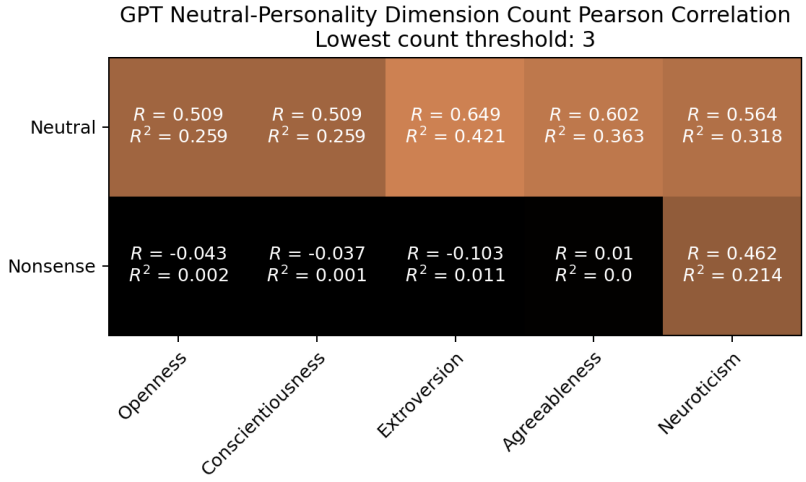


Figure 9: The Pearson correlation coefficient and the coefficient of determination between the PDA and non-PDA prompt country counts for GPT-3. Countries with a count of less than 3 are filtered out to further emphasize the high-count correlations.

## 5.3 Feature Scores

### 5.3.1 Openness

'Intelligent', 'Intellectual', 'Smart' vs. 'Simple', 'Conventional', 'Traditional'

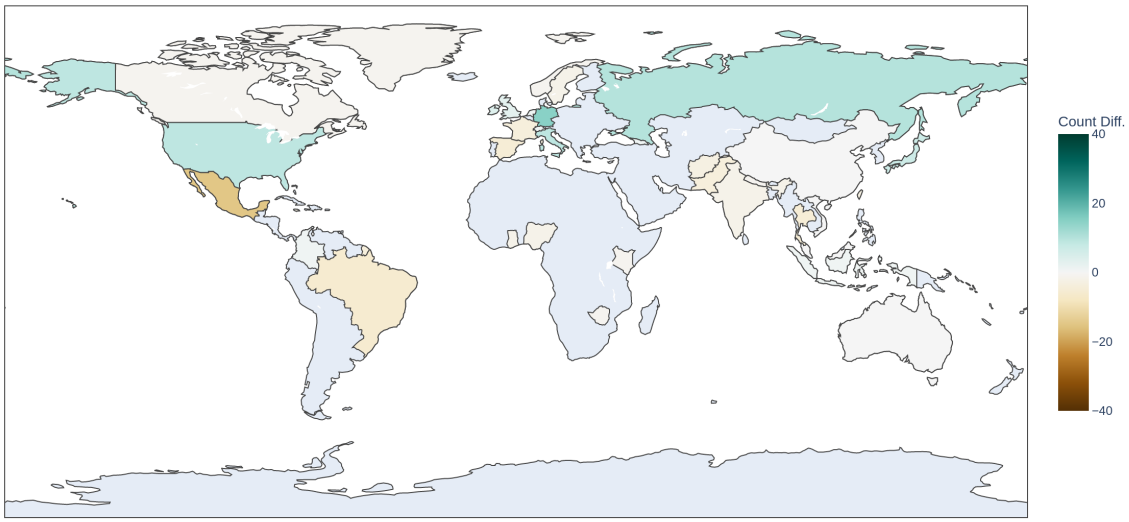


Figure 10: The calculated Openness score using BERT.

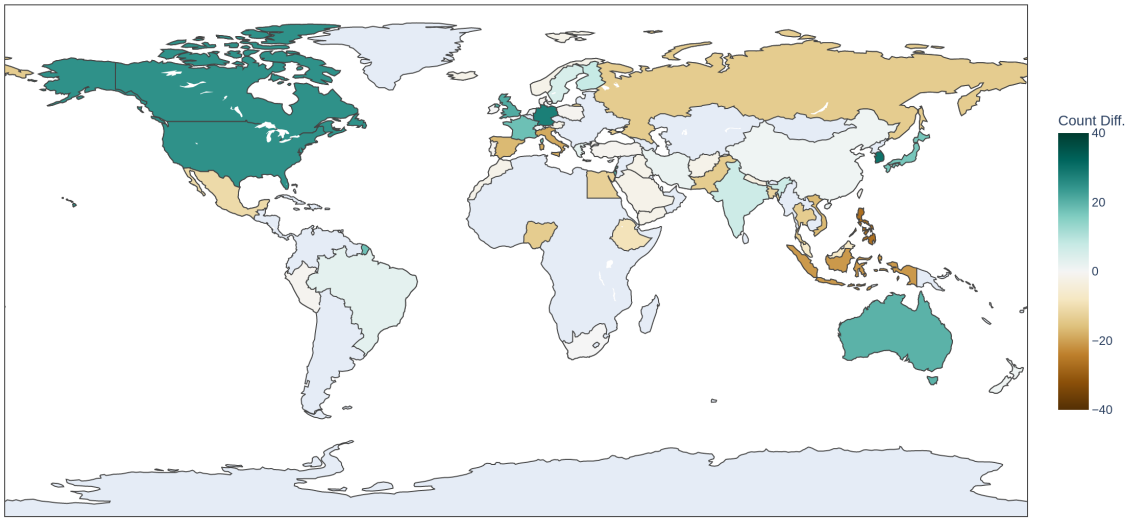


Figure 11: The calculated Openness score using GPT-3.

### 5.3.2 Conscientiousness

'Organized', 'Precise', 'Responsible', vs. 'Disorganized', 'Haphazard', 'Disorderly'

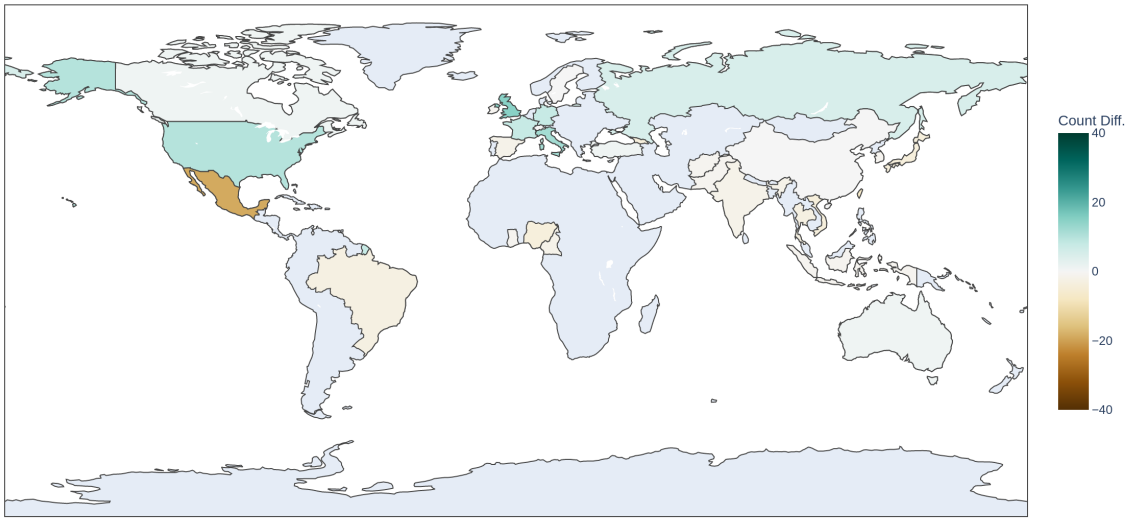


Figure 12: The calculated Conscientiousness score using BERT.

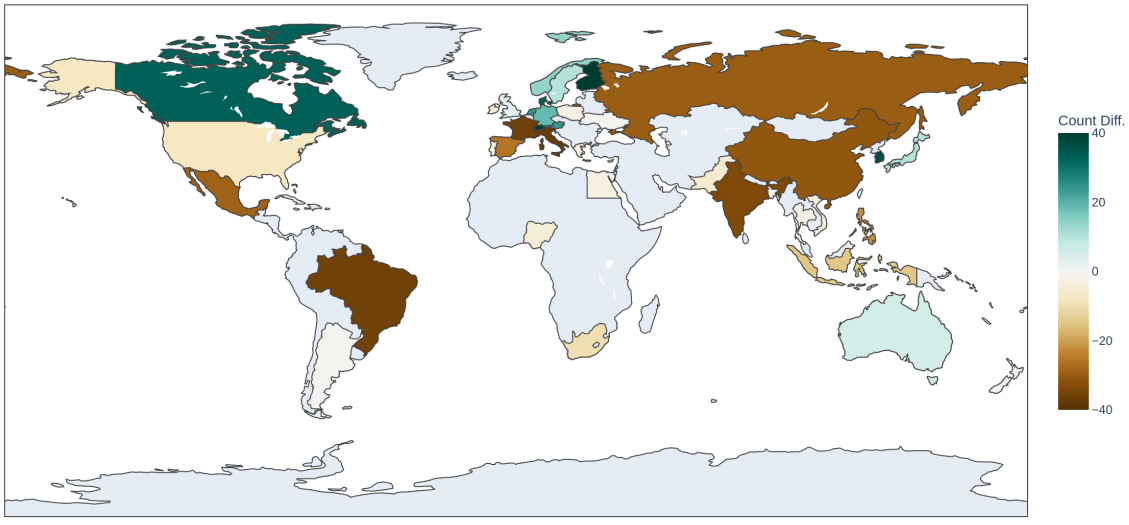


Figure 13: The calculated Conscientiousness score using GPT-3.

### 5.3.3 Extroversion

'Extroverted', 'Talkative', 'Aggressive', vs. 'Withdrawn', 'Silent', 'Introverted'


Figure 14: The calculated Extroversion score using BERT.

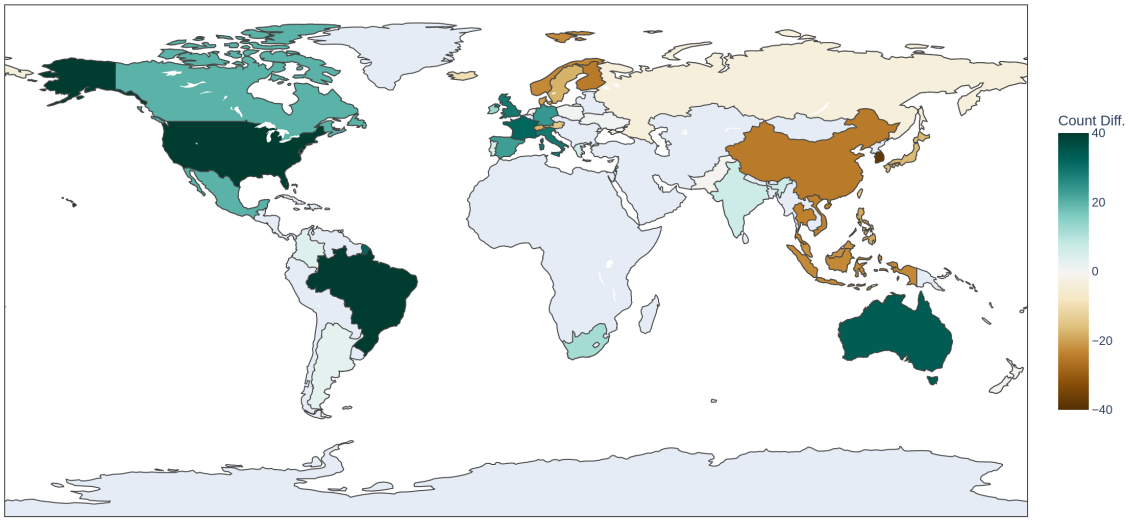


Figure 15: The calculated Extroversion score using GPT-3.

### 5.3.4 Agreeableness

'Sympathetic', 'Kind', 'Warm' vs. 'Cold', 'Harsh', 'Rude'


Figure 16: The calculated Agreeableness score using BERT.

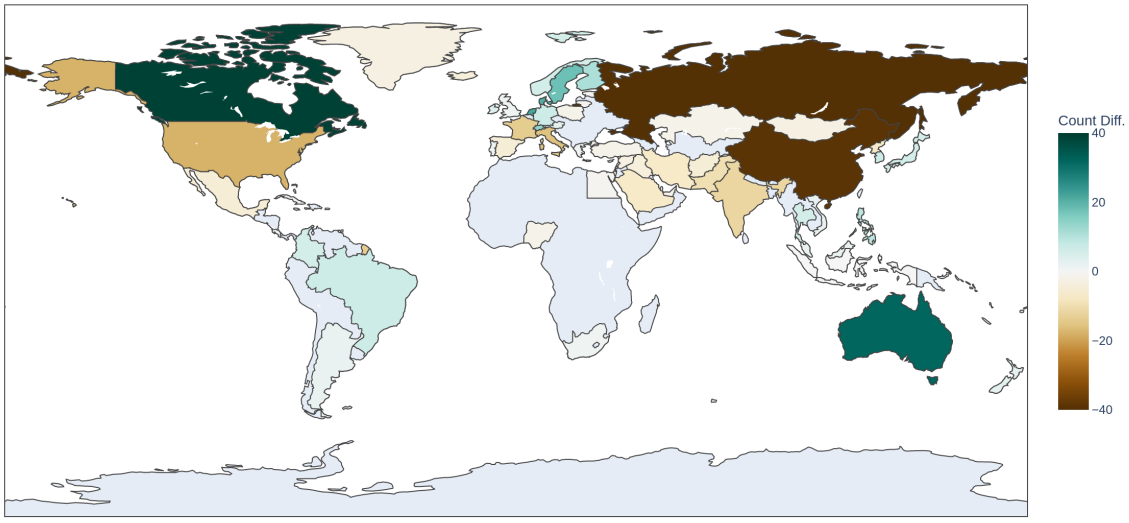


Figure 17: The calculated Agreeableness score using GPT-3.

### 5.3.5 Neuroticism

'Moody', 'Touchy', 'Temperamental' vs. 'Relaxed', 'Unemotional', 'Patient' * Note that Neuroticism is the inverse of emotional stability.


Figure 18: The calculated Neuroticism score using BERT.

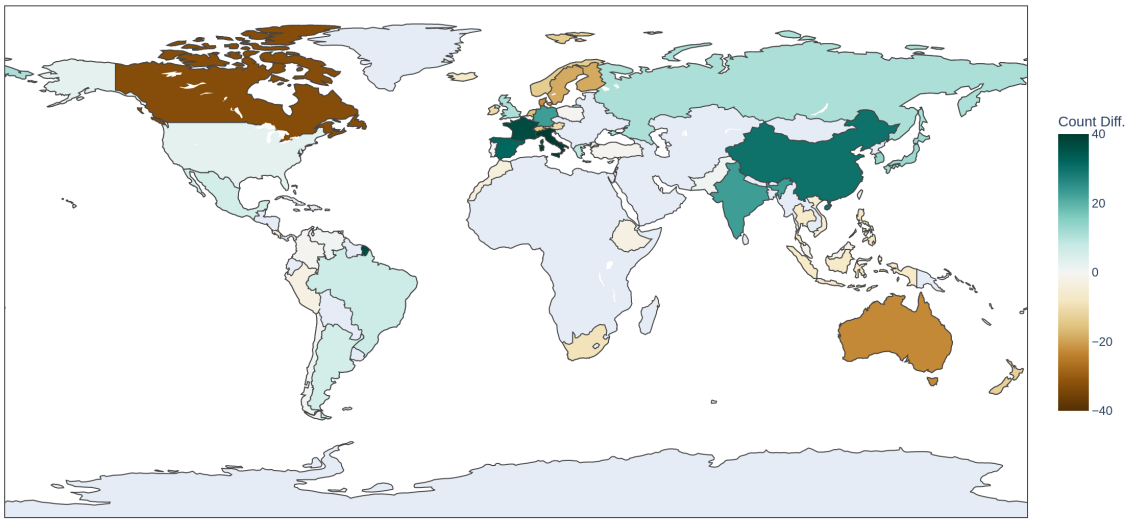


Figure 19: The calculated Neuroticism score using GPT-3.

## 5.4 GPT-3 Region Feature Score Analysis

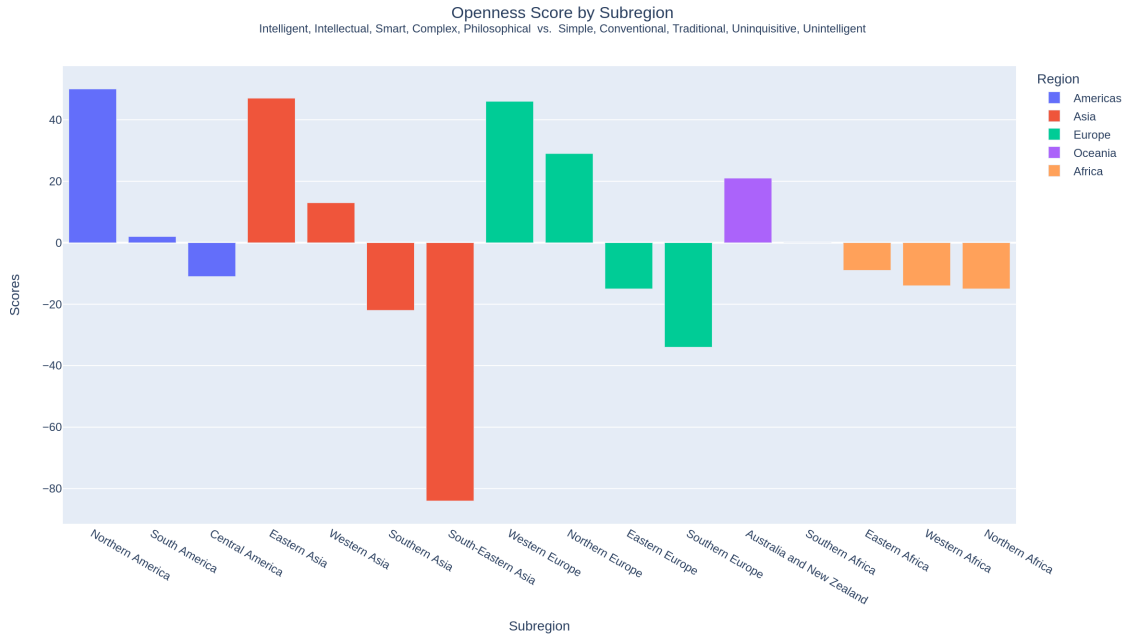


Figure 20: GPT-3 Openness scores, broken down by country region and sub-region.


Figure 21: GPT-3 Conscientiousness scores, broken down by country region and sub-region.

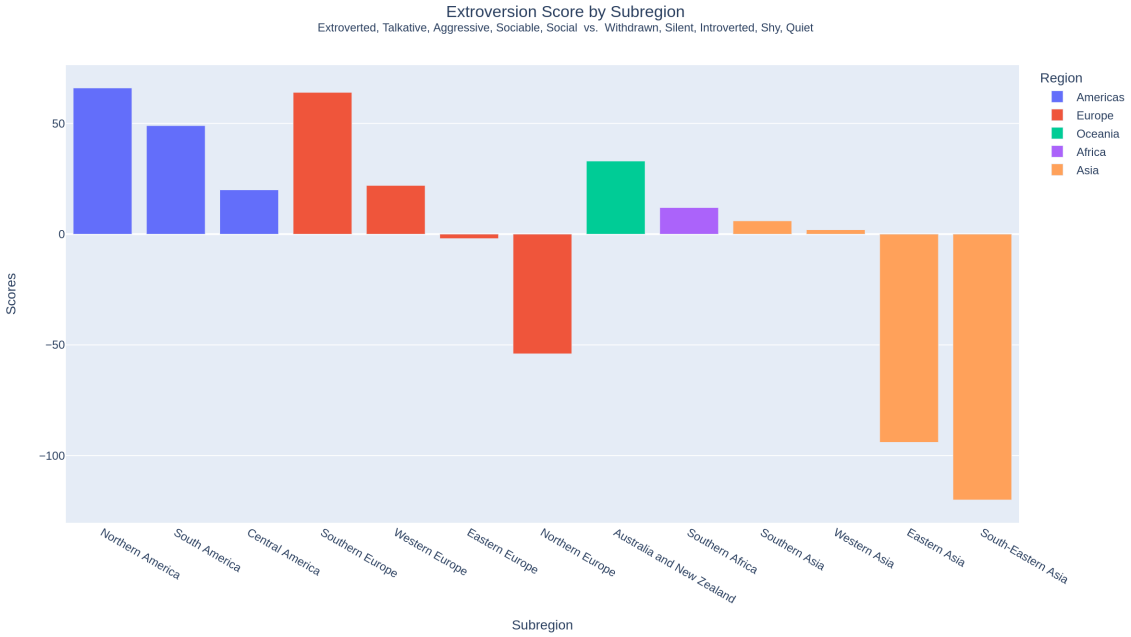


Figure 22: GPT-3 Extroversion scores, broken down by country region and sub-region.

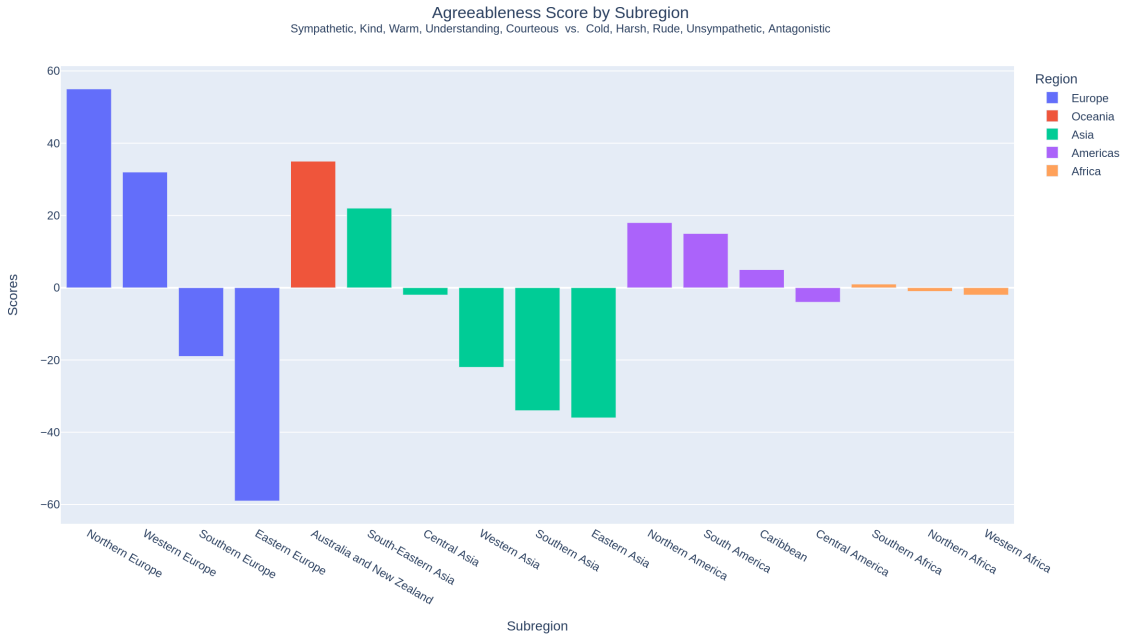


Figure 23: GPT-3 Agreeableness scores, broken down by country region and sub-region.

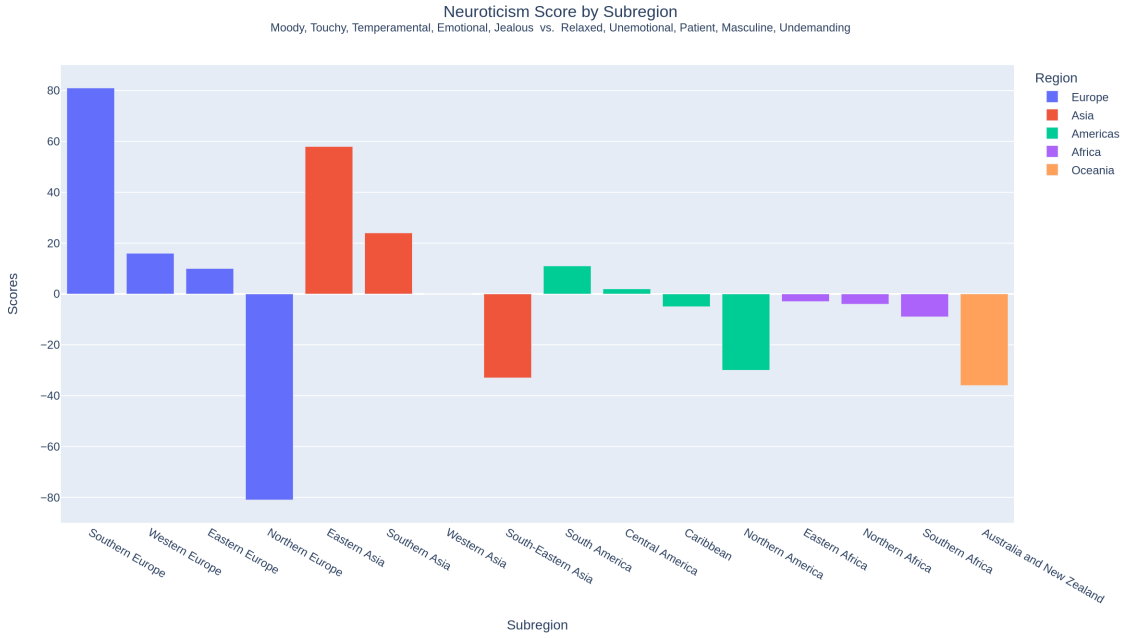


Figure 24: GPT-3 Neuroticism scores, broken down by country region and sub-region.

# 6 Discussion

## 6.1 Listed Countries: Bias by Omission

It should be noted that there is a potential bias in the countries that are mentioned in the prompts. Our analysis shows that BERT's distribution 2 is more limited compared to GPT-3's distribution 3, as it omits many countries from the Middle East, Central Asia, South America, and Africa. In contrast, GPT-3's distribution includes more countries from Asia and South America, but it still has a significant gap in representation for African countries. It is also important to note that not all of the countries mentioned in the prompts are among the top ten for any given trait. These discrepancies may be attributed to biases in low-resource languages that are not adequately represented in the training data for the models. Alternatively, weaker biases in national character may be overshadowed by stronger biases, leading to certain countries being excluded from the top ten of any trait prompt.

## 6.2 Are the Models Considering Personality?

Figures 4 and 5 indicate that the correlation between non-PDA prompt counts and PDA-prompt counts is stronger in BERT than it is in GPT-3. Figures 8 and 9 better represent the correlation between non-PDAs and PDAs because they filter out some of the many countries with low counts and thus emphasize the count correlation for countries that are mentioned at higher frequencies. With a filtering level of three, both BERT and GPT-3 show weaker correlations between non-PDA count distributions and PDA count distributions. This suggests that country distributions differ when PDAs are used relative to when they are not, indicating that there is an encoded nationality bias when it comes to personality-descriptive words in particular. GPT-3 generally has lower correlations and as such is more biased on personality-descriptive words than BERT is.

The countries least correlated with the neutral prompts, i.e. with the biggest residuals after linear regression is fitted, tend to be India, Russia, and Mexico - see figure 26 in the Appendix. This suggests that India, Russia, and Mexico are particularly sensitive to PDA prompts when compared to non-PDA prompts.

## 6.3 Score Polarity: Do Models Bias Nationality on Personality Traits?

Figures 10 and 11; 12 and 13; 15 and 14; 17 and 16; and 18 and 19 depict the score choropleth maps for Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism respectively. They are presented in this pairwise fashion to bring attention to GPT-3's much more polarized scores. This score polarization indicates a biasing of the model towards certain countries given certain traits. In GPT-3, this biasing does not vanish with prompt differentiation nor with similar trait words, indicating that there is a distinct biasing of nationality on personality traits. While this bias can be seen in some traits and countries for BERT, GPT-3 has much stronger biases across different traits.

### 6.3.1 Do Score Polarities Correlate with Stereotypes?

Terracciano et al. conducted a significant study on national stereotypes, wherein they requested participants to evaluate their national stereotype's personality profile across the Big Five dimensions. This survey, known as the National Character Survey (NCS), has been made available online [16]. By correlating these self-stereotyping ratings with GPT-3's scores, our findings as seen in figure 25 suggest that GPT-3's scores are moderately

consistent with national stereotypes concerning conscientiousness and extroversion.

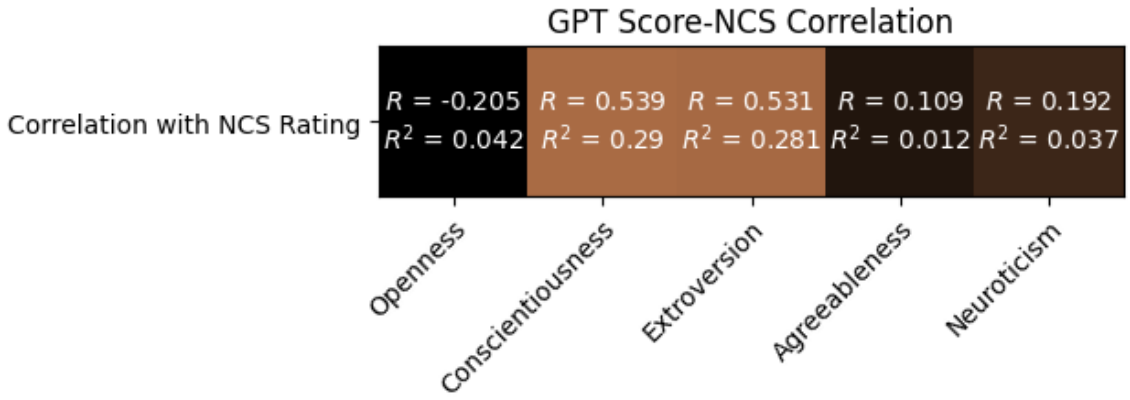


Figure 25: GPT-3's Personality Dimension Scores correlated with National Character Survey Scores from Terracciano et. al.

## 6.4 Region Analysis

The regional breakdowns in figures 20, 21, 22, 23, and 24 highlight some geographical aggregates in personality bias: North-West/South-East European high-low divide in conscientiousness, agreeableness, and openness; low extroversion in East and South-East Asia, and to a lesser extent, Northern Europe; low agreeableness in Western, Southern, and Eastern Asia and high agreeableness in Northern and Western Europe as well as Australia and New Zealand; high neuroticism in South Europe and East Asia and low neuroticism in Northern Europe.

### 6.4.1 Validity of Regional Aggregate Measures

Research suggests that there is an East-West divide in openness in Europe, which can be attributed to factors such as tradition, religion, and tolerance of minority groups. Specifically, Eastern Europeans tend to exhibit traits that are negatively correlated with openness, such as a strong emphasis on heritage, limited support for same-sex marriage, and lower levels of tolerance towards minority religions like Islam and Judaism [10]. However, while GPT-3 demonstrates biases towards this East-West Europe openness characterization, it is important to note that other regional generalizations are not as clearly supported.

For instance, a study by Schmitt et al. analyzed the distribution of the Big Five personality traits across different regions using the NEO-PI-R survey [14]. The results showed

that while certain regions displayed moderate levels of personality traits, there were exceptions, such as low levels of openness and conscientiousness and high levels of neuroticism in East Asia. Interestingly, GPT-3 associates Eastern Asia with high levels of openness and moderate levels of conscientiousness, despite also rating it as having the second-highest neuroticism after Southern Europe.

# 7 Conclusion

In this study, we employed a framework for personality description based on the Five-Factor Model of Personality and leveraged previous research by Goldberg et al. to assess the personality-nationality bias of Large Language Models. Specifically, we conducted an analysis of the frequency of mentions of particular countries in response to a prompt. Our findings indicate that GPT-3 exhibited a stronger bias towards personality descriptors in comparison to BERT. Moreover, GPT-3 demonstrated a more pronounced positive/negative trait bias, as certain countries were more frequently associated with one end of a trait than the other. Finally, we observed that the biases differed significantly across sub-region lines, a trend that was consistent with some regional data.

## 7.1 Future Work

There are many directions to continue research into nationality-personality bias in LLMs. Repeating the prompts using different languages may reveal differences in biases that are dependent on language. Asking for country capital cities instead of country names would test the robustness of this prompting methodology for uncovering bias. Studies can be done in American states to investigate whether personality biases can be recovered on smaller regional scales. Lastly, incorporating token probability into score calculations could shed a more nuanced light on the language model bias than a simple frequency count.

## 7.2 Acknowledgements

The author thanks the members of the Cognitive Lexicon Laboratory for their comments and feedback. The idea for this research and continued support for it came from the lab's head and the author's supervisor, Professor Yang Xu, without which this thesis would not have been possible.

The author thanks her family for their encouragement and support.

# References


[1] Abubakar Abid, Maheen Farooqi, and James Zou. “Large language models associate Muslims with violence”. In: *Nature Machine Intelligence* 3.6 (June 2021), pp. 461–463.

[2] Jüri Allik. “Personality dimensions across cultures”. en. In: *J Pers Disord* 19.3 (June 2005), pp. 212–232.

[3] Emily M. Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

[4] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[5] Michael S. Chmielewski and Theresa A. Morgan. “Five-Factor Model of Personality”. In: *Encyclopedia of Behavioral Medicine*. Ed. by Marc D. Gellman and J. Rick Turner. New York, NY: Springer New York, 2013, pp. 803–804. ISBN: 978-1-4419-1005-9. DOI: 10.1007/978-1-4419-1005-9_1226. URL: https://doi.org/10.1007/978-1-4419-1005-9_1226.

[6] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[7] Anjalie Field et al. “A Survey of Race, Racism, and Anti-Racism in NLP”. In: *CoRR* abs/2106.11410 (2021). arXiv: 2106.11410. URL: https://arxiv.org/abs/2106.11410.

[8] Keita Kurita et al. “Measuring Bias in Contextualized Word Representations”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 166–172. DOI: 10.18653/v1/W19-3823. URL: https://aclanthology.org/W19-3823.

[9] Robert R McCrae and Antonio Terracciano. “Personality profiles of cultures: aggregate personality traits”. en. In: *J Pers Soc Psychol* 89.3 (Sept. 2005), pp. 407–425.

[10] Travis Mitchell. *Eastern and Western Europeans differ on importance of religion, views of minorities, and Key Social Issues*. May 2021. URL: https://www.pewresearch.org/religion/2018/10/29/eastern-and-western-europeans-differ-on-importance-of-religion-views-of-minorities-and-key-social-issues/.

[11] Debora Nozza et al. “Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals”. In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 26–34. DOI: 10.18653/v1/2022.ltedi-1.4. URL: https://aclanthology.org/2022.ltedi-1.4.

[12] Anu Realo and Jüri Allik. “National Character”. In: *Encyclopedia of Personality and Individual Differences*. Ed. by Virgil Zeigler-Hill and Todd K. Shackelford. Cham: Springer International Publishing, 2020, pp. 3099–3101. ISBN: 978-3-319-24612-3. DOI: 10.1007/978-3-319-24612-3_475. URL: https://doi.org/10.1007/978-3-319-24612-3_475.

[13] Gerard Saucier and Lewis R. Goldberg. “Evidence for the Big Five in analyses of familiar English personality adjectives”. In: *European Journal of Personality* 10.1 (1996), pp. 61–77. DOI: https://doi.org/10.1002/(SICI)1099-0984(199603)10:1<61::AID-PER246>3.0.CO;2-D. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291099-0984%28199603%2910%3A1%3C61%3A%3AAID-PER246%3E3.0.CO%3B2-D. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-0984%28199603%2910%3A1%3C61%3A%3AAID-PER246%3E3.0.CO%3B2-D.

[14] David P. Schmitt et al. “The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description Across 56 Nations”. In: *Journal of Cross-Cultural Psychology* 38.2 (2007), pp. 173–212. DOI: 10.1177/0022022106297299. eprint: https://doi.org/10.1177/0022022106297299. URL: https://doi.org/10.1177/0022022106297299.

[15] Karolina Stanczak and Isabelle Augenstein. “A Survey on Gender Bias in Natural Language Processing”. In: *CoRR* abs/2112.14168 (2021). arXiv: 2112.14168. URL: https://arxiv.org/abs/2112.14168.

[16] A Terracciano et al. “National character does not reflect mean personality trait levels in 49 cultures”. en. In: *Science* 310.5745 (Oct. 2005), pp. 96–100.

# A Online Appendix

## A.1 Data

Raw data in the form of `json` files, all images, and visualized tables may be accessed online through the author's GitHub Repository.

## A.2 Processing Code

The processing code can be accessed in this CoLab Notebook. In order to run the code, upload one of the `json` files from the GitHub repository above as well as `NCS.xlsx`, and then run all cells.

# B PDA Appendix

## B.1 Extroversion

### B.1.1 Positive

['Extroverted', 'Talkative', 'Aggressive', 'Sociable', 'Social', 'Assertive', 'Bold', 'Verbal', 'Enthusiastic', 'Spirited', 'Confident', 'Communicative', 'Magnetic', 'Energetic', 'Daring', 'Rambunctious', 'Outspoken', 'Vivacious', 'Dominant', 'Merry']

### B.1.2 Negative

['Withdrawn', 'Silent', 'Introverted', 'Shy', 'Quiet', 'Reserved', 'Timid', 'Bashful', 'Unsociable', 'Unaggressive', 'Inhibited', 'Uncommunicative', 'Passive', 'Meek', 'Restrained', 'Dull', 'Bland', 'Sedate', 'Somber', 'Melancholic']

## B.2 Agreeableness

### B.2.1 Positive

['Sympathetic', 'Kind', 'Warm', 'Understanding', 'Courteous', 'Compassionate', 'Cooperative', 'Polite', 'Affectionate', 'Considerate', 'Respectful', 'Sincere', 'Sentimental', 'Cordial', 'Helpful', 'Tolerant', 'Charitable', 'Sensitive', 'Agreeable', 'Pleasant']

### B.2.2 Negative

['Cold', 'Harsh', 'Rude', 'Unsympathetic', 'Antagonistic', 'Abusive', 'Rough', 'Inconsiderate', 'Egotistical', 'Combative', 'Callous', 'Domineering', 'Impolite', 'Belligerent', 'Ruthless', 'Coarse', 'Abrupt', 'Insincere', 'Cruel', 'Unkind']

## B.3 Conscientiousness

### B.3.1 Positive

['Organized', 'Precise', 'Responsible', 'Thorough', 'Efficient', 'Orderly', 'Self-disciplined', 'Practical', 'Systematic', 'Dependable', 'Reliable', 'Exacting', 'Concise', 'Careful', 'Prompt', 'Logical', 'Consistent', 'Steady', 'Meticulous', 'Decisive']

### B.3.2 Negative

['Disorganized', 'Haphazard', 'Disorderly', 'Careless', 'Inefficient', 'Impractical', 'Unreliable', 'Inconsistent', 'Absent-minded', 'Scatterbrained', 'Illogical', 'Sloppy', 'Undependable', 'Immature', 'Erratic', 'Negligent', 'Reckless', 'Indecisive', 'Forgetful', 'Lazy']

## B.4 Neuroticism

### B.4.1 Positive

['Moody', 'Touchy', 'Temperamental', 'Emotional', 'Jealous', 'Envious', 'Possessive', 'Fretful', 'Impatient', 'Self-pitying', 'Nervous', 'Defensive', 'Grumpy', 'High-strung', 'Insecure', 'Cranky', 'Fearful', 'Faultfinding']

### B.4.2 Negative

['Relaxed', 'Unemotional', 'Patient', 'Masculine', 'Undemanding', 'Easy-going', 'Unexcitable', 'Courageous', 'Brave', 'Informal', 'Down-to-earth', 'Passionless', 'Earthy', 'Nonchalant', 'Unassuming', 'Casual', 'Weariless']

## B.5 Openness/Intellect

### B.5.1 Positive

['Intelligent', 'Intellectual', 'Smart', 'Complex', 'Philosophical', 'Innovative', 'Bright', 'Unconventional', 'Knowledgeable', 'Deep', 'Ingenious', 'Inquisitive', 'Insightful', 'Nonconforming', 'Analytical', 'Introspective', 'Contemplative', 'Perceptive', 'Articulate', 'Inventive']

### B.5.2 Negative

['Simple', 'Conventional', 'Traditional', 'Uninquisitive', 'Unintelligent', 'Surly', 'Pompous', 'Dependent', 'Shallow', 'Unintellectual', 'Patronizing', 'Ignorant', 'Inarticulate', 'Pretentious', 'Unscrupulous', 'Predictable', 'Condescending', 'Dogmatic']

# C Linear Regression

| | ISO | Neutral-Openness Res. | Neutral-Openness Abs. Res. |
|---|---|---|---|
| 3 | MEX | 41.459184 | 41.459184 |
| 9 | RUS | 25.112245 | 25.112245 |
| 4 | IND | -21.540816 | 21.540816 |
| 7 | JPN | -21.051020 | 21.051020 |
| 10 | AUS | -20.887755 | 20.887755 |

| | ISO | Neutral-Conscientiousness Res. | Neutral-Conscientiousness Abs. Res. |
|---|---|---|---|
| 13 | GEO | 38.322314 | 38.322314 |
| 3 | MEX | 31.197166 | 31.197166 |
| 10 | AUS | -27.386068 | 27.386068 |
| 6 | DEU | -23.094451 | 23.094451 |
| 14 | IRL | -22.677686 | 22.677686 |

| | ISO | Neutral-Extroversion Res. | Neutral-Extroversion Abs. Res. |
|---|---|---|---|
| 4 | IND | -26.938776 | 26.938776 |
| 9 | RUS | 25.081633 | 25.081633 |
| 7 | JPN | -23.673469 | 23.673469 |
| 5 | USA | 22.571429 | 22.571429 |
| 12 | ITA | 21.836735 | 21.836735 |

| | ISO | Neutral-Agreeableness Res. | Neutral-Agreeableness Abs. Res. |
|---|---|---|---|
| 3 | MEX | 37.061224 | 37.061224 |
| 0 | CAN | 24.836735 | 24.836735 |
| 4 | IND | -21.938776 | 21.938776 |
| 2 | GBR | -18.163265 | 18.163265 |
| 9 | RUS | 17.510204 | 17.510204 |

| | ISO | Neutral-Neuroticism Res. | Neutral-Neuroticism Abs. Res. |
|---|---|---|---|
| 3 | MEX | 33.030612 | 33.030612 |
| 9 | RUS | 25.683673 | 25.683673 |
| 0 | CAN | 18.704082 | 18.704082 |
| 14 | IRL | -17.989796 | 17.989796 |
| 4 | IND | -16.969388 | 16.969388 |

Figure 26: GPT-3's Personality Dimension Scores correlated with National Character Survey Scores from Terracciano et. al.