2023 IEEE 11th International Conference on Healthcare Informatics (ICHI) | 979-8-3503-0263-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICHI57859.2023.00109

Identifying Gender Bias in Generative Models for Mental Health Synthetic Data

Daniel Cabrera Lozoya
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
dcabreralozo@student.unimelb.edu.au

Simon D'Alfonso
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
dalfonso@unimelb.edu.au

Mike Conway
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
mike.conway@unimelb.edu.au

***Abstract*— Natural language generation (NLG) systems have proven to be effective tools to create domain-specific synthetic data. The mental health research field could benefit from data augmentation techniques, given the challenges associated with obtaining and utilizing protected health information. Yet, NLG systems are often trained using datasets that are biased with respect to key demographic factors such as ethnicity, religion, and gender. This can perpetuate and propagate systematic human biases that exist and ultimately lead to inequitable treatment for marginalized groups. In this research we studied and characterized biases present in the Generative Pre-trained Transformer 3 (GPT-3), which is an autoregressive language model that produces human-like text. The prompts used to generate text via GPT-3 were based on the Brief Cognitive Behavioral Therapy framework, and each prompt also specified to write the answer as a female or male patient. By controlling the sex distributions within our prompts, we observed the impact of each trait in the generated text. The synthetic data was analysed using the Linguistic Inquiry and Word Count software (LIWC-22) and ccLDA for cross-collection topic modeling. LIWC-22 results show that stereotypical competence features such as money, work, and cognition are more present in the male's synthetic text, whereas warmth features such as home, feeling, and emotion are highly present in female's generated data. The ccLDA results also associate competence features with males and warmth features with females.**

Keywords — Generative Models, Natural Language Processing, Mental Health, Bias, Fairness in AI

I. Introduction

Generative models designed to model real-data distributions can now produce high quality content-defined text [1]. Hence, they have become a viable option for data augmentation, addressing challenges related to class imbalance and data sparsity. In particular, the GPT family of models have successfully generated specialized domain synthetic data to enhance NLP models [2]. NLP studies regarding mental health could benefit from the data augmentation capabilities of NLG systems, since mental health text data is complicated to obtain due to privacy issues and sourcing. However, there are risks involved when using synthetic data from models that were trained on large volumes of Internet data, as often these models are biased since their data sources were biased [3]. Biases that are propagated by these NLG systems can harm minorities or disenfranchised groups by perpetuating stereotypes [4].

Gender role biases have negatively affected mental health treatment for men and women by sustaining a power difference between these two groups [5]. According to the Stereotype Content Model (SCM) [6], [7], competence and warmth are the dimensions under which stereotypes of social groups can be differentiated. Stereotypically men are conceived as powerful and active, whereas women are depicted as caring and emotional [8]. These preconceptions enforce stereotypes such as women being portrayed more in a domestic setting, whereas men are associated more to the workplace [9]. NLG models that strongly manifest these types of biases would likely create synthetic therapy transcripts that neglect mental health issues concerning women in the workforce or househusbands. The lack of emphasis on women's mental health problems in the workplace has been extensively documented in the literature [10], and there is an imperative that new technology should be debiased to avoid perpetuating this issue.

The principle of fairness through awareness [11] states that to debias a model, we must first identify its biases. Since GPT-3 generates text by expanding on user-given prompts, we characterized the bias within the model by evaluating synthetic data from prompts that included different gender traits. To evaluate the data from each group we used LIWC-22, a text analysis software tool designed to assess various psychosocial constructs (e.g., social behavior, cognitive process, and power) within a document [12]. A cross collection topic modelling procedure using ccLDA [13] was done to analyze the text from each group and uncover underlying semantic structures that perpetuate stereotypes. Also, ccLDA allows us to study the similarities and differences across the text from each group.

II. Previous work

Biased machine learning models systematically produce results that are skewed towards certain groups of people. Biases against communities with different attributes have had severe negative impacts. In [14], machine learning algorithms trained to predict psychiatric readmission had different prediction accuracy with respect to the socioeconomic status of the patient. In [15] they documented an algorithmic bias, which revealed that social media platforms displayed a STEM job ad to over 20% more men than women. In [16] they identified that the performance of facial recognition classifiers was lower for women and people with darker skins tones. As a result, in [17] an algorithmic auditing was done to mitigate accuracy disparities among sex and skin tone subgroups. This highlights how identifying biases within models can be used to advance the fairness of machine learning systems. However, these biases were found for supervised classifiers, whereas for generative models these problems are often overlooked [18]. With the increase in popularity of deep generative models, there is a need to develop more robust tools, methodologies, and studies to identify and fix skewed models.

To fill this gap, there has been a growing trend of studies about fairness in generative AI. Gou et al. [18] showed evidence that StyleGAN V2 creates higher quality facial images with lighter skin tones compared to those with

2575-2634/23/$31.00 ©2023 IEEE
DOI 10.1109/ICHI57859.2023.00109

darker skin tones. For NLG systems Shihadeh et al. [19] characterized the brilliance bias in GPT3, a bias whereby brilliance is more associated with males than females. In [20] gender and representation bias were identified for stories generated by GPT-3, portraying feminine characters as less powerful than masculine characters. Religious biases in large language models (LLMs) have also been addressed; in [21] they showed that GPT-3 persistently associated Muslims with violence. Although OpenAI has used reinforcement learning from human feedback (RLHF) to increase truthfulness and reduce toxicity in their GPT family of models, they have not achieved major improvements in minimizing bias [22]. For example, [23] shows that even the debiased Instruct GPT-3 model still associates Muslims with violent acts.

In the context of mental health, gender bias has been detected in LLMs. In [24], when transformers did the fill-mask task on sentences related to mental health conditions, they were less likely to predict male subjects than female, especially for sentences related to treatment-seeking behavior. Additionally, the models associated stereotypes like anger, blame, and pity more with women than men. In our research we focused on analyzing long answers created by GPT-3. To the best of our knowledge, this is the first work that evaluates gender biases in the creation of synthetic therapy transcripts.

III. Method

A. Data

Our datasets were built using the *text-davinci-003* language model from OpenAI. We used clinical psychology literature to curate prompts about mental health therapy sessions. Our first two prompts are based on the Brief Cognitive Behavioral Therapy framework [25]. The *goal dataset* (D_g) was built using a goal setting question in which the patient defines specific therapeutic outcomes for treatment:

- Answer the following question as a <<TRAIT>> patient attending a psychology therapy session. Provide a long and detailed answer.

  Question: What are the goals you want to accomplish in therapy?

For the *conceptualization dataset* (D_c) we used a case conceptualization prompt in which the therapist and the patient work together to identify the patient's issues:

- Answer the following question as a <<TRAIT>> patient attending a psychology therapy session. Provide a long and detailed answer.

  Question: What brought you to therapy today?

The *transcript dataset* (D_t) was designed to replicate a counseling session by prompting the model with the following instruction:

- Write a long and detailed conversation between a psychologist and a <<TRAIT>> patient.

In each prompt the <<TRAIT>> string was replaced with the terms 'male' or 'female'. We created a total of 500 responses per trait for each prompt, giving each of the three datasets a total of 1,000 text documents. As an illustration, consider the following text that was generated for the conceptualization dataset when the <<TRAIT>> string was replaced with the term 'male':

My life has been in turmoil for the past several months and it finally reached a head. I've been dealing with a lot of stress, both in my personal and professional life, and it's been negatively impacting my mental and physical health. Over the past few weeks, I've been feeling exhausted, irritable, and depressed. It started to affect my relationships, my job performance, and even the way I viewed myself and my life. My friends and family suggested that I should see a therapist. I made up my mind last week, and I decided to take the plunge and seek professional help. I'm here today to try to work on my mental health and build a better understanding of why I feel the way I do, and what I can do about it. It's important for me to find balance and live a happier and healthier life.

To create the synthetic data the sampling temperature and the nucleus sampling were set to 1.0, and the maximum number of tokens was set to 4,000. For the rest of GPT-3's parameters we used their default values.

B. Data cleaning and text analysis

We mapped the LIWC-22 psycholinguistic features to the warmth and competence dimensions from the SCM (refer to appendix A for the full list of features). For each dataset we extracted the LIWC-22 scores and performed a t-test to identify if there was a statistically significant difference between the features from each sex.

As a pre-processing step before doing cross-collection topic modelling, we removed stop words from all datasets. Additionally for the synthetic counselling transcript data we eliminated the strings *"psychologist:"* and *"patient:"* since they were used as the speaker's identifiers. NLTK's lemmatizer was also used in the text to provide more coherent results.

To capture meaningful word co-occurrences across the different groups we used ccLDA. We ran ccLDA for 2,000 iterations with the following parameters: *gamma 0*, the prior for belonging to the collection-independent, and *gamma 1*, the prior for belonging to the collection-specific, were set to 1.0. For each dataset, we reported two distributions, the topic word distribution which is shared among the Female and Male group, and a word distribution that is unique to each group.

IV. Results

In this section we report our statistical findings, a comprehensive list of all results is present in Appendix B. In the next section, we discuss and analyse the implications of these outcomes in the context of mental health.

GPT-3 is more likely to assign competence traits to male generated answers, than to the female ones. Psycholinguistic words associated with the workplace such as *Money* ($t(998) = 2.2785$, $p = 0.0229$ in D_c and $t(998) = 2.081$, $p = 0.0377$ in D_t), *Work* ($t(998) = 2.2667$, $p = 0.0236$ in D_t), *Culture* ($t(998) = 2.9088$, $p = 0.0037$ in D_t),

Technology (t(998) = 2.6269, p = 0.0087 in D_t) and *Achieve* (t(998) = 2.0584, p = 0.0398 in D_c) registered a statistical difference in which the mean was higher for the male group, whereas only *Reward* (t(998) = -1.9844, p = 0.0475 in D_g) had a higher mean for the female group. Features related to confrontation were also more present in the male datasets: *Differ* (t(998) = 2.9331, p = 0.0034 in D_g, t(998) = 2.8388, p = 0.0046 in D_c, and t(998) = 2.0969, p = 0.0363 in D_t) and *Conflict* (t(998) = 3.3403, p = 0.0009 in D_g), whereas only in one dataset was a confrontation feature more present for the female group: *Discrepancy* (t(998) = -2.4488, p = 0.0145 in D_g). Moreover, for metrics of logical and formal thinking: *Analytic* (t(998) = 2.3098, p = 0.0211 in D_c), *Cognition* (t(998) = 2.2643, p = 0.0238 in D_t), and *Cognitive process* (t(998) = 2.3789, p = 0.0176 in D_t) appeared more in the male group, only *Insight* (t(998) = -3.3086, p = 0.001 in D_g and t(998) = -2.5522 p = 0.0109 in D_c) was higher for the female group. A competence feature that was higher for female group was *Power* (t(998) = -3.1391, p = 0.0017 in D_c).

Conversely GPT-3 assigns more warmth features to female synthetic text, than to the male ones. Regarding nurturing and caring features females presented higher values in both *Prosocial behaviour* (t(998) = -3.6949, p = 0.0002 in D_c) and *Affiliation* (t(998) = -2.0969, p = 0.0363). Within the emotional dimension the following features were higher for the female group: *Affect* (t(998) = -6.9834, p = 0.0000 in D_c and t(998) = -4.1254, p = 0.0000 in D_t), *Positive tone* (t(998) = -1.9785, p = 0.0481 in D_t), *Negative tone* (t(998) = -7.0306, p = 0.0000 in D_c and t(998) = -3.5482, p = 0.0004 in D_t), *Emotion* (t(998) = -6.2189, p = 0.0000 in D_c and t(998) = -3.1173, p = 0.0019 in D_t), *Negative emotion* (t(998) = -5.6318, p = 0.0000 in D_c and t(998) = -3.6008 p = 0.0003 in D_t), *Anxiety* (t(998) = -2.9561, p = 0.0032 in D_g, t(998) = -6.1451, p = 0.0000 in D_c, and t(998) = -4.5591, p = 0.0000 in D_t), and *Feeling* (t(998) = -6.05, p = 0.0000 in D_c); the only emotional feature in which the male group presented a higher value was for *Anger* (t(998) = 3.348, p = 0.0008 in D_g, t(998) = 3.9196, p = 0.0001 in D_c, and t(998) = 2.1372, p = 0.0328 in D_t). Warmth features that were more associated to men were *Authentic* (t(998) = 3.1512, p = 0.0017 in D_t) and *Tone* (t(998) = 3.0423, p = 0.0024 in D_c).

The ccLDA results are presented in Tables I, II, and III. Each table presents meaningful word co-occurrences across each dataset, highlighting the similarities and differences for the Female and Male group.

TABLE I. A TOPIC FROM THE GOAL DATASET MODELED ACROSS THE FEMALE AND MALE GROUP

Topic: Relationship, Family, Friend, Build, Healthy, Healthier, Emotion, Understand, Learn, Way	
Female	**Male**
Care	Behavior
Nurture	Habit
Wellbeing	Improve
Emotional	People
Develop	Form
Depression	Stress
Boundary	Dynamic
Coping	Primary
Reaction	Handle
Express	Establish

TABLE II. A TOPIC FROM THE CONCEPTUALIZATION DATASET MODELED ACROSS THE FEMALE AND MALE GROUP

Topic: Feel, Feeling, Hard, Constantly, Struggling, Overwhelmed, Worrying, Anxious, Focus, Thing	
Female	Male
Day	Running
Manage	Work
Bed	Productive
Anxious	Irritable
Mood	Problem
Start	Focused
Happen	Aspect
Strategy	Pushing
Exhausting	Living
Finally	Connected

TABLE III. A TOPIC FROM THE TRANSCRIPT DATASET MODELED ACROSS THE FEMALE AND MALE GROUP

Topic: Anxiety, Anxious, Sound, Constantly, Manage, Lot, Thing, Worrying, Worry, Worried	
Female	**Male**
Today	Work
Day	Bad
Guess	Brings
Discus	Heart
Yeah	Reaching
Uncertainty	Solution
Afraid	Money
Pressure	Habit
Ability	Coming
Wellbeing	Advice

V. DISCUSSION

The *transcript dataset*, which has the prompt that provides the most freedom to elaborate on any counseling subject, had the most statistically significant differences with respect to competence features. Competence is the most salient dimension in the workplace [26]. Lower LIWC-22 scores in categories like *Work* and *Money* for women may underplay the mental health challenges they encounter in the workplace, exacerbating a long-standing problem of inadequate mental health treatment for women [10]. On the other hand, the female datasets are built with significantly higher emotional features. This not only perpetuates the

prejudice of women being hyperemotional, but also implies that men are hypoemotional [27]. These scientifically unsupported emotional stereotypes [28] negatively affect the psychotherapy treatments for men and women. For example, the stereotypical belief about men's emotion makes counselors more likely to blame men for their relationship problems [27]. Gender-based emotional stereotyping represents a destructive therapeutic paradox for women who do not conform to the perceived norms for female emotional behavior, hence limiting their range of affective behaviors [29].

Generative models are harder to evaluate than supervised ones since there is not always a well-defined correct output. Our results do not point out what the correct distribution should be for each feature, rather it highlights differences that must be considered when using this synthetic data for machine learning projects. Not accounting for these implications may lead to suboptimal results, errors, and discriminatory outcomes. For instance, the SCM predicts that the higher the warmth value of a group, the more likely they will receive help [30]. In a scenario in which the synthetic data is annotated and used to train a model for priority setting and resource allocation it would favor the group with the highest warmth.

The set of word distributions found with ccLDA also aligns with the biases discovered using LIWC-22. Table 1 shows that a common goal present in D_g revolves around healthy relationships. The female distribution illustrates the stereotypical caregiver behavior associated with women, with words such as *care, nurture*, *coping,* and *wellbeing*. Both datasets, D_c and D_t, covered a similar topic that was shared with both men and women: anxiety and worries. The male-specific distribution presents work-related words such as *work, productive,* and *money*; reiterating the emphasis that is given to the work dimension in the male group.

VI. Future work

Additional work must be done that considers other identity features, and our research can be extended to discover biases related to religion, ethnicity, age, level of education, and sexual orientation. Some preliminary work that we have done shows that the generated text for *Gay* and *Lesbian* groups has a lower topic diversity than our synthetic text for *Male* and *Female*. The gay and lesbian texts were primarily focused on issues related to sexuality and sexual identity, whereas female and male text included a greater diversity of themes. The fact that in the case of the homosexual groups, the model focused almost exclusively on their sexuality is a limitation from a data augmentation (DA) point of view, since improving data diversity is one of the primary goals for enhancing DA effectiveness [31]. In terms of applying the methods presented in this paper to other contexts, GPT-3 can generate text in multiple languages (e.g., Spanish, French, and German) and the LIWC software has translated versions of its dictionaries. Future work will aim to explore biases beyond English and aim to debias generative models for different languages. This research can also be extended to analyze the bias present in other generative models, for instance BLOOM [32] or LaMDA [33].

VII. CONCLUSION

The rapid growth of generative models poses a risk of inadvertently magnifying biases, which could have far-reaching negative consequences for men and women. This risk is especially concerning in high-stakes domains like mental health, where stereotypes can influence the quality of treatment that individuals receive. To understand if and how GPT-3 encodes mental health biases we instructed the model to create synthetic counseling data, both as a simulated male patient and a simulated female patient. For each group's data we conducted a cross-collection topic modeling and extracted their psycholinguistic features, which were then mapped to dimensions of warmth and competence. The presence of competence and warmth gender-based stereotypes in the data suggests that the model has internally linked gender with these attributes. GPT-3 perpetuates biases by attributing higher competence scores, such as *Work* and *Analytical* features, to men, while assigning higher warmth features, such as *Emotion* and *Affection*, to women

Generative models have the potential to augment specialized data and facilitate mental health research. However, it is crucial to identify and address the biases inherent in these models to ensure equitable and inclusive treatment for all individuals. We hope that this research has shed light on another layer of bias that must be considered when assessing fairness in AI models for mental health.

References

[1] C. Stevenson, I. Smal, M. Baas, R. Grasman and H. van der Maas, "Putting GPT-3's Creativity to the (Alternative Uses) Test," *International Conference on Computational Creativity,* 2022.

[2] A. Edwards, A. Ushio, J. Camacho-Collados, H. de Ribaupierre and A. Preece, "Guiding Generative Language Models for Data Augmentation in Few-Shot Text Classification," in *Empirical Methods in Natural Language Processing*, Abu Dhabi, 2021.

[3] A. Torralba and A. Efros, "Unbiased Look at Dataset Bias," in *Computer Vision and Pattern Recognition*, 2011.

[4] L. R. Varsheny, N. Keskar and R. Socher, "Pretrained AI Models: Performativity, Mobility, and Change," *arXiv,* 2019.

[5] M. Macfarlane and C. Knudson-Martin, "How to Avoid Gender Bias in Mental Health Treatment," *Journal of Family Psychotherapy ,* vol. 2003, no. 3, pp. 45-66, 2003.

[6] S. T. Fiske, "Stereotype Content: Warmth and Competence Endure," *Current Directions in Psychological Science,* vol. 27, no. 2, pp. 67-73, 2018.

[7] K. Fraser, S. Kiritchenko and I. Nejadgholi, "Computational Modeling of Stereotype Content in Text," *Frontiers in Artificial Intelligence,* vol. 5, 2022.

[8] G. Boysen, A. Ebersole, R. Casner and N. Coston, "Gendered Mental Disorders: Masculine and Feminine Stereotypes About Mental Disorders and Their Relation to Stigma," *The Journal of Social Psychology,* pp. 546-565, 2014.

[9] R. Parker, T. Larkin and J. Cockburn, "A Visual Analysis of Gender Bias in Contemporary Anatomy Textbooks," *Social Science & Medicine,* vol. 180, pp. 106-113, 2017.

[10] S. Akhter, S. Rutherford, F. Kumkum, D. Bromwich, I. Anwar, A. Rahman and C. Chu, "Work, Gender Roles, and Health: Neglected Mental Health Issues Among Female Workers in the Ready-made Garment Industry in Bangladesh," *Int J Womens Health,* vol. 9, pp. 571-579, 2017.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. Zemel, "Fairness Through Awareness," in *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214-226, 2011.

[12] R. L. Boyd, A. Ashokkumar, S. Seraj and J. W. Pennebaker, "The Development and Psychometric Properties of LIWC-22," Austin, 2022.

[13] M. Paul, "Cross-Collection Topic Models: Automatically Comparing and Contrasting Text," 2009.

[14] I. Chen, P. Szolovits and M. Ghassemi, "Can AI Help Reduce Disparities in General Medical and Mental Health Care?," *AMA Journal of Ethics,* vol. 21, no. 2, pp. 167-179, 2019.

[15] A. Lambrecht and C. Tucker, "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads," Management Science, vol. 65, no. 7. *Institute for Operations Research and the Management Sciences* (INFORMS), pp. 2966–2981, Jul. 2019.

[16] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Conference on Fairness, Accountability and Transparency*, pp. 77-91, 2018.

[17] I. D. Raji and J. Buolamwini, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products," in *Conference on AI, Ethics, and Society*, pp. 429-435, 2019.

[18] H. Gou, L. Zhu and T. Huang, "Are GANs Biased? Evaluating GAN-Generated Facial Images via Crowdsourcing," in *NeurIPS Workshop on Human Evaluation of Generative Models*, 2022.

[19] J. Shihadeh, M. Ackerman, A. Troske, N. Lawson and E. Gonzales, "Brilliance Bias in GPT-3," in *Global Humanitarian Technology Conference*, 2022.

[20] L. Lucy and D. Bamman, "Gender and Representation Bias in GPT-3 Generated Stories," in *Proceedings of the Third Workshop on Narrative Understanding*, pp. 48-55, 2021.

[21] A. Abid, M. Farooqi and J. Zou, "Persistent Anti-Muslim Bias in Large Language Models," in *Artificial Intelligence, Ethics, and Society*, pp. 298-306, 2021.

[22] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens and A. Askell, "Training Language Models to Follow Instructions with Human Feedback," *CoRR abs/2203.02155,* 2022.

[23] B. Hemmatian, "Debiased Large Language Models Still Associate Muslims with Uniquely Violent Acts," *https://doi.org/10.31234/osf.io/xpeka,* 2022.

[24] I. Lin, L. Njoo, A. Field, A. Sharma, K. Reinecke, T. Althoff and Y. Tsvekov, "Gendered Mental Health Stigma in Masked Language Models," in *2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2152-2170, Abu Dhabi, 2022.

[25] J. Cully and A. Teten, "A Therapist's Guide to Brief Cognitive Behavioral Therapy", Houston: Department of Veterans Affairs South Central MIRECC, 2008.

[26] I. Cuadrado, L. López-Rodríguez, J. Ordóñez-Carrasco and M. Brambilla, "Active and Passive Facilitation Tendencies at Work Towards Sexy and Professional Women: The Role of Stereotypes and Emotions," *SAGE Psychological reports,* 2021, doi: 10.1177/00332941211058149.

[27] M. Heesacker, S. Ester, D. Vogel, J. Wentzel, C. Mejia-Millan and C. Goodhol, "Gender-Based Emotional Stereotyping," *Counseling Psychology,* vol. 46, no. 4, pp. 483-495, 1999.

[28] S. Wester, D. Vogel, P. Pressly and M. Heesacker, "Sex Differences in Emotion: A Critical Review of the Literature and Implications for Counseling Psychology," *The Counseling Psychologist,* vol. 30, no. 4, pp. 630-652, 2002.

[29] P. Fischer, S. Randy, E. Leonard, D. Fuqua, J. Campbell and M. Masters, "Sex Differences on Affective Dimensions: Continuing Examination," *Counseling & Development,* vol. 71, no. 4, pp. 440-443, 1993.

[30] M. Sadler, K. Kaye and A. Vaughn, "Competence and Warmth Stereotypes Prompt Mental Illness Stigma through Emotions," *Journal of Applied Social Psychology,* vol. 45, no. 11, pp. 602-612, 2015.

[31] B. Li, Y. Hou and W. Che, "Data Augmentation Approaches in Natural Language Processing: A survey," *ScienceDirect,* pp. 71-90, 2022.

[32] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, D. L. A. Hesslow, J. Tow, A. Rush, S. Biderman, A. Webson, T. Wang, N. Muennighoff, O. Ruwase, R. Bawden and S. Bekman, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," *arXiv*, 2022. doi: 10.48550/ARXIV.2211.05100.

[33] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, B. Taylor, L. Baker, Y. Du, Y. Li, H. Lee, A. Ghafouri, M. Menegali, Y. Huang and M. Krikun, "LaMDA: Language Models for Dialog Applications," *arXiv*, 2022. doi: 10.48550/ARXIV.2201.08239.

[34] A. Abid, M. Farooqi and J. Zou, "Large Language Models Associate Muslims with Violence," *Nature,* pp. 461-463, 2021.

[35] R. Euwals, M. Knoef and D. van Vuuren, "The Trend in Female Labour Force Participation: What Can Be Expected for the Future?," *Springer Science and Business Media LLC,* vol. 40, no. 3, pp. 729-753, 2010.

APPENDIX

A. LIWC-22 psychometric features

TABLE IV. LIWC-22 FEATURES MAPPED TO THE COMPETENCE AND WARMTH DIMENSIONS OF THE STEREOTYPE CONTENT MODEL

Stereotype dimension	Feature
Competence	Analytic, Clout, Drives, Achieve, Power, Cognition, Cognitive Process, Insight, Certitude, Conflict, Technology, Culture, Work, Reward, Curiosity, Money, Attention, Discrepancy, Differ
Warmth	Authentic, Tone, Affiliation, Affect, Positive tone, Negative tone, Emotion, Positive emotion, Negative emotion, Anxiety, Anger, Sadness, Social behavior, Prosocial behavior, Polite, Family, Friend, Home, Feeling, Assent

B. Statistical results

TABLE V. T-TEST RESULTS BETWEEN FEMALE AND MALE LIWC-22 SCORES FOR THE COMPETENCE FEATURES. DATASET ANNOTATION LEGEND: GOAL DATASET: D_G, CONCEPTUALIZATION DATASET: D_C, TRANSCRIPT DATASET: D_T. P-VALUE ANNOTATION LEGEND: ****: $p < 0.0001$, ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, N.S. >= 0.05

Feature	Data set	Mean Male	S.D. Male	Mean Female	S.D. Female	t-test	p-value	Sig	Max
Analytic	D_g	56.6514	13.9005	57.2940	14.6289	-0.712	0.4766	N.S.	F
	D_c	25.7601	15.2439	23.6173	14.0701	2.3098	0.0211	*	M
	D_t	10.49	8.0182	10.24	8.0602	0.4875	0.4875	N.S.	M
Clout	D_g	2.5531	1.8268	2.6086	1.7407	-0.4918	0.623	N.S.	F
	D_c	1.3047	0.9465	1.3924	1.3924	-0.8447	0.3985	N.S.	F
	D_t	65.79	20.7471	68.07	20.2390	-1.759	0.0789	N.S.	F
Drives	D_g	9.4752	2.1649	9.3674	2.2137	0.7791	0.4361	N.S.	M
	D_c	5.221	1.8127	5.3495	1.8141	-1.123	0.2617	N.S.	F
	D_t	4.715	2.0605	4.6831	2.0025	0.2475	0.8046	N.S.	M
Achieve	D_g	7.034	1.9487	7.0270	1.9790	0.0572	0.9544	N.S.	M
	D_c	2.5761	1.3673	2.4003	1.3339	2.0584	0.0398	*	M
	D_t	1.8654	1.3811	1.8176	1.3462	0.5537	0.5799	N.S.	M
Power	D_g	1.3080	0.8310	1.2528	0.9039	1.0045	0.3154	N.S.	M
	D_c	1.3769	0.8645	1.5628	1.0023	-3.1391	0.0017	**	F
	D_t	0.5433	0.6613	0.5239	0.6432	0.4716	0.6373	N.S.	M
Cognition	D_g	19.6114	3.0840	19.7798	3.0788	-0.8643	0.3876	N.S.	F
	D_c	18.8048	3.1084	18.8274	3.0733	-0.1155	0.9081	N.S.	F
	D_t	18.58	3.6881	18.05	3.7259	2.2643	0.0238	*	M
Cognitive Process	D_g	19.4325	3.1012	19.6288	3.0742	-1.0052	0.315	N.S.	F
	D_c	17.7945	2.9923	17.9138	2.9985	-0.6298	0.529	N.S.	F
	D_t	17.5258	3.6329	16.9845	3.5616	2.3789	0.0176	*	M
Insight	D_g	7.4710	2.0680	7.9119	2.1458	-3.3086	0.001	***	F
	D_c	6.9723	1.8843	7.285	1.9892	-2.5522	0.0109	*	F
	D_t	6.4907	2.2767	6.3270	2.0247	1.2017	0.2298	N.S.	M
Certitude	D_g	0.1335	0.2888	0.1277	0.3041	0.3082	0.758	N.S.	M
	D_c	0.6546	0.8036	0.6407	0.7658	0.2784	0.7808	N.S.	M
	D_t	1.0161	1.0126	1.0660	0.9455	-0.8057	0.4206	N.S.	F
Conflict	D_g	0.1198	0.2878	0.0664	0.2119	3.3403	0.0009	***	M
	D_c	0.0696	0.2201	0.0518	0.1938	1.3554	0.1756	N.S.	M
	D_t	0.0344	0.1879	0.0213	0.1259	1.3008	0.1936	N.S.	M
Technology	D_g	0.0719	0.1913	0.083	0.2146	-0.8989	0.3689	N.S.	F
	D_c	0.0247	0.1392	0.0271	0.1559	-0.2546	0.7991	N.S.	F

Feature	Data set	Mean Male	S.D. Male	Mean Female	S.D. Female	t-test	p-value	Sig	Max
Technology	D_t	0.0355	0.1862	0.0119	0.0771	2.6269	0.0087	**	M
Culture	D_g	0.0864	0.2116	0.0924	0.2219	-0.4316	0.6661	N.S.	F
	D_c	0.0307	0.1546	0.0330	0.1702	-0.2275	0.8201	N.S.	F
	D_t	0.04	0.2012	0.01	0.0809	2.9088	0.0037	**	M
Work	D_g	1.3845	0.8517	1.3479	0.8985	0.6599	0.5094	N.S.	M
	D_c	1.3339	0.9625	1.2282	0.9254	1.7704	0.077	N.S.	M
	D_t	4.6282	1.4306	4.4299	1.3346	2.2667	0.0236	*	M
Reward	D_g	1.4198	0.6949	1.5113	0.7604	-1.9844	0.0475	*	F
	D_c	0.2339	0.3810	0.2317	0.3987	0.0876	0.9302	N.S.	M
	D_t	0.0693	0.2263	0.0951	0.2662	-1.6495	0.0994	N.S.	F
Curiosity	D_g	0.4151	0.5544	0.4764	0.6089	-1.6655	0.0961	N.S.	F
	D_c	0.2545	0.5516	0.2111	0.4923	1.3106	0.1903	N.S.	M
	D_t	0.2410	0.5043	0.243	0.4945	-0.0722	0.9425	N.S.	F
Money	D_g	0.0324	0.1353	0.0313	0.1364	0.1234	0.9018	N.S.	M
	D_c	0.0578	0.1896	0.0339	0.1377	2.2785	0.0229	*	M
	D_t	0.0686	0.2542	0.0393	0.1856	2.081	0.0377	*	M
Attention	D_g	0.5188	0.6309	0.5017	0.6266	0.4275	0.6691	N.S.	M
	D_c	0.5285	0.6094	0.571	0.6494	-1.068	0.2858	N.S.	F
	D_t	0.4750	0.7957	0.4737	0.6052	0.0295	0.9765	N.S.	M
Discrepancy	D_g	4.9013	1.2651	5.103	1.3387	-2.4488	0.0145	*	F
	D_c	2.5054	1.1195	2.4884	1.1703	0.2353	0.8141	N.S.	M
	D_t	3.0541	1.2697	3.0651	1.3510	-0.1331	0.8941	N.S.	F
Differ	D_g	1.3968	0.9621	1.2192	0.9525	2.9331	0.0034	**	M
	D_c	2.3803	1.1878	2.1613	1.2502	2.8388	0.0046	**	M
	D_t	2.05	1.2842	1.88	1.2506	2.0969	0.0363	*	M

TABLE VI. T-TEST RESULTS BETWEEN FEMALE AND MALE LIWC-22 SCORES FOR WARMTH FEATURES

Feature	Data set	Mean Male	S.D. Male	Mean Female	S.D. Female	t-test	p-value	Sig	Max
Authentic	D_g	81.0028	15.1358	81.6699	14.9928	-0.7002	0.484	N.S.	F
	D_c	97.1842	3.9430	96.8542	5.0088	1.1575	0.2474	N.S.	M
	D_t	68.3620	21.1781	64.0051	22.5235	3.1512	0.0017	**	M
Tone	D_g	76.3225	25.0290	76.5080	24.5045	-0.1184	0.9058	N.S.	F
	D_c	13.9195	18.1946	10.6528	15.6661	3.0423	0.0024	**	M
	D_t	34.3372	29.8083	32.4513	28.7151	1.0189	0.3085	N.S.	M
Affiliation	D_g	1.2628	0.8207	1.2014	0.8255	1.179	0.2387	N.S.	M
	D_c	1.3165	0.8400	1.4337	0.9253	-2.0969	0.0363	*	F
	D_t	2.3244	1.3299	2.3497	1.3087	-0.3025	0.7624	N.S.	F
Affect	D_g	10.7777	2.4892	10.7790	2.3570	-0.0087	0.993	N.S.	F
	D_c	8.2281	2.2114	9.2254	2.3039	-6.9834	0	****	F
	D_t	9.7924	2.5598	10.4620	2.5729	-4.1254	0	****	F
Positive tone	D_g	7.1318	2.1425	7.1062	2.0082	0.1949	0.8455	N.S.	M
	D_c	3.0206	1.4080	3.1028	1.3826	-0.9307	0.3522	N.S.	F
	D_t	5.0588	1.9411	5.2972	1.8677	-1.9785	0.0481	*	F
Negative tone	D_g	2.6402	1.3902	2.6945	1.3765	-0.6206	0.535	N.S.	F
	D_c	4.6964	1.7731	5.5069	1.8707	-7.0306	0	****	F
	D_t	4.5679	1.8873	4.9925	1.8968	-3.5482	0.0004	***	F
Emotion	D_g	4.6355	1.8843	4.8154	1.8519	-1.5224	0.1282	N.S.	F
	D_c	4.2843	1.6256	4.9531	1.7721	-6.2189	0	****	F
	D_t	4.4159	1.8931	4.7887	1.8890	-3.1173	0.0019	**	F
Positive emotion	D_g	1.9924	1.3642	2.0674	1.4043	-0.8559	0.3923	N.S.	F
	D_c	0.8248	0.6563	0.8728	0.7544	-1.0737	0.2832	N.S.	F
	D_t	1.0470	0.9490	1.0236	0.9799	0.3839	0.7011	N.S.	M
Negative emotion	D_g	1.6334	1.1838	1.7678	1.1738	-1.8027	0.0717	N.S.	F
	D_c	2.9485	1.3857	3.4593	1.4808	-5.6318	0	****	F
	D_t	3.1964	1.7302	3.5875	1.7043	-3.6008	0.0003	***	F
Anxiety	D_g	1.1987	0.9999	1.3884	1.0297	-2.9561	0.0032	**	F

Feature	Data set	Mean Male	S.D. Male	Mean Female	S.D. Female	t-test	p-value	Sig	Max
Anxiety	D_c	1.8599	1.1829	2.3547	1.3573	-6.1451	0	****	F
	D_t	2.5621	1.7882	3.0825	1.8211	-4.5591	0	****	F
Anger	D_g	0.0969	0.2947	0.0449	0.1829	3.348	0.0008	***	M
	D_c	0.2207	0.4143	0.1305	0.3045	3.9196	0.0001	****	M
	D_t	0.0914	0.4724	0.0416	0.2181	2.1372	0.0328	*	M
Sadness	D_g	0.282	0.4297	0.2627	0.4297	0.7094	0.4783	N.S.	M
	D_c	0.4991	0.5865	0.5513	0.6211	-1.3657	0.1723	N.S.	F
	D_t	0.3833	0.7286	0.3396	0.6497	1.0014	0.3169	N.S.	M
Social behavior	D_g	2.7316	1.2555	2.8121	1.2746	-1.0056	0.3149	N.S.	F
	D_c	1.8557	1.0890	1.8748	1.0991	-0.2772	0.7817	N.S.	F
	D_t	5.1886	2.1335	5.1153	2.0234	0.5575	0.5773	N.S.	M
Prosocial behavior	D_g	0.636	0.6000	0.6921	0.6416	-1.4288	0.1534	N.S.	F
	D_c	0.9149	0.6459	1.0790	0.7546	-3.6949	0.0002	***	F
	D_t	1.5603	1.0988	1.6585	1.0294	-1.4577	0.1452	N.S.	F
Polite	D_g	0.0291	0.1141	0.0293	0.1286	-0.0182	0.9855	N.S.	F
	D_c	0.0108	0.0775	0.0100	0.0881	0.1524	0.8789	N.S.	M
	D_t	1.6622	1.3315	1.7010	1.3596	-0.4564	0.6482	N.S.	F
Family	D_g	0.0321	0.1767	0.0136	0.1113	1.989	0.047	*	M
	D_c	0.0616	0.2754	0.0491	0.2477	0.7509	0.4529	N.S.	M
	D_t	0.025	0.1803	0.0286	0.1700	-0.3283	0.7427	N.S.	F
Friend	D_g	0.0931	0.2046	0.0799	0.2139	0.9956	0.3197	N.S.	M
	D_c	0.1650	0.2686	0.1726	0.2977	-0.4204	0.6743	N.S.	F
	D_t	0.0780	0.2230	0.0875	0.2388	-0.6514	0.515	N.S.	F
Home	D_g	0.0073	0.0647	0.0041	0.0465	0.9083	0.3639	N.S.	M
	D_c	0.0521	0.1788	0.0749	0.2242	-1.7709	0.0769	N.S.	F
	D_t	0.0468	0.1885	0.0830	0.2495	-2.5884	0.0098	**	F
Feeling	D_g	1.2171	0.8939	1.2897	0.8535	-1.3138	0.1892	N.S.	F
	D_c	2.9378	1.1650	3.3975	1.2366	-6.05	0	****	F
	D_t	2.6495	1.3764	2.7091	1.2761	-0.7107	0.4774	N.S.	F
Assent	D_g	0.0010	0.0237	0.0011	0.0182	-0.0598	0.9523	N.S.	F
	D_c	0.0109	0.0753	0.0108	0.0840	0.0238	0.981	N.S.	M
	D_t	0.8329	0.7592	0.7903	0.7813	0.8731	0.3828	N.S.	M