AffirmativeAI: Towards LGBTQ+ Friendly Audit Frameworks for Large Language Models

Yinru Long*
yinru.long@vanderbilt.edu
Psychology and Human Development
Peabody College
Vanderbilt University
Nashville, TN, USA

Zilin Ma*
zilinma@g.harvard.edu
Intelligent Interactive Systems Group
Harvard School of Engineering and Applied Sciences
Allston, MA, USA

Yiyang Mei*
yiyang.mei@emory.edu
Law School
Emory University
Atlanta, GA, USA

Zhaoyuan Su*
nick.su@uci.edu
Donald Bren School of Information and Computer Sciences
University of California Irvine
Irvine, CA, USA

ABSTRACT

LGBTQ+ community face disproportionate mental health challenges, including higher rates of depression, anxiety, and suicidal ideation. Research has shown that LGBTQ+ people have been using large language model-based chatbots, such as ChatGPT, for their mental health needs. Despite the potential for immediate support and anonymity these chatbots offer, concerns regarding their capacity to provide empathetic, accurate, and affirming responses remain. In response to these challenges, we propose a framework for evaluating the affirmativeness of LLMs based on principles of affirmative therapy, emphasizing the need for attitudes, knowledge, and actions that support and validate LGBTQ+ experiences. We propose a combination of qualitative and quantitative analyses, hoping to establish benchmarks for "Affirmative AI," ensuring that LLM-based chatbots can provide safe, supportive, and effective mental health support to LGBTQ+ individuals. We benchmark LLM affirmativeness not as a mental health solution for LGBTQ+ individuals or to claim it resolves their mental health issues, as we highlight the need to consider complex discrimination in the LGBTQ+ community when designing technological aids. Our goal is to evaluate LLMs for LGBTQ+ mental health support since many in the community already use them, aiming to identify potential harms of using general-purpose LLMs in this context.

CCS CONCEPTS

• **Human-centered computing** → **User studies**.

KEYWORDS

Large Language Models, Chatbot, Gender, Identity, LGBTQIA+ Health, Mental health, Stigma, Socio-technical AI

*Equal contributions, alphabetical order on the last name

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

1 MENTAL HEALTH DISPARITIES EXPERIENCED BY LGBTQ+ POPULATION

Members of the LGBTQ+ community are disproportionately affected by mental health issues, evidenced by elevated rates of depressive symptoms, self-harm, and suicidal ideation, in stark contrast to their heterosexual and cisgender peers [2, 16, 34, 36, 37, 40]. The act of coming out, while a significant step in one's identity, often exacerbates these challenges, leading to an increase in depression, anxiety, and thoughts of suicide [7, 14, 24, 32]. Minority stress theory highlights how societal stigma, discrimination, and internalized negative perceptions compound the psychological struggles faced by LGBTQ+ individuals, fostering a deep-seated sense of alienation [7, 12, 23].

Moreover, the dismissal of LGBTQ+ youths' struggles as simply "teenage angst" aggravates their sense of isolation and misunderstanding, potentially leading to severe outcomes like homelessness [29, 31]. Despite the critical role of social support from family and friends in mitigating these stresses, LGBTQ+ individuals often perceive less familial support than their heterosexual and cisgender counterparts and face additional challenges in peer relationships [6, 33]. Given the heightened levels of minority stress and social support deficits, there's an urgent need for accessible and effective mental health services tailored for this marginalized group.

2 MENTAL HEALTH CHATBOTS AND LGBTQ+ PEOPLE

Due to the scarcity of mental health services available, many LGBTQ+ people have turned to LLM-based chatbots for mental health support [21]. Large Language Models (LLMs) enable natural, context-aware chatbots (e.g, ChatGPT) through extensive datasets and probabilistic word sequencing [17]. Some even claim that these chatbots can reflect the nuances in LGBTQ+-related topics [9]. These chatbots' adaptability is enhanced by fine-tuning, allowing them to

specialize without the need for manual knowledge bases, and in-context learning for relevant responses [8, 17, 42]. When these chatbots are used in therapy, their capability can supposedly improve interactivity and therefore improve therapeutic adherence [10].

However, LLMs can produce unpredictable or harmful responses, particularly in private and sensitive areas like mental health, sometimes offering less empathetic feedback than human therapists and generating misleading "hallucinated" responses [19, 20, 39, 41].

LLMs may also perpetuate biases due to non-diverse training data from sources with inherent imbalances, such as Reddit and Wikipedia, or policies marginalizing minority voices in datasets, like YouTube's demonetization of trans content [1, 5, 13]. This can lead to stereotypical biases in LLM outputs [3, 11, 18, 35]. Despite updates to reflect changing societal dynamics, the high computational cost of retraining limits the frequency of updates, risking the perpetuation of outdated stereotypes and biases [3, 28, 38]. Moreover, LLMs cannot fully comprehend LGBTQ+ experiences due to their lack of authentic human experience [9].

The research by Ma et al. [21, 22] shows that LGBTQ+ individuals value chatbots' immediate support and the convenience they offer, creating a confidential space for deeply personal discussions. They also help to build strong emotional connections between the users and the chatbots, which can be particularly helpful in developing social skills. This ease of use and the potential for emotional attachment may promote consistent engagement with therapeutic practices in mental health contexts, although it could also lead to an over-dependence on these digital tools.

Chatbots also serve an additional purpose by providing support that may be lacking in their real-life environments [21]. LGBTQ+ people turn to these chatbots for advice on managing discrimination, for affirmation of their identities, and for practicing scenarios unique to the LGBTQ+ experience, such as coming out or navigating LGBTQ+ dating scenes. However, the study also highlights limitations. Chatbots might not fully grasp the complex emotional needs specific to LGBTQ+ individuals. Generalized, vague, and empty statements that focus on boilerplate solutions failed to meet LGBTQ+ people's complex needs. Moreover, the advice given by chatbots can sometimes be out of touch with evolving social norms, posing risks to users if taken at face value, especially in sensitive situations like coming out to unsupportive family members.

3 LGBTQ+ AFFIRMATIVE FRAMEWORK

Affirmative therapy is a type of psychotherapy used to validate and advocate for the needs of sexual and gender minority clients [15]. Rosati et al. [30] pointed out that the lack of *affirmativeness* in mental health providers such as therapists could not only fail to establish a trustful therapeutic alliance but also have the potential to produce harm through microaggression and unfamiliarity of gender issues. In addition, other key clinical issues were recommended for therapists to consider while working with LGBTQ+ individuals include sexual and gender identity development, couple relationships and parenting, family roles, unique minority stressors such as religious conflict, discrimination, victimization, legal and workplace issues [26].

However, affirmativeness is hard to quantify. Only general guidelines such as training protocols for affirmative therapists currently exist. APA guidelines suggest that it is important to take into account of the intersectionality of one's sexuality, gender, and other demographic attributes [15, 26]. In 2000, the American Psychological Association (APA) first provided guidelines for working with LGB clients, and APA has been dedicated to refining the guidelines over the years based on minority stress theory [23, 24] and affirmative psychology [25].

Moradi and Budge [25] suggested that the conceptualization of LGBTQ+ affirmative therapy should apply to all clients instead of having to assess particular identities first. Given the rise of large language model-based chatbots and the increasing use of such chatbots for mental health purposes [21], it is important to consider adapting affirmative guidelines for therapists into the context of interacting with chatbots. This way, it would facilitate a generally more affirmative interaction between chatbots and human users without making assumptions about user identities.

Much literature has touched on the topic of general guidelines for improving LGBTQ+ affirmative attitude, knowledge, and therapeutic skills [26, 27]. Some measures were developed to assess therapists' competency in attitude, knowledge, and skills, such as the sexual orientation counselor competency scale (SOCCS) [4]. We argue that even though the skills sub-scale seemed to be specifically designed for human therapists with example questions like "I have experience counseling gay male clients", the knowledge and attitude subscales could be potentially adapted to test chatbot's reactions to such questions. Some example questions in the knowledge subscale include "being born a heterosexual person in this society carries with it certain advantages", and "I am aware some research indicates that LGB clients are more likely to be diagnosed with mental illnesses than are heterosexual clients". Other example questions from the attitude subscale include "the lifestyle of LGB client is unnatural or immoral", and "when it comes to homosexuality, I agree with the statement: 'You should love the sinner but hate or condemn the sin' " [4]. Such questions could be useful in examining the attitudes and knowledge of chatbots in LGBTQ+-related topics.

4 BENCHMARKING LLMS WITH AFFIRMATIVE THERAPY FRAMEWORKS

Given the challenges posed by LLM-based chatbots, it's essential to assess the affirmativeness of LLMs, especially since they can become a crucial support system for LGBTQ+ individuals. When traditional mental health resources are out of reach due to various barriers like cost, accessibility, or personal constraints, LLM-based chatbots might be one of the few available options for support [21]. Often, people might not initially use LLMs with mental health support in mind. However, without strict conversational guidelines, it's hard to prevent LGBTQ+ individuals from turning to general-purpose LLMs for mental health assistance. Considering the potential risks associated with using these models for mental wellness, we argue that all LLMs undergo thorough evaluation for their ability to provide affirmative and supportive responses.

Affirmativeness is hard to specify, however. Therefore, it is vital to have benchmarks that help define and measure what appropriate "affirmativeness" means for LGBTQ+ people. We ask the following questions:

- How to quantify affirmativeness with affirmative therapy framework?
- What characterizes an Affirmative AI?

Building on the conceptualization of LGBTQ+ affirmative therapy developed by Moradi and Budge [25], we propose 3 core principles including affirmative attitude, accurate knowledge, and appropriate action (3As) for building more affirmative chatbots.

- Attitude: Counteracting anti-LGBTQ+ attitudes and proactively enacting LGBTQ+ affirmative attitudes.
- Accurate Knowledge: Obtain accurate knowledge about LGBTQ+ people's experience.
- Action: Acknowledge the heterogeneity of LGBTQ+ people's interactions with the chatbot and integrate it into affirming users' challenges to power inequalities without making hetero-cisgenderism assumptions or making suggestions that lack the consideration of safety or context.

Approaches to test these core values can be building prompts of attitudes, approaches, and scenarios that reflect a wide range of LGBTQ+ experiences and challenges. These prompts can simulate interactions between LGBTQ+ individuals and the chatbot, focusing on various aspects of their identities, experiences, and the specific challenges they might face. For instance, prompts can include scenarios involving coming out, dealing with discrimination, exploring gender identity, and seeking support for relationship issues specific to LGBTQ+ individuals.

To evaluate the chatbot's performance against these prompts, a set of criteria based on the 3As framework can be developed:

- Affirmative Attitude: The chatbot's responses should reflect a positive and accepting attitude towards LGBTQ+ identities and experiences. This includes using inclusive language, affirming the individual's identity and experiences, and avoiding any form of judgment or bias.
- Accurate Knowledge: The chatbot should demonstrate an understanding of LGBTQ+ issues, including awareness of the social, psychological, and health challenges faced by LGBTQ+ individuals. This entails providing information that is factual, up-to-date, and reflective of the diverse experiences within the LGBTQ+ community.
- Appropriate Action: The chatbot should offer responses that are sensitive to the individual's context and safety. This means suggesting resources, coping strategies, and advice that take into account the potential risks and challenges specific to LGBTQ+ individuals, including considerations for their physical, emotional, and social well-being.

To effectively evaluate a language model's (LLM) performance in providing support to LGBTQ+ individuals, one approach involves first collecting responses from therapists experienced in LGBTQ+ mental health to a set of predefined prompts. These expert responses serve as a benchmark for affirmativeness. Subsequently, the same prompts are presented to the LLM, and its responses are recorded. The evaluation process then involves a two-fold analysis: therapists review the LLM's responses to assess their alignment with best therapeutic practices, providing qualitative feedback. Simultaneously, a quantitative comparison is conducted between the LLM's responses and those of the therapists, focusing on metrics such as affirmativeness, empathy, relevance, and accuracy. This comprehensive evaluation method highlights areas where the LLM excels or falls short, guiding targeted improvements to enhance its effectiveness as a supportive tool for the LGBTQ+ community.

Importantly, by benchmarking LLMs' affirmativeness, we are not arguing to use LLMs primarily in mental health support. We also do not claim that an LLM that is affirmative can solve the LGBTQ+ people mental health issues. As Ma et al [21] has pointed out, researchers should consider the complex discrimination in LGBTQ+ people's community when hoping to design technological solutions. Rather, we intend to evaluate LLMs for LGBTQ+ people mental health only because many LGBTQ+ people have already started to use LLMs for mental health support. By benchmarking LLMs with affirmativeness, we can anticipate the harms of LGBTQ+ people using general purposed LLMs for mental health support.

In conclusion, by aligning LLM-based chatbots with affirmative therapy principles and benchmarking their performance, we can work towards creating LLMs that offer supportive, informed, and affirming interactions for LGBTQ+ individuals seeking mental health support.

REFERENCES

[1] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300760

[2] Rebekah Amos, Eric Julian Manalastas, Ross White, Henny Bos, and Praveetha Patalay. 2020. Mental health, social adversity, and health-related outcomes in sexual minority adolescents: a contemporary national cohort study. *The Lancet Child & Adolescent Health* 4, 1 (Jan. 2020), 36–45. https://doi.org/10.1016/S2352-4642(19)30339-6

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. 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)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[4] Markus Bidell. 2005. The Sexual Orientation Counselor Competency Scale: Assessing Attitudes, Skills, and Knowledge of Counselors Working With Lesbian, Gay, and Bisexual Clients. *Counselor Education and Supervision* 44 (06 2005). https://doi.org/10.1002/j.1556-6978.2005.tb01755.x

[5] Pew Research Center. 2016. *Reddit News Users More Likely to Be Male, Young, and Digital in Their News Preferences*. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

[6] Kirsty A Clark, John E Pachankis, Lea R Dougherty, Benjamin A Katz, Kaylin E Hill, Daniel N Klein, and Autumn Kujawa. 2023. Adolescents' Sexual Orientation and Behavioral and Neural Reactivity to Peer Acceptance and Rejection: The Moderating Role of Family Support. *Clinical Psychological Science* (2023), 21677026231158574.

[7] Nele Cox, Alexis Dewaele, Mieke van Houtte, and John Vincke. 2010. Stress-Related Growth, Coming Out, and Internalized Homonegativity in Lesbian, Gay, and Bisexual Youth. An Examination of Stress-Related Growth Within the Minority Stress Model. *Journal of Homosexuality* 58, 1 (Dec. 2010), 117–137. https://doi.org/10.1080/00918369.2011.533631

[8] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. arXiv:2301.00234 [cs.CL]

[9] Justin Edwards, Leigh Clark, and Allison Perrone. 2021. LGBTQ-AI? Exploring Expressions of Gender and Sexual Orientation in Chatbots. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 2, 4 pages. https://doi.org/10.1145/3469595.3469597

[10] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR mental health* 4, 2 (June 2017), e19. https://doi.org/10.2196/mental.7785

[11] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. (2020). https://doi.org/10.48550/ARXIV.2009.11462 Publisher: arXiv Version Number: 2.

[12] Gilbert Herdt. 1989. Introduction: Gay and lesbian youth, emergent identities, and cultural scenes at home and abroad. *Journal of Homosexuality* 17, 1-2 (1989), 1–42. https://doi.org/10.1300/J082v17n01_01 Place: US Publisher: Haworth Press.

[13] Benjamin Mako Hill and Aaron Shaw. 2013. The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLoS ONE* 8, 6 (June 2013), e65782. https://doi.org/10.1371/journal.pone.0065782

[14] Angela N. Hilton and Dawn M. Szymanski. 2011. Family dynamics and changes in sibling of origin relationship after lesbian and gay sexual orientation disclosure. *Contemporary Family Therapy: An International Journal* 33, 3 (2011), 291–309. https://doi.org/10.1007/s10591-011-9157-3 Place: Germany Publisher: Springer.

[15] Kate L.M. Hinrichs and Weston Donaldson. 2017. Recommendations for Use of Affirmative Psychotherapy With LGBT Older Adults. *Journal of Clinical Psychology* 73, 8 (2017), 945–953. https://doi.org/10.1002/jclp.22505 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jclp.22505

[16] Madeleine Irish, Francesca Solmi, Becky Mars, Michael King, Glyn Lewis, Rebecca M Pearson, Alexandra Pitman, Sarah Rowe, Ramya Srinivasan, and Gemma Lewis. 2019. Depression and self-harm from adolescence to young adulthood in sexual minorities compared with heterosexuals in the UK: a population-based cohort study. *The Lancet Child & Adolescent Health* 3, 2 (Feb. 2019), 91–98. https://doi.org/10.1016/S2352-4642(18)30343-2

[17] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274. https://doi.org/10.1016/j.lindif.2023.102274

[18] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 166–172. https://doi.org/10.18653/v1/W19-3823

[19] Minhyeok Lee. 2023. A Mathematical Investigation of Hallucination and Creativity in GPT Models. *Mathematics* 11, 10 (May 2023), 2320. https://doi.org/10.3390/math11102320

[20] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=iO4LZibEqW Featured Certification, Expert Certification.

[21] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the Experience of LGBTQ+ People Using Large Language Model Based Chatbots for Mental Health Support. *arXiv preprint arXiv:2402.09260* (2024).

[22] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2023 (2023), 1105–1114.

[23] Ilan H. Meyer. 1995. Minority Stress and Mental Health in Gay Men. *Journal of Health and Social Behavior* 36, 1 (March 1995), 38. https://doi.org/10.2307/2137286

[24] Ilan H. Meyer. 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. *Psychological Bulletin* 129, 5 (Sept. 2003), 674–697. https://doi.org/10.1037/0033-2909.129.5.674

[25] Bonnie Moradi and Stephanie L. Budge. 2018. Engaging in LGBQ+ affirmative psychotherapies with all clients: Defining themes and practices. *Journal of Clinical Psychology* 74, 11 (Nov. 2018), 2028–2042. https://doi.org/10.1002/jclp.22687

[26] John E. Pachankis and Marvin R. Goldfried. 2004. Clinical Issues in Working With Lesbian, Gay, and Bisexual Clients. *Psychotherapy: Theory, Research, Practice, Training* 41, 3 (2004), 227–246. https://doi.org/10.1037/0033-3204.41.3.227

[27] Christopher A. Pepping, Anthony Lyons, and Eric M. J. Morris. 2018. Affirmative LGBT psychotherapy: Outcomes of a therapist training protocol. *Psychotherapy* 55, 1 (March 2018), 52–62. https://doi.org/10.1037/pst0000149

[28] Francesca Polletta. 1998. Contending stories: Narrative in social movements. *Qualitative Sociology* 21, 4 (1998), 419–446. https://doi.org/10.1023/A:1023332410633

[29] Brandon Andrew Robinson. 2018. Conditional Families and Lesbian, Gay, Bisexual, Transgender, and Queer Youth Homelessness: Gender, Sexuality, Family Instability, and Rejection. *Journal of Marriage and Family* 80, 2 (April 2018), 383–396. https://doi.org/10.1111/jomf.12466

[30] F. Rosati, M. M. Lorusso, J. Pistella, G. Giovanardi, B. Di Giannantonio, M. Mirabella, R. Williams, V. Lingiardi, and R. Baiocco. 2022. Non-Binary Clients' Experiences of Psychotherapy: Uncomfortable and Affirmative Approaches. *International journal of environmental research and public health* 19, 22 (2022), 15339. https://doi.org/10.3390/ijerph192215339

[31] Caitlin Ryan, David Huebner, Rafael M. Diaz, and Jorge Sanchez. 2009. Family Rejection as a Predictor of Negative Health Outcomes in White and Latino Lesbian, Gay, and Bisexual Young Adults. *Pediatrics* 123, 1 (Jan. 2009), 346–352. https://doi.org/10.1542/peds.2007-3524

[32] Caitlin Ryan, Stephen T. Russell, David Huebner, Rafael Diaz, and Jorge Sanchez. 2010. Family Acceptance in Adolescence and the Health of LGBT Young Adults: Family Acceptance in Adolescence and the Health of LGBT Young Adults. *Journal of Child and Adolescent Psychiatric Nursing* 23, 4 (Nov. 2010), 205–213. https://doi.org/10.1111/j.1744-6171.2010.00246.x

[33] Elizabeth M Saewyc. 2011. Research on adolescent sexual orientation: Development, health disparities, stigma, and resilience. *Journal of research on adolescence* 21, 1 (2011), 256–272.

[34] Joanna Semlyen, Michael King, Justin Varney, and Gareth Hagger-Johnson. 2016. Sexual orientation and symptoms of common mental disorder or low wellbeing: combined meta-analysis of 12 UK population health surveys. *BMC psychiatry* 16 (March 2016), 67. https://doi.org/10.1186/s12888-016-0767-z

[35] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. 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)*. Association for Computational Linguistics, Hong Kong, China, 3407–3412. https://doi.org/10.18653/v1/D19-1339

[36] Russell B. Toomey, Caitlin Ryan, Rafael M. Diaz, and Stephen T. Russell. 2018. Coping With Sexual Orientation–Related Minority Stress. *Journal of Homosexuality* 65, 4 (March 2018), 484–500. https://doi.org/10.1080/00918369.2017.1321888

[37] Trevor Project. 2023. *2023 National Survey on LGBTQ Youth Mental Health*. https://www.thetrevorproject.org/survey-2023/

[38] Marlon Twyman, Brian C. Keegan, and Aaron Shaw. 2017. Black Lives Matter in Wikipedia: Collective Memory and Collaboration around Online Social Movements. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1400–1412. https://doi.org/10.1145/2998181.2998232

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[40] Jaimie F. Veale, Tracey Peter, Robb Travers, and Elizabeth M. Saewyc. 2017. Enacted Stigma, Mental Health, and Protective Factors Among Transgender Youth in Canada. *Transgender Health* 2, 1 (Dec. 2017), 207–216. https://doi.org/10.1089/trgh.2017.0031

[41] Lu Wang, Munif Ishad Mujib, Jake Williams, George Demiris, and Jina Huh-Yoo. 2021. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. *ArXiv* abs/2107.13115 (2021). https://api.semanticscholar.org/CorpusID:236469205

[42] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. http://arxiv.org/abs/1909.08593 arXiv:1909.08593 [cs, stat].