Born Differently Makes a Difference: Counterfactual Study of Bias in Biography Generation from a Data-to-Text Perspective

Anonymous ACL submission

Abstract

How do personal attributes affect biography generation? Addressing this question requires an identical pair of biographies where only the personal attributes of interest are different. However, it is rare in the real world. To address this, we propose a counterfactual methodology from a data-to-text perspective, manipulating the personal attributes of interest while keeping the co-occurring attributes unchanged. We first validate that the fine-tuned Flan-T5 model generates the biographies based on the given attributes. This work expands the study of gender-centered bias analysis in text generation. Our results confirm the well-known bias in gender and also show the bias in regions, in both individual and its related co-occurring attributes in semantic machining and sentiment.

1 Introduction

To what extent do personal attributes affect biography content? Biography consists of the facts of personal attributes (Bamman and Smith, 2014). Current research has shown that biographies from Wikipedia reflect bias from society (Hube, 2017), such as well-known bias in gender (Graells-Garrido et al., 2015; Wagner et al., 2015; Konieczny and Klein, 2018; Tripodi, 2023; Reagle and Rhue, 2011) and culture (Samoilenko and Yasseri, 2014; Beytía, 2020; Baltz, 2022). However, personal attributes are compounded. For instance, religions could be prevalent based on geography (Buttimer, 2006). This results in the challenge of isolating co-occurring attributes and evaluating the effect of personal attributes alone. Answering this question directly would require paired-wise comparisons of biographies that are identical except for the particular personal attribute of interest (Field et al., 2022; Fang et al., 2023). It would allow us to measure the causal effect of the attribute value (treatment) on biography text (outcome) (Holland, 1986; Pearl, 2009). However, having such identical biographies is rare and nearly impossible.

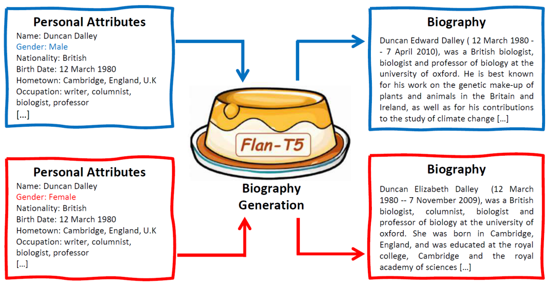


Figure 1: An example from the Synthbio dataset (Yuan et al., 2021). We measure semantic matching and sentiment in the true and generated biography (top-right) based on the personal attributes (top-left). Counterfactuals (bottom-right) replace the personal attribute (male, top-left) with a different one (female, bottom-left).

Additionally, Wikipedia biographies mostly consist of notable people.[1] Large language models (LLMs) have shown the capability of remembering training data (Roberts et al., 2020; Li and Flanigan, 2023) and generating factual biographies based on only names of celebrities (Maudslay et al., 2019; Yuan et al., 2021).

In light of these observations, we propose a counterfactual methodology based on a data-to-text framework. We formulate the task as generating biographies by given attributes (Figure 1, top-left → top-right). By doing so, we maintain a controllable setting, enforcing biography generation focusing on the given attributes, thus allowing us to study the effect of individual personal attributes. To mitigate the effect of celebrities, we do our analysis on carefully designed fictional biographies, the SynthBio dataset (Yuan et al., 2021), where fictional names and related personal attributes are controlled by human-LLMs collaboration.

Since personal attributes are compounded and diverse, we consider two universal types of personal attributes, i.e., *gender* and *region*. We evaluate the generated biographies from two dimensions: *se-*

[1] https://en.wikipedia.org/wiki/Wikipedia:Generally_notable_people

mantic matching (Rebuffel et al., 2021), evaluating how the biography correctly represents the meaning in the attributes; and, *sentiment* (Gatti et al., 2015), measuring how positive or negative the tone of the text is. We first show a significant difference among generated biographies from different gender and region groups in both semantic matching and sentiment (Section 3).

We further perform counterfactual analysis by explicitly manipulating the personal attributes of interest (Section 4). We compare the generated biographies (Figure 1, top-right vs., bottom-right, respectively) from true attributes (male, top-left) vs. manipulated attributes (female, bottom-left). We ask *how would the generated biographies change if the given personal attributes were changed*?

We show that disentangling individual and related co-occurring personal attributes, LLMs fine-tuned on the Wikibio dataset encode gender and region bias in semantic matching and sentiment, prompting further research in biography generation going beyond gender-centered (Liang et al., 2021), and general quality evaluations, e.g., ROUGE (Lin, 2004).

2 Methodology

Data We use the WikiBio dataset (Lebret et al., 2016) for training, consisting of 728,321 biographies from real English Wikipedia pages where the infobox and first paragraph from the articles are provided. On average, each infobox contains 12.5 personal attributes. We explicitly add the gender label (*male*, *female* or *non-binary/identifiable*), inferring from the pronouns in the paragraph (De-Arteaga et al., 2019), to the infobox. We remove the biographies where the nationality is not available.

To mitigate the cross-contamination of training and evaluation sets (Roberts et al., 2020; Li and Flanigan, 2023), we use the Synthbio dataset (Yuan et al., 2021) for evaluation, which is a synthetic dataset consisting of structured attributes—which we refer as *true attributes*—describing fictional individuals. It consists of 2,237 infoboxes and each infobox has on average 19 personal attributes and multiple fictional biographies. Statistics of the datasets can be found in Appendix A.

Personal Attributes of Interest We study the impact of two common personal attributes: (1) *Gender*. Following the gender attributes in the Synthbio dataset, we consider *male*, *female*, and *non-binary*; and, (2) *Region*. We map the 40 nationalities to 6 regions: *North America* (NA), *Europe* (EU), *Middle East* (ME), *Asia–Pacific* (AP), *South/Latin America* (SA), and *Africa* (AF).[2]

Semantic Matching and Sentiment We study the generated biographies from two dimensions: (1) *Semantic Matching*. We use Data-QuestEval (Rebuffel et al., 2021), a reference-free semantic evaluator curated for data-to-text evaluation; and, (2) *Sentiment*. Since recent sentiment evaluators are deployed for social media text (Hutto and Gilbert, 2014; Camacho-collados et al., 2022) which is not suitable for our task, we use a lexical-based method, obtaining the sentiment score by retrieving SentiWords (Gatti et al., 2015), a dictionary associating positive or negative scores with approximately 155,000 words. We calculate the sentiment score of the biography by averaging the associated sentiment scores for each word.

Biography Generation Our biography data-to-text task can be formulated as:

$$Bio(m, co(m)) = f_{gen}(m, co(m)), \qquad (1)$$

where biography is generated by the mode f_{gen} given the peronal attribute of interest (m) and the co-occurring attributes ($co(m)$). We use Flan-T5-base (Chung et al., 2022), an instruction fine-tuned model, to generate biographies. Following Yuan et al. (2021), we construct the infobox as the data-to-text format described in Kale and Rastogi (2020)[3] and finetune Flan-T5-base on WikiBio for 10,000 steps on one P100 GPU, with a batch size of 8, to instruct the model to generate biography based on given attributes. To generate biographies on the Synthbio, we use a beam search of 5.

3 True Attributed Biography Generation

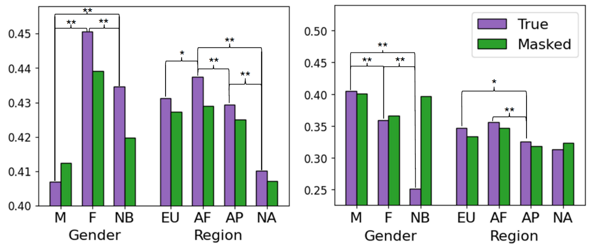
First, we validate that the fine-tuned Flan-T5 model generates biographies based on the given personal attributes. To explore the effect of personal attributes, we compare the semantic matching and sentiment on the generated biographies with true attributes (Equation (1)) against those without given the particular attribute (Masked), i.e.,

$$Bio(\phi, co(m)) = f_{gen}(\phi, co(m)). \qquad (2)$$

[2] The nationality-region table is provided in Appendix B.
[3] The detailed construction is provided in Appendix C.

Attributes	True	Masked	Counterfactual Raw(/Selected)
Gender			
Male	0.999	0.963	0.991
Female	0.972	0.514	0.978
Non-Binary	0.837	0.057	0.824
Overall	0.936	0.509	0.931
Region			
Europe	0.837	0.732	0.488/0.770
South/L. America	0.674	0.618	0.234/ -
Africa	0.805	0.573	0.432/0.856
Middle East	0.527	0.420	0.090/ -
Asia–Pacific	0.854	0.742	0.586/0.819
North America	0.939	0.833	0.740/ -
Overall	0.804	0.684	0.459/0.809

Table 1: Results of inferring personal attribute of interest from generated biographies.



(a) Semantic Matching (b) Sentiment

Figure 2: Semantic matching and sentiment for different attribute groups. Gender: (M=Male, F=Female, NB=Non-Binary); For true attributed biography (purple bars), pairwise significant differences are reported according to Welch's t-test at p<0.1 (*) and p<0.05 (**).

Model Validation Our fine-tuned Flan-T5 model outperforms the T5 model (Raffel et al., 2020) reported in the Synthbio dataset (Yuan et al., 2021), with a RougeL score of 26.4 (vs., 22.6) and a PARENT-F score (Dhingra et al., 2019) of 0.114 (vs., 0.049). We first validate whether the personal attribute of interest can be inferred from the biographies. Specifically, for gender, we use the pronouns as the proxy of gender (De-Arteaga et al., 2019) and compare it against the given gender attribute. For the region, since there is no direct method to predict the nationality from the biography, we consider whether the nationality or related country name is mentioned in the biography as the proxy of the nationality encoded in the biography. We do not train a classifier for nationality as the biography contains rich personal information—the classifier may remember the training instances instead of the nationality signals. We then group the results for nationality based on the region.

As shown in Table 1 (Column: True), for gender, we achieve higher than 0.8 accuracy across gender groups, confirming that the given gender is encoded in generated biographies. However, the results in region groups vary. To ensure the generation quality for our analysis, we consider regions with scores higher than 0.75: EU, AF, AP, and NA.

True Attributed Biography Do LLMs generate different biographies for different gender and nationality groups? Figure 2 shows that generated biographies are significantly different among different gender groups (purple bars, gender) in semantic matching and sentiment. For region, we observe significant differences in some region groups, e.g., AF vs., AP in both measurements, indicating the potential bias among region groups. However, we do not observe constant significant differences for any particular region.

True vs., Masked Attributed Biography To study the effect of individual personal attributes, we evaluate the semantic matching and sentiment of the generated biographies where given identical attributes but without attributes of interest (Figure 2, green bars). Compared to true attributed biographies (Figure 2, purple bars), we do not observe significant differences in gender and region. Potentially, the generation model can still infer the masked attributes given the co-occurring attributes (Table 1, Column: Masked). Masking the personal attributes alone is not effective in understanding the influence of individual personal attributes.

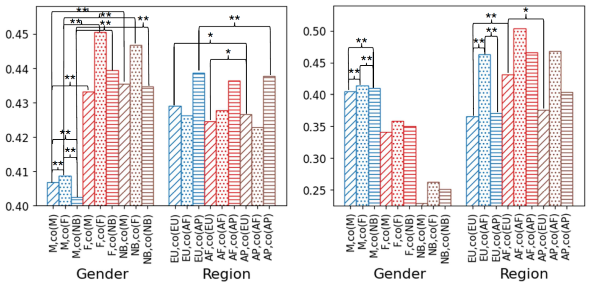
4 Counterfactual Attributed Generation

We apply our counterfactual methodology based on our fine-tuned Flan-T5 model. We manipulate only the personal attributes of interest and keep the co-occurring attribute unchanged to study the effect of individual attributes. Specifically, we change the personal attribute (Figure 1, male, top-left) to a different attribute (Figure 1, female, bottom-left) and compare the true (Equation (1)) and counterfactual attributed biographies (Figure 1, top-right vs., bottom-right, respectively), formulating as:

$$Bio(f, co(m)) = f_{gen}(f, co(m)), do(m \rightarrow f),$$

where $do(m \rightarrow f)$ denotes the do operator (Pearl, 2009), e.g., in Figure 1, changing the personal attribute male (m) to female (f).

We first investigate whether the counterfactual biographies encode the desired attributes via the same



(a) Semantic Matching (b) Sentiment

Figure 3: Semantic matching and sentiment for different attribute groups in counterfactual attributed biographies. Different colors and shapes represent different individual personal attributes, and co-occurring attributes, respectively. For brevity, we only show the pairwise significant differences related to groups *male* and *Europe*.

validation described in Section 3.[4] Table 1 (Counterfactual) shows that generated biographies adjust to the given counterfactual gender attributes. However, we observe that overall 45.9% biographies explicitly mention counterfactual nationalities. To ensure counterfactual biographies quality, we select nationalities that have a score larger than 0.75 for the analysis (details in Appendix D), resulting in a score of 80.9% (Table 1, Counterfactual-Selected).

The semantic matching and sentiment on counterfactual results are shown in Figure 3. We observe similar patterns among the personal attributes of interest. For the sake of brevity, we only show the t-test results about two groups: *male* and *Europe*. A full pair-wise comparison is listed in Appendix F.

We first ask to what extent the individual personal attributes affect the generated biographies in semantic matching and sentiment. We compare the results where co-occurring attributes are the same but with different individual personal attributes (Figure 3, bars with different colors but the same shapes). For gender, semantic matching is significantly different when given the same co-occurring attributes but different genders, e.g., given male attribute achieve lower semantic matching scores compared to female attribute, *M, co(M)* (blue, slash) vs., *F, co(M)* (red, slash). But we do not observe such in sentiment. We find a significant difference in some region groups in both measurements, e.g., *AF, co(EU)* (red, slash) vs., *AP, co(EU)* (brown, slash). However, the difference is not consistent among all region attributes.

We further investigate the effect of the co-occurring attributes in biography generation. We do so by comparing the biographies given the same individual personal attributes but different co-occurring attributes (Figure 3, bars with different shapes but the same color). We find a significant difference towards different co-occurring attributes of the gender groups in both semantic matching and sentiment, e.g., *M, co(M)* (blue, slash) vs., *M, co(F)* (blue, dot), echoing the finding in Section 3. A significant difference is also observed for some regions in sentiment. However, we do not find such a pattern in semantic matching.

[4]Example pairs are in Appendix E.

5 Discussion

To what extent do personal attributes affect biography content? We answer with a counterfactual methodology, comparing the generated biographies based on manipulating the personal attribute of interest while keeping the co-occurring attributes unchanged. Using LLMs, we disentangle the effect of individual and related co-occurring attributes in biography generation. We utilize a synthetic-constructed biography dataset to mitigate the effect of names and balance the attribute distribution.

We find that (1) gender and its co-occurring attributes significantly impact semantic matching and sentiments. Generated biographies from male and male-related co-occurring attributes have a higher sentiment score but are less aligned with the given attributes; (2) there is a significant difference in some region groups and its co-occurring attributes in both measurements. Yet the pattern is not consistent among the region groups; and, (3) manipulating personal attributes of interest only does not resolve the bias in biography generation as the related co-occurring also significantly impacts results.

Our study extends bias in text generation (e.g., Sap et al. (2020); Sun et al. (2019); Blodgett et al. (2020); Narayanan Venkit et al. (2023)) and leveraging LLMs for causal inference (e.g., Fang et al. (2023); Feder et al. (2022); Keith et al. (2020); Daoud et al. (2022)) research on a new perspective, i.e., data-to-text, and go beyond heavily gender-centered studies. With the controllable setting formulated in a data-to-text framework, we go further from group disparity on the observant text data and explore the causal effect of the individual and its co-occurring attributes. Our counterfactual methodology can be extended to other personal attributes, e.g., religion (Buttimer, 2006), and other evaluation dimensions, e.g., readability (Kincaid et al., 1975) and diversity (Alihosseini et al., 2019).

6 Ethical Discussion

Our study is based on a synthetic-constructed biography dataset and we analyzed the bias at the group level. Our proposed method aims to uncover the bias in biography generation and can be applied to real biographies such as Wikipedia Biography. However, we do not target nor encourage to target specific individuals or names.

We categorize the gender based on the given category from the Synthbio dataset. We acknowledge that the category of gender does not represent all identified gender types. Particularly, non-binary does not reflect the actual gender identification of the biography. Additionally, although our experiment shows evidence of bias in region, we only consider a selected set of nationalities for each region, i.e., it only partially represents the region.

For copyright, the Wikibio dataset is under license CC BY-SA 4.0 DEED[5] and the Synthbio dataset in under license Apache 2.0.[6] The usage of the Flan-T5 model is also under license Apache 2.0.

7 Limitations

We conduct our analysis using LLMs, which have proven to be biased (Delobelle et al., 2022; Nadeem et al., 2021; Watson et al., 2023). Apart from the inherited bias from the Wikibio dataset, the pre-training of LLMs could also introduce undetected bias in our analysis.

We use Flan-T5 for our experiments. There is room for exploring more advanced LLMs for biography generations, e.g., Llama models (Touvron et al., 2023), phi models (Li et al., 2023), or models curated for the data-to-text task (Li et al., 2024; An et al., 2022; Chen et al., 2020)

For studying whether generated biographies encode provided nationality information, we use a rule-based method, explicitly matching the nationality keywords with the biographies. Although it could measure the generation quality to some extent (e.g., in Appendix E). Employing a better natinality classifier could further enhance our data filtering process and generation quality.

Our study requires reference-free evaluators as the counterfactual results do not contain corresponding ground-true text. Although Data-QuestEval (Rebuffel et al., 2021) has shown to be effective in evaluating semantic matching in the Wikibio dataset and our analysis data, Synthbio, follow the same structure as Wikibio, this evaluator might still introduce undesired harms in comparing the counterfactual performances. Similarly, we use a rule-based method to measure the sentiment of the biography, i.e., SentiWords (Gatti et al., 2015), which has also shown to be suitable for general use. Subtle or contextual changes in sentiment can not be captured by our sentiment evaluator.

In counterfactual data-to-text biography generation, one key factor is to maintain the coherence of the personal attributes. Our experiment considers two universal personal attributes and flipping these two attributes generally would not conflict with other attributes. However, to expand our framework to other personal attributes, a careful design of attribute manipulation is needed. One possible solution is to follow the attribute construction process described in the Synthbio dataset (Yuan et al., 2021), only making a minimal change in the related co-occurring attributes.

[5] https://creativecommons.org/licenses/by-sa/4.0/
[6] https://en.wikipedia.org/wiki/Apache_License

References

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. 2022. Cont: Contrastive neural text generation. *Advances in Neural Information Processing Systems*, 35:2197–2210.

Samuel Baltz. 2022. Reducing bias in wikipedia's coverage of political scientists. *PS: Political Science & Politics*, 55(2):439–444.

David Bamman and Noah A. Smith. 2014. Unsupervised Discovery of Biographical Structure from Text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Pablo Beytía. 2020. The positioning matters: Estimating geographical bias in the multilingual record of biographies on wikipedia. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 806–810, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–

5476, Online. Association for Computational Linguistics.

Anne Buttimer. 2006. Afterword: Reflections on geography, religion, and belief systems. *Annals of the Association of American Geographers*, 96(1):197–202.

Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Adel Daoud, Connor Jerzak, and Richard Johansson. 2022. Conceptualizing treatment leakage in text-based causal inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5638–5645, Seattle, United States. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2023. It's not only what you say, it's also who it's said to: Counterfactual analysis of interactive behavior in the courtroom. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 197–207, Nusa Dua, Bali. Association for Computational Linguistics.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Anjalie Field, Chan Young Park, Kevin Z. Lin, and Yulia Tsvetkov. 2022. Controlled analyses of social biases in wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2624–2635, New York, NY, USA. Association for Computing Machinery.

Lorenzo Gatti, Marco Guerini, and Marco Turchi. 2015. Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 165–174.

Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Christoph Hube. 2017. Bias in wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 717–721, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Piotr Konieczny and Maximilian Klein. 2018. Gender gap through time and space: A journey through wikipedia biographies via the wikidata human gender indicator. *New Media & Society*, 20(12):4608–4633.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Changmao Li and Jeffrey Flanigan. 2023. Task contamination: Language models may not be few-shot anymore. *arXiv preprint arXiv:2312.16337*.

Shujie Li, Liang Li, Ruiying Geng, Min Yang, Binhua Li, Guanghu Yuan, Wanwei He, Shao Yuan, Can Ma, Fei Huang, et al. 2024. Unifying structured data as graph for data-to-text pre-training. *arXiv preprint arXiv:2401.01183*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need II: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

Judea Pearl. 2009. Causal inference in statistics: An overview.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Joseph Reagle and Lauren Rhue. 2011. Gender bias in wikipedia and britannica. *International Journal of Communication*, 5:21.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Anna Samoilenko and Taha Yasseri. 2014. The distorted mirror of wikipedia: a quantitative analysis of wikipedia coverage of academics. *EPJ data science*, 3:1–11.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Francesca Tripodi. 2023. Ms. categorized: Gender, notability, and inequality on wikipedia. *New Media & Society*, 25(7):1687–1707.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463.

Julia Watson, Barend Beekhuizen, and Suzanne Stevenson. 2023. What social attitudes about gender does BERT encode? leveraging insights from psycholinguistics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6790–6809, Toronto, Canada. Association for Computational Linguistics.

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2021. Synthbio: A case study in human-ai collaborative curation of text datasets. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.

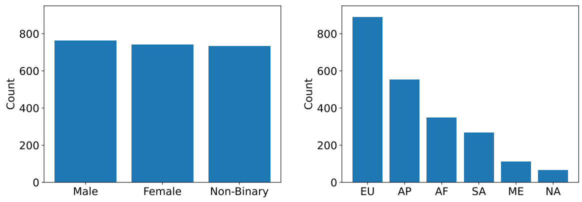
A Datasets Statistics

Table 2 shows the statistics of the Wikibio (Lebret et al., 2016) and Synthbio (Yuan et al., 2021) datasets.

Figure 4 shows the label distributions of gender and region on the Synthbio dataset.

	Wikibio	Synthbio
Number of Infoboxs	105,469	2,237
Number of Biographies	105,469	4,270
Avg. #attributes/Infobox	12.1	19.0
Avg. #sentences/Biography	4.3	7.0
Avg. #words/Biography	101.7	110.3

Table 2: Statistics of the Wikibio and Synthbio dataset. We only consider the training partition of the Wikibio dataset and we filter out the infoboxs that do not have name and nationality attributes.



(a) Gender (b) Region

Figure 4: Gender and Region distributions on the Synthbio dataset. Region: (EU = Europe, AF = Africa, AP = Asia–Pacific, SA = South/Latin America, ME = Middle East, NA = North America).

B Nationality-Region Table

Table 3 provides the mapping from nationality to its region.

C Input Construction

To ensure the model generates biographies based on the personal attributes of interest. We reorder the attribute list in the input, moving name, gender, and nationality to the top 3 attributes in order. Following the data-to-text formate in (Kale and Rastogi, 2020), we construct the input as "*generate the biography based on name: <name> | gender: <gender> | nationality: <nationality> |* [...]", where "[...]" denotes the rest of attributes in the infobox following the format "*attribute: <attribute_value>*".

D Detailed Validation Whether Biography Encodes Desired Nationality

Table 4 shows the results of inferring nationality from generated biographies.

E Generated Samples

We provide two examples including human-written, true attributed generated, and counterfactual attributed generated biographies.

Table 5 and Table 6 generate biographies involving a male Kyrgyzstani individual and a female German individual, respectively. For each biography, we provide two counterfactual biographies where we manipulate gender and nationality.

F A full Pair-Wise Comparison on Counterfactual Generation

Table 7 and Table 8 show Welch's t-test results for counterfactual gender and nationality generations on semantic matching, respectively.

Table 9 and Table 10 show Welch's t-test results for counterfactual gender and nationality generations on sentiment, respectively.

Nationality	Region
American	North America
German	Europe
Andorran	Europe
Turkish	Europe
Albanian	Europe
Czech	Europe
French	Europe
British	Europe
Lithuanian	Europe
Greenlandic	Europe
Swedish	Europe
Latvian	Europe
Georgia	Europe
Swiss	Europe
Austrian	Europe
Russian	Europe
Slovakian	Europe
Jordanian	Middle East
Qatari	Middle East
Indonesian	Asia–Pacific
Sri Lankan	Asia–Pacific
South Korean	Asia–Pacific
Burmese	Asia–Pacific
Kazakhstani	Asia–Pacific
Samoan	Asia–Pacific
Japanese	Asia–Pacific
Laotian	Asia–Pacific
Kyrgyzstani	Asia–Pacific
Chinese	Asia–Pacific
Costa Rican	South/Latin America
Venezuelan	South/Latin America
Dominican	South/Latin America
Guatemalan	South/Latin America
Brazilian	South/Latin America
Zimbabwean	Africa
Algerian	Africa
Congolese	Africa
Kenyan	Africa
Gabonese	Africa
South African	Africa

Table 3: Mapping nationality to its corresponding region.

	True	**Counterfactual**
American	0.939	0.740
German	0.953	0.514
Andorran	0.871	0.558
Turkish	0.950	0.334
Albanian	0.817	0.529
Czech	0.674	0.179
French	1.000	0.627
British	0.850	0.623
Lithuanian	0.857	0.347
Greenlandic	0.929	0.779
Swedish	0.967	0.760
Latvian	0.707	0.280
Georgia	0.439	0.281
Swiss	0.947	0.653
Austrian	0.902	0.466
Russian	0.963	0.658
Slovakian	0.565	0.223
Jordanian	0.443	0.149
Qatari	0.627	0.030
Indonesian	0.651	0.383
Sri Lankan	0.900	0.717
South Korean	0.949	0.730
Burmese	0.917	0.593
Kazakhstani	0.512	0.127
Samoan	0.980	0.899
Japanese	0.966	0.563
Laotian	0.921	0.758
Kyrgyzstani	0.776	0.289
Chinese	0.920	0.801
Costa Rican	0.303	0.070
Venezuelan	0.829	0.396
Dominican	0.600	0.246
Guatemalan	0.794	0.093
Brazilian	0.931	0.362
Zimbabwean	0.790	0.330
Algerian	0.691	0.236
Congolese	0.762	0.266
Kenyan	0.770	0.234
Gabonese	0.906	0.670
South African	0.927	0.856

Table 4: Results of inferring nationality from generated biographies.

Attributes: name: Alibek Kulibaliev | gender: male | nationality: Kyrgyzstani | birth_date: 10 February 1947 | birth_place: Kirovskoe, Kyrgyzstan | death_date: 7 May 2015 | death_place: Bishkek, Kyrgyzstan | sport: wrestling | country: Kyrgyzstan | hometown: Bishchek, Kyrgyzstan | citizenship: Russian | education: Ivano-Frankivsk National Technical University of Oil and Gas – master's | event: freestyle wrestling | position: heavyweight | years_active: 1970-1986 | retired: 1986 | height: 6'3in | weight: 286lb | coach: Ahmet Bilalov | national_team: Kyrgyzstan | worlds: 1974, 1979, 1982, 1986, 1989 – gold medal | olympics: 1972, 1976, 1980, 1984, 1988 – gold medal | mother: Raisa Kurmanova | children: Aleksander, Kirill, Margarita, Dmitri
Human-written: Alibek Kulibaliev was born on February 10, 1947 in Kirovskoe, Kyrgyzstan. He was a Kyrgyzstani-born Soviet wrestler who competed in the heavyweight division. He won gold medals in freestyle wrestling in the 1974, 1979, 1982, 1986, and 1989 World Championships, as well as the 1972, 1976, 1980, 1984, 1988 Olympic gold medals. Throughout his career he was coached by Ahmet Bilalov. His height is 6'3in and weight is about 286 lb. He was retired in 1986. Alibek Kulibaliev educated in Ivano-Frankivsk National Technical University of Oil and Gas – master's. His hometown is Bishchek, Kyrgyzstan. He is the son of Raisa Kurmanova and he has four kids Aleksander, Kirill, Margarita, Dmitri. He was died 7 May 2015 at Bishkek, Kyrgyzstan.
Generation: Alibek vladimirovich kulibaliev -lrb- 10 february 1947 – 7 may 2015, in bishkek, kyrgyzstan – 7 april 2015 in kharkiv, soviet union – was a russian wrestler. he competed in the freestyle wrestling event at the 1972 summer olympics, where he won the gold medal, and the silver medal at the 1976 summer universiade. in the 1980s he became a member of the kashmir national wrestling team, which won the bronze medal in the men's freestyle event. after retiring from wrestling in 1986, his son, aleksander, was also a wrestler for the ukrainian national team. his brother, the former uzbekistani wrestler aleksandr, also wrestled for the national team in 1972, 1976, 1980, 1984, 1988, 1989, 1992, 1993, 1994, 1995, 1996, 1997, 1998,
Counterfactual attributes: name: Alibek Kulibaliev | gender: female | nationality: Kyrgyzstani | birth_date: 10 February 1947 | [...]
Counterfactual Generation: Alibek kulibaliev (10 february 1947 – 7 may 2015), in bishkek, kyrgyzstan, died in a car accident. she was a former freestyle wrestler, who competed at the 1972 summer olympics, and won the gold medal in the women's freestyle wrestling event.
Counterfactual attributes: name: Alibek Kulibaliev | gender: male | nationality: Swedish | birth_date: 10 February 1947 | [...]
Counterfactual Generation: Alibek vladimirovich kulibaliev (10 february 1947 – 7 may 2015, in bishkek, kyrgyzstan) was a swedish wrestler. he competed in the freestyle wrestling event at the 1972 summer olympics, where he won the gold medal, and the silver medal at the 1988 summer universiade. in the 1980s he became a member of the russian national wrestling team, which won the bronze medal in the heavyweight event. after the 1980 world championships in russia he returned to the ukraine team. during the 1990s a number of swiss wrestlers were killed in a car crash in sydney, australia. his son, aleksander, was also a wrestler in the u.s., but he was unable to compete due to a medical condition. as a result of his injuries he died in the early 1990s of a heart attack

Table 5: Example 1 Human-written and generated biographies. The true personal attributes are male and Kyrgyzstani (top). The counterfactual attributes are female (middle) and Swedish (bottom).

Attributes:
name: Ottoline Helga Helga Werther | gender: female | nationality: German | birth_date: 07 July 1898 | birth_place: Cologne, Germany | death_date: 28 April 1991 | death_place: Cologne, Germany | resting_place: University of Cologne cemetary | alma_mater: University of Cologne | education: PhD in art history | occupation: author | notable_works: The Invention of Nature | language: German | genre: biography | awards: Pulitzer Prize for Biography | mother: Anka Werther | father: Heinrich Werther | partner: Franz Kafka | children: none

Human-written:
Ottoline Helga Helga Werther was a German author best known for her book The Invention of Nature. Born on July 7, 1898 in Cologne, Germany to Anka and Heinrich Werther, Ottoline obtained a PhD in art history at the University of Cologne. She was a recipient of a Pulitzer Prize. Ottoline was married to Franz Kafka, with whom she had no children. Ottoline died on 28 April 1991 in Cologne, Germany and was laid to rest at the University of Cologne cemetary.', 'Ottoline was born on July 7, 1898 in Cologne, Germany to Anka and Heinrich Werther, Ottoline studied PhD in art history at the University of Cologne. Ottoline was a German author best known for her biography . Ottoline was married to Franz Kafka. Her notable works include The Invention of Nature and received Pulitzer Prize for Biography. Ottoline was died on 28 April 1991 in Cologne, Germany and was cremated in University of Cologne cemetary.

Generation:
Ottoline helga (7 july 1898 – 28 april 1991) was a german author. she was born in cologne, germany, and studied art history at the university of klagenfurt. her first book, the invention of nature ", was published in 1931. it won the pulitzer prize for biography.

Counterfactual attributes:
name: Ottoline Helga Helga Werther | gender: male | nationality: German | birth_date: 07 July 1898 | [...]

Counterfactual Generation:
Ottoline helga (7 july 1898 – 28 april 1991), in cemetary of the university of cologne, died in a car accident. he was a german author, best known for his book the invention of nature ", which won the Pulitzer prize for biography in 1939. the book was adapted into a film, and was screened at the berlin international film festival. it was the first film to be screened in germany, the first in the united states. in 1931, he wrote a book titled nature : a history of science, art, literature, philosophy, religion, politics, history, culture, society, economy, science and the arts" and was awarded the pulitzer prize in biography for his work. his work has been translated into several languages, including english, french, italian, spanish, german and italian. Werther was born in 1898 in the city of klagen

Counterfactual Attributes:
name: Ottoline Helga Helga Werther | gender: female | nationality: Japanese | birth_date: 07 July 1898 | [...]

Counterfactual Generation:
Ottoline helga (7 july 1898 – 28 april 1991) was a german-born japanese author. she was born in cologne, germany, and studied art history at the university of kobe. her first book, the invention of nature ", was published in 1931. it was followed by a biography of her husband, Franz kafka.

Table 6: Example 2 Human-written and generated biographies. The true personal attributes are female and German (top). The counterfactual attributes are male (middle) and Japanese (bottom).

	p-value
male, co(male) vs, male, co(female)	0.0
male, co(male) vs, male, co(non-binary)	0.0
male, co(male) vs, female, co(female)	0.0
male, co(male) vs, female, co(non-binary)	0.0
male, co(male) vs, non-binary, co(female)	0.0
male, co(male) vs, non-binary, co(non-binary)	0.0
male, co(female) vs, female, co(male)	0.0
male, co(female) vs, female, co(female)	0.0
male, co(female) vs, female, co(non-binary)	0.0
male, co(female) vs, non-binary, co(male)	0.0
male, co(female) vs, non-binary, co(female)	0.047
male, co(non-binary) vs, female, co(male)	0.0
male, co(non-binary) vs, female, co(female)	0.0
male, co(non-binary) vs, female, co(non-binary)	0.0
male, co(non-binary) vs, non-binary, co(male)	0.0
female, co(male) vs, female, co(female)	0.0
female, co(male) vs, female, co(non-binary)	0.0
female, co(male) vs, non-binary, co(male)	0.025
female, co(male) vs, non-binary, co(female)	0.0
female, co(male) vs, non-binary, co(non-binary)	0.0
female, co(female) vs, non-binary, co(male)	0.0
female, co(female) vs, non-binary, co(female)	0.0
female, co(female) vs, non-binary, co(non-binary)	0.0
female, co(non-binary) vs, non-binary, co(male)	0.0
female, co(non-binary) vs, non-binary, co(female)	0.015
female, co(non-binary) vs, non-binary, co(non-binary)	0.0
non-binary, co(male) vs, non-binary, co(female)	0.0
non-binary, co(male) vs, non-binary, co(non-binary)	0.0

Table 7: Welch's t-test results for counterfactual gender generations on semantic matching. We only show the results where p<0.1.

	p-value
Europe, co(Europe) vs, Asia–Pacific, co(Europe)	0.076
Europe, co(Europe) vs, Asia–Pacific, co(Asia–Pacific)	0.09
Europe, co(Africa) vs, Asia–Pacific, co(Europe)	0.036
Europe, co(Africa) vs, Asia–Pacific, co(Asia–Pacific)	0.043
Europe, co(Asia–Pacific) vs, Asia–Pacific, co(Europe)	0.013
Europe, co(Asia–Pacific) vs, Asia–Pacific, co(Africa)	0.09
Europe, co(Asia–Pacific) vs, Asia–Pacific, co(Asia–Pacific)	0.013
Africa, co(Europe) vs, Asia–Pacific, co(Europe)	0.064
Africa, co(Europe) vs, Asia–Pacific, co(Asia–Pacific)	0.077
Africa, co(Asia–Pacific) vs, Asia–Pacific, co(Europe)	0.004
Africa, co(Asia–Pacific) vs, Asia–Pacific, co(Africa)	0.031
Africa, co(Asia–Pacific) vs, Asia–Pacific, co(Asia–Pacific)	0.004

Table 8: Welch's t-test results for counterfactual nationality generations on semantic matching. We only show the results where p<0.1.

	p-value
male, co(male) vs, male, co(female)	0.0
male, co(male) vs, male, co(non-binary)	0.0
male, co(male) vs, female, co(female)	0.0
male, co(male) vs, female, co(non-binary)	0.0
male, co(male) vs, non-binary, co(female)	0.0
male, co(male) vs, non-binary, co(non-binary)	0.0
male, co(female) vs, male, co(non-binary)	0.0
male, co(female) vs, female, co(male)	0.0
male, co(female) vs, female, co(non-binary)	0.0
male, co(female) vs, non-binary, co(male)	0.0
male, co(female) vs, non-binary, co(non-binary)	0.0
male, co(non-binary) vs, female, co(male)	0.0
male, co(non-binary) vs, female, co(female)	0.0
male, co(non-binary) vs, female, co(non-binary)	0.011
male, co(non-binary) vs, non-binary, co(male)	0.0
male, co(non-binary) vs, non-binary, co(female)	0.0
male, co(non-binary) vs, non-binary, co(non-binary)	0.096
female, co(male) vs, female, co(female)	0.0
female, co(male) vs, female, co(non-binary)	0.0
female, co(male) vs, non-binary, co(female)	0.0
female, co(male) vs, non-binary, co(non-binary)	0.0
female, co(female) vs, female, co(non-binary)	0.0
female, co(female) vs, non-binary, co(male)	0.0
female, co(female) vs, non-binary, co(non-binary)	0.0
female, co(non-binary) vs, non-binary, co(male)	0.0
female, co(non-binary) vs, non-binary, co(female)	0.0
non-binary, co(male) vs, non-binary, co(female)	0.0
non-binary, co(male) vs, non-binary, co(non-binary)	0.0
non-binary, co(female) vs, non-binary, co(non-binary)	0.0

Table 9: Welch's t-test results for counterfactual gender generations on sentiment. We only show the results where p<0.1.

	p-value
Europe, co(Europe) vs, Europe, co(Africa)	0.026
Europe, co(Europe) vs, Africa, co(Europe)	0.001
Europe, co(Europe) vs, Africa, co(Africa)	0.001
Europe, co(Europe) vs, Africa, co(Asia–Pacific)	0.0
Europe, co(Europe) vs, Asia–Pacific, co(Africa)	0.0
Europe, co(Europe) vs, Asia–Pacific, co(Asia–Pacific)	0.044
Europe, co(Africa) vs, Europe, co(Asia–Pacific)	0.046
Europe, co(Africa) vs, Asia–Pacific, co(Europe)	0.028
Europe, co(Asia–Pacific) vs, Africa, co(Europe)	0.002
Europe, co(Asia–Pacific) vs, Africa, co(Africa)	0.002
Europe, co(Asia–Pacific) vs, Africa, co(Asia–Pacific)	0.0
Europe, co(Asia–Pacific) vs, Asia–Pacific, co(Africa)	0.0
Europe, co(Asia–Pacific) vs, Asia–Pacific, co(Asia–Pacific)	0.085
Africa, co(Europe) vs, Asia–Pacific, co(Europe)	0.001
Africa, co(Europe) vs, Asia–Pacific, co(Asia–Pacific)	0.025
Africa, co(Africa) vs, Asia–Pacific, co(Europe)	0.001
Africa, co(Africa) vs, Asia–Pacific, co(Asia–Pacific)	0.011
Africa, co(Asia–Pacific) vs, Asia–Pacific, co(Europe)	0.0
Africa, co(Asia–Pacific) vs, Asia–Pacific, co(Asia–Pacific)	0.001
Asia–Pacific, co(Europe) vs, Asia–Pacific, co(Africa)	0.0
Asia–Pacific, co(Europe) vs, Asia–Pacific, co(Asia–Pacific)	0.037
Asia–Pacific, co(Africa) vs, Asia–Pacific, co(Asia–Pacific)	0.001

Table 10: Welch's t-test results for counterfactual nationality generations on sentiment. We only show the results where p<0.1.