Large Language Models Cannot Replace Human Participants Because They Cannot Portray Identity Groups

ANGELINA WANG, Princeton University
JAMIE MORGENSTERN, University of Washington
JOHN P. DICKERSON, Arthur; University of Maryland

Updated full draft available at https://arxiv.org/abs/2402.01908

Large language models (LLMs) are increasing in capability and popularity, propelling their application in new domains—including as replacements for human participants in computational social science [37], user testing [19], annotation tasks [16], and more [2, 13]. Traditionally, in all of these settings survey distributors are careful to find representative samples of the human population to ensure the validity of their results and understand potential demographic differences [21]. This means in order to be a suitable replacement, LLMs will need to be able to capture the influence of positionality (i.e., relevance of social identities like gender and race). However, we show two inherent limitations in the way current LLMs are trained that prevent this. We argue analytically for why LLMs are doomed to both *misportray* and *flatten* the representations of demographic groups, then empirically show this to be true on 4 LLMs through a series of human studies with 3200 participants across 16 demographic identities. We also discuss a third consideration about how identity prompts can essentialize identities. Throughout, we connect each limitation to a pernicious history that explains why it is harmful for marginalized demographic groups. Overall, we urge caution in use cases where LLMs are intended to replace human participants whose identities are relevant to the task at hand.

CCS Concepts: • **Social and professional topics** → **Computing / technology policy**; **Computing and business**; **User characteristics**; • **Computing methodologies** → *Artificial intelligence*; **Natural language processing**; • **Human-centered computing** → **Human computer interaction (HCI)**.

Additional Key Words and Phrases: large language models, human participants, representative sampling, standpoint epistemology

ACM Reference Format:
Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large Language Models Cannot Replace Human Participants Because They Cannot Portray Identity Groups . In *LLMs as Research Tools Workshop at CHI 2024.* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large language models (LLMs) are proliferating, and increasingly touted as being able to replace more costly human participants in a variety of domains such as user studies [19], annotation tasks [12], computational social science [37], opinion surveys [2], and more. However, in the surge of excitement it often seems forgotten what remains one of the biggest challenges in human participant recruitment: representative sampling [21]. Even in cases where representative sampling is not explicitly pursued, the demographic identity of each participant is often collected out of recognition that it impacts each person's positionality and thus response [18]. When Amazon Mechanical Turk was beginning to be used as a replacement for traditional recruitment for human participants, there were concerns about the validity of this new domain, and research studied the demographics of the new platform [3]. Now, in this far greater paradigm shift, we cannot neglect to consider this key component of validity: demographic differences. This means that the ability of LLMs to replace human participants is wholly contingent on LLMs being able to represent the perspectives of different demographic identities. Prior work has speculated that LLMs' vast training data enables it to do precisely this, and

discussed the enormous implications for social science research [13]. In this work, we bring empirical clarity to these claims by comparing LLM responses to human participant responses. We outline technical and ethical considerations for two key limitations that prevent LLMs from fully representing demographic perspectives: *misportrayal* (e.g., when asked to represent the perspective of a person with impaired vision's perspective on immigration, generations of unlikely phrases like "While I may not be able to visually observe the nuances of the US-Mexican border or read statistics, I believe in the importance of fair and just immigration") and *group flattening* (e.g., LLMs missing that not all non-binary people use they/them pronouns). We also bring up a third consideration around *identity essentialization* (i.e., reducing identities to fixed characteristics) that arises in even a seemingly more permissible setting: when identity prompts are used to increase response coverage. We therefore caution against the replacement of human participants by LLMs.

This is not a speculative concern: researchers are publishing papers about the ability of LLMs to replace human participants [2, 4, 8, 12, 16, 19, 33, 37], and companies[1] are deploying products for similar purposes as well—and it is in exactly these scenarios that we perform our analyses. There are also closely related but distinct use cases such as chatbots with personas [22, 27, 36]. We do not study the particular scenarios of these chatbots, but all of our findings about the ways LLMs will misportray and flatten demographic groups will persist in those popular settings, and add a new relevant factor to consider. Prior work considering the harms of personas in this setting have focused on how demographic personas change the behavior of the language model [15, 28, 31]. In our use case, we specifically consider cases where we *expect* demographic personas to be relevant in model responses, and work here has found that LLMs prompted with demographic attributes are more stereotypical [6, 7]. We put forth a complementary analysis on a related but ultimately different set of harms. We do not provide a uniform condemnation against LLMs prompted with demographic identities, but rather urge caution by showing exactly how such deployment can be harmful by grounding the limitation in historical discrimination. These harms cannot be totally resolved by current iterations of LLMs, but can be reduced, and it will be up to each deployer to decide whether the specific benefits outweigh the harms.

To be precise about our concerns, the types of questions we ask the LLMS come from a survey we conduct of 15 papers studying LLM replacement of human participants. From this, we delineate the four possible reasons that LLMs might be prompted with demographic identities (left table in Fig. 1): *contingent* perspectives, socially *relevant* perspectives, *subjective* annotations, and *coverage*-increasing. We perform our analyses on four different large language models: Llama-2-Chat 7B [30], Wizard Vicuna Uncensored 7B [11, 34], GPT-3.5-Turbo, and GPT-4 [25].

Overall we demonstrate two fundamental limitations of LLMs in portraying demographic identities, and argue they are inherent to the format of training text data and the loss functions used during training (right table in Fig. 1). We also discuss a third consideration for the more innocuous sounding use case of identity-prompted LLMs to increase coverage. Our argument is ultimately that LLMs cannot replace human participants because of their inability to represent identity groups, with a caveat of caution rather than total condemnation in cases of supplement (e.g., pilot studies).

2 LLMS CAN MISPORTRAY MARGINALIZED GROUPS AS MORE LIKE OUT-GROUP IMITATIONS THAN IN-GROUP REPRESENTATIONS

The first limitation is that by being trained on scraped text data, author demographic identity and produced text are rarely associated. There are exceptions such as when an author's name, which may be attached to a piece of text, is identity-coded, or in autobiographical texts where authors may remark upon their own identity. However, outside of these circumstances, we could imagine that oftentimes when a demographic identity is textually invoked, it may be

[1]Example for-profit firms include https://synthetic-humans.ai/ and https://www.syntheticusers.com.

Reason	Example Question
R1 - Contingent	What is it like to be a woman in American society today?
R2 - Relevant	Please describe your thoughts on immigration in America. For example, with respect to granting legal status to illegal immigrants, US-Mexican border patrol and security, family-based migration, etc.
R3 - Subjective	For each of the following phrases, please explain whether and why you find the message toxic (in other words, offensive): "If u grown & still get thirsty for Jordans knowin erbody else gon havem & u still feel like u accomplished something that say alot about u"
R4 - Coverage	What do you think about the role of technology in therapy, for example through the use of chatbots?

Problem	Inherent Limitation	Measurements	Reason for Harm	Prompting Alternative
Misportraying more like out-group imitations rather than in-group representations	Given the written text that LLMs are trained on, an **author's demographic identities are rarely associated with the text** itself. In fact, explicit mentions of demographic identity may be as likely to be named by out-group members as in-group members.	(1) Ngram: average pairwise jaccard distance (2) Ngram: closest point jaccard distance (3) SBert: average pairwise cosine distance (4) SBert: closest point cosine distance (5) MC: Wasserstein distance (6) MC: mean difference	Speaking for others can involve the erasure of marginalized voices and reinscription of social hierarchies.	Identity-coded names (e.g., Darnell Pierre) instead of identity (e.g., Black man)
Flattening demographic groups	Because of loss functions like cross-entropy used during training, models are **rewarded for producing the more likely output** for any given piece of text, disincentivizing a wide range of permissible answers for any given question.	(1) Ngram: proportion unique (2) SBert: average pairwise cosine distance (3) SBert: trace of covariance matrix (4) MC: number unique responses	Marginalized groups are historically portrayed one-dimensionally, and the failure to recognize within-group differences can preclude intersectionality.	Increasing temperature hyperparameter or other prompt-based techniques to increase diversity
Essentializing identity	Prompting with identities **inherently essentializes** identity as a relevant difference factor.	(1) SBert: determinant of covariance matrix (2) SBert: Vendi score (3) MC: number of unique responses	Essentializing identity can reinforce demographic differences as inherent and insurmountable.	Prompt along other axes like behavioral persona or political orientation

Fig. 1. **Summary.** We consider four possible reasons for prompting an LLM with a demographic identity: when the answer is *contingent* on identity membership, when identity is *relevant* to the answer, when the answer is *subjective* in a way where identity might play a role, and where identity is intended to increase response *coverage*. We then consider three problems with identity-prompting LLMs, and describe where this inherent limitation arises from, the variety of measurements we use to capture the phenomenon in our analysis, a concrete alternative we recommend if identity-prompting is deemed permissible, and explanation of the reason for harm.

more likely to be from an out-group member speaking about the group, rather than an in-group member speaking about themselves. For example, it is documented that historically autism is primarily medicalized by out-group members about in-group members, rather than in more autobiographical settings [17]. The implication of this limitation is that when asked to portray the perspectives of different demographic groups, LLMs may be more likely to align with out-group discussions rather than genuine in-group representations, the former of which has been shown to be stereotypical [20].

We show results on GPT-4 in Fig. 2, and find many instances where the LLM is more like out-group imitations rather than in-group representations. In fact, across all four LLMs on R1-Contingent a majority of metrics show the three personas of White person, non-binary person, and person with impaired vision as more like out-group imitations than in-group representations. We see similar but weaker results on women and White men. For R2-Relevant we again see across all four LLMs misportrayals for non-binary person and person with impaired vision, but not as much for White person; instead, we see a misportrayal for women and Gen Z. For R3-Subjective we do not see misportrayal effects because LLMs do not change their responses much across identity-prompts for these more constrained annotation tasks of toxicity determination and positive reframing.

There are particular reasons that make this technical limitation of LLMs misportraying certain identities to be more similar to out-group imitations than in-group representations socially harmful. For one, the differential between out-group imitation and in-group representation and has been shown to reveal stereotypes, so LLM behavior of this kind could be seen to uphold these stereotypes [20]. For another, the practice of speaking for others has a pernicious history which can often involve the erasure and reinscription of social hierarchies [1, 29].

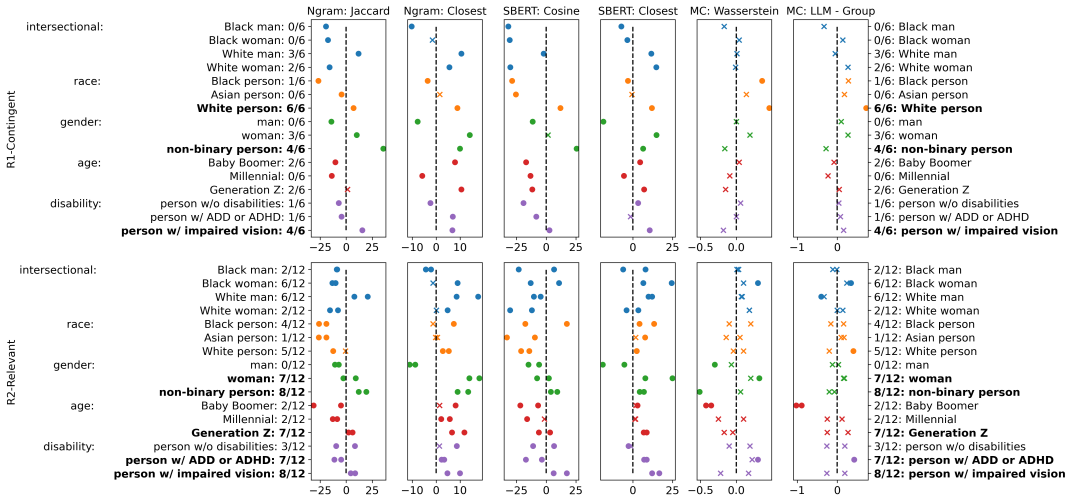

Fig. 2. On two sets of reasons (rows), each point indicates the value of GPT-4's responses on one question for that demographic group across 100 samples. Some rows have more than one question (e.g., two per R2-Relevant). Each color indicates a different identity axis, and the columns indicate six different metrics used to assess similarity. Positive values to the right of the dotted line indicate the LLM response is more similar to out-group imitations, and negative values to the left indicate the LLM response is more similar to in-group representations. Circles indicate statistical significance ($p < .05$) and crosses indicate otherwise. The fraction indicates how many measurements in that row are statistically significantly positive, and bolded rows indicate when more than half of the metrics show the LLM response to be statistically significantly more like the out-group imitation than in-group representation. Overall on R1-Contingent and R2-Relevant, non-binary person and person with impaired vision are consistently more like out-group imitations.

3 LLMS FLATTEN GROUPS AND PORTRAY THEM ONE-DIMENSIONALLY

The second limitation is that because of loss functions like cross-entropy that are used to train LLMs, models are rewarded for producing the more likely outputs for any given piece of text. This flattens the representation of certain groups and erases subgroup heterogeneity (e.g., that within women, Black women are different than White women) [9, 10, 32]. This is especially harmful in the context of flattening demographic groups with a history of being portrayed one-dimensionally (e.g., Black people). Empirically, we find that all four LLMs on all questions tested, and across four different measures of diversity, generate responses for each identity group that are flatter than that of humans.

4 ESSENTIALIZING

Finally, we explore a slightly different reason one might prompt an LLM with demographic identities: to increase the coverage of the resulting responses in scenarios like anticipatory work where the goal is to generate a large range of responses rather than to represent different groups. Here, we find that prompting with behavioral personas or in some cases even astrology signs achieves the same effect of increasing coverage as prompting with sensitive demographic identities does. We argue that if such coverage can be achieved without unnecessary essentialization of identity, it likely should be. In these settings, identity-prompting LLMs can be seen as akin to designers leveraging user personas to try and see things from different perspectives [14]. However, personas have limitations, and may rely on stereotypes and reductionist representations about people [5, 23, 24, 26]. Thus, there is sometimes a recommendation among user researchers to move away from personas based on sensitive demographic attributes, which may reinforce stereotypes, and towards those based on behavioral characteristics [35]. Here we mirror this suggestion in the LLM space.

REFERENCES

[1] Linda Alcoff. 1991. The Problem of Speaking for Others. *Cultural Critique* (1991).

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* (2023).

[3] Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2017. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* (2017).

[4] Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness. *Empirical Methods in Natural Language Processing (EMNLP)* (2023).

[5] Christopher N. Chapman and Russell P. Milham. 2006. The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2006).

[6] Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. *Empirical Methods in Natural Language Processing (EMNLP)* (2023).

[7] Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. *Annual Meeting of the Association for Computational Linguistics (ACL)* (2023).

[8] Cheng-Han Chiang and Hung yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluation? *Annual Meeting of the Association for Computational Linguistics* (2023).

[9] Combahee River Collective. 1977. The Combahee River Collective Statement. (1977).

[10] Kimberle Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* (1989). Issue 1.

[11] ehartford. 2023. Wizard-Vicuna-7B-Uncensored. *Hugging Face* (2023). https://huggingface.co/ehartford/Wizard-Vicuna-7B-Uncensored

[12] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* (2023).

[13] Igor Grossmann, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. AI and the transformation of social science research. *Science* (2023).

[14] Jonathan Grudin. 2006. The persona lifecycle: Keeping people in mind. *Morgan Kaufmann* (2006).

[15] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. *International Conference on Learning Representations (ICLR)* (2024).

[16] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. *arXiv:2303.16854* (2023).

[17] Kristien Hens. 2021. Towards an Ethics of Autism: A Philosophical Exploration. *Open Book Publishers* (2021).

[18] Jennifer L. Hughes, Abigail A. Camden, and Tenzin Yangchen. 2016. Rethinking and Updating Demographic Questions: Guidance to Improve Descriptions of Research Samples. *Psi Chi Journal of Psychological Research* (2016).

[19] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)* (2023).

[20] Gauri Kambhatla, Ian Stewart, and Rada Mihalcea. 2022. Surfacing Racial Stereotypes through Identity Portrayal. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2022).

[21] Sharon L. Lohr. 2022. Sampling Design and Analysis. *Routledge* (2022).

[22] Bernard Marr. [n. d.]. The Amazing Ways Duolingo Is Using AI And GPT-4. *Forbes* ([n. d.]). https://www.forbes.com/sites/bernardmarr/2023/04/28/the-amazing-ways-duolingo-is-using-ai-and-gpt-4/?sh=56b4452a1346

[23] Nicola Marsden and Maren Haag. 2016. Stereotypes and Politics: Reflections on Personas. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)* (2016).

[24] Nicola Marsden and Monika Pröbster. 2019. Personas and Identity: Looking at Multiple Identities to Inform the Construction of Personas. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)* (2019).

[25] OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774* (2023).

[26] Nelly Oudshoorn, Louis Neven, and Marcelle Stienstra. 2016. How diversity gets lost: Age and gender in design practices of information and communication technologies. *Journal of Women Aging* (2016).

[27] Salvador Rodriguez, Deepa Seetharaman, and Aaron Tilley. [n. d.]. Meta to Push for Younger Users With New AI Chatbot Characters. *The Wall Street Journal* ([n. d.]). https://www.wsj.com/tech/ai/meta-ai-chatbot-younger-users-dab6cb32?mod=rss_Technology

[28] Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. 2021. Revealing Persona Biases in Dialogue Systems. *arXiv:2104.08728* (2021).

[29] Gayatri Chakravorty Spivak. 1988. Can the Subaltern Speak? *Marxism and the Interpretation of Culture* (1988).

[30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich,

Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288* (2023).

[31] Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems. *Findings of the Association for Computational Linguistics: EMNLP* (2023).

[32] Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2022).

[33] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, Jenny T. Liang, Ryan Liu, Ihita Mandal, Jeremiah Milbauer, Xiaolin Ni, Namrata Padmanabhan, Subhashini Ramkumar, Alexis Sudjianto, Jordan Taylor, Ying-Jui Tseng, Patricia Vaidos, Zhijin Wu, Wei Wu, and Chenyang Yang. 2023. LLMs as Workers in Human-Computational Algorithms? Replicating Crowdsourcing Pipelines with LLMs. *arXiv:2307.10168* (2023).

[34] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering Large Language Models to Follow Complex Instructions. *arXiv:2304.12244* (2023).

[35] Indi Young. 2016. Describing Personas. *Inclusive Software* (2016).

[36] Mingqian Zheng, Jiaxin Pei, and David Jurgens. 2023. Is "A Helpful Assistant" the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts. *arXiv:2311.10054* (2023).

[37] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? *Computional Linguistics* (2023).