Can AI Serve as a Substitute for Human Subjects in Software Engineering Research?

Marco Gerosa[1*], Bianca Trinkenreich[2], Igor Steinmacher[1], Anita Sarma[2]

[1*]Northern Arizona University, USA.
[2]Oregon State University, USA.

*Corresponding author(s). E-mail(s): Marco.Gerosa@nau.edu;
Contributing authors: Bianca.Trinkenreich@oregonstate.edu;
Igor.Steinmacher@nau.edu; Anita.Sarma@oregonstate.edu;

Abstract

Research within sociotechnical domains, such as Software Engineering, fundamentally requires the human perspective. Nevertheless, traditional qualitative data collection methods suffer from difficulties in participant recruitment, scaling, and labor intensity. This vision paper proposes a novel approach to qualitative data collection in software engineering research by harnessing the capabilities of artificial intelligence (AI), especially large language models (LLMs) like ChatGPT and multimodal foundation models. We explore the potential of AI-generated synthetic text as an alternative source of qualitative data, discussing how LLMs can replicate human responses and behaviors in research settings. We discuss AI applications in emulating humans in interviews, focus groups, surveys, observational studies, and user evaluations. We discuss open problems and research opportunities to implement this vision. In the future, an integrated approach where both AI and human-generated data coexist will likely yield the most effective outcomes.

1 Introduction

Software engineering (SE) is inherently a sociotechnical discipline [1]—considering the human perspective in research ensures that technological advancements are informed by the nuanced needs and complexities of those affected by the software. Therefore, it is not surprising that SE research commonly includes a qualitative component based on data collected through interviews, focus groups, surveys, observation, user studies, etc. Nevertheless, the recruitment and engagement of human participants, particularly from underrepresented groups, pose increasing challenges [2, 3].

AI has the potential to revolutionize human factors research. For instance, large language models (LLMs) have been explored to help qualitative data analysis by processing large amounts of text and identifying patterns and categories [4–6]—an LLM can quickly sift through thousands of pages of interview transcripts, observational notes, or social media posts, extracting and categorizing key phrases and topics.

But, what if we push the boundaries further? What if LLMs and other foundational AI models could be harnessed to substitute human subjects in qualitative research? These advanced AI systems, trained with vast amounts of data, can generate responses that closely mimic human-generated content. Embedded within these models are intricate patterns that reveal themselves in the content they generate in response to prompts. Changes in these prompts can dramatically alter the response's nature, affecting not just the conveyed information but also the structure, style, and diction of the generated text. Therefore, with strategic prompting, a foundation model can simulate a particular demographic profile. Researchers could then engage with these models, posing questions to elicit responses and behaviors that could serve as qualitative data. This data could then be analyzed to develop theories and validate software engineering tools, offering a novel dimension to qualitative research.

Before we proceed—*a disclaimer—we neither believe nor desire for AI to completely replace human subjects in software engineering research.* The purpose of this vision paper is to explore the concept and consider the possible supportive roles of AI. Looking ahead, the research community will establish practices to determine the optimal balance between AI-generated synthetic text and human-sourced data in qualitative software engineering research.

2 AI-Based Foundation Models as Alternates to Human Data Sources

2.1 Interviews: persona-based prompting

By constructing detailed personas and corresponding prompts, researchers can guide the AI to produce text that provides insights into how different groups might think, feel, or react to various stimuli or situations. By applying this approach, the AI could simulate the perspectives and feedback such personas might provide about different research topics. In essence, through persona-based prompt engineering, researchers can virtually 'interview' these AI-generated personae.

As illustrated in Table 1, researchers may specify extensive details about the persona. The language model connects this information to its pre-existing knowledge base to generate text that aligns coherently with the defined profile. In the table, we tweak the prompt to man and woman personas, keeping all the other characteristics. Interestingly, the responses generated by the large language model exhibit coherence with findings from previous research [7], which found that men are more likely than women to contribute to open source projects motivated by fun and for overcoming technical challenges, while women are more likely than men to be motivated by reciprocity and kinship. As we can observe in the synthetic interviews, the man starts his answer by focusing on technical hurdles and his journey through complex problem-solving, while the woman begins her dialogue by emphasizing community collaboration, learning experiences, and the value of contributing back to the community. Intriguingly, the woman even highlights her involvement in non-code contributions, particularly taking the lead in documentation—a point that also aligns with previous work indicating women's propensity towards engaging in non-coding roles in OSS [8].

Prompt:	**Response:**
You are John, a 45 years-old man living in Arizona, United States, who has worked in a big IT company for 20 years and started contributing to open source projects in Java during the weekends. I am going to interview you. Can you briefly tell me about your experience contributing to open source projects?	Absolutely. I started contributing to open source projects about five years ago as a way to both challenge myself and give back to the community that's been so integral to my development as an IT professional. Initially, it was a bit daunting. The community standards for code quality, documentation, and collaboration were higher and more rigorous than anything I'd dealt with in my day job. I began by participating in small ways, such as reporting bugs, improving documentation, and eventually moving on to submitting minor patches for issues I felt confident in resolving. Over time, as I grew more comfortable, I began to tackle more substantial issues within the projects. (...)
You are Mary, a 45 years-old woman living in Arizona, United States, who has worked in a big IT company for 20 years and started contributing to open source projects in Java during the weekends. I am going to interview you. Can you briefly tell me about your experience contributing to open source projects?	Sure, I've found contributing to open source projects to be an incredibly rewarding experience. Over the past few years, I've dedicated my weekends to working on various Java-based projects, which has allowed me to not only refine my coding skills but also to engage with a community of like-minded individuals who are passionate about collaborative software development. My contributions have ranged from fixing bugs to adding new features, and I've even taken the lead on documentation efforts for a couple of smaller projects. This has involved a lot of learning on my part, as I've had to familiarize myself with different project codes and collaborate with developers from around the world. I've found that working on open source projects has greatly complemented my day job at the IT company, where I also use Java extensively. It has been fulfilling to give back to the community that has been a cornerstone of my professional development for so many years. Plus, it's exciting to see code I've worked on being used in real-world applications.

Table 1 Using prompt engineering to interview specific personas. Responses generated by GPT4.

With the continued expansion of training datasets and model parameters, the AI models will become progressively more adept at capturing the nuanced distinctions in response generation for diverse personas. In this context, researchers can even explore intersectionality, thereby gaining a deeper comprehension of how overlapping social categorizations influence individual experiences and identities.

2.2 Focus Groups: multi-persona prompting

Focus groups convene a carefully selected group of individuals with expertise in relevant areas and capitalize on the group dynamics to capture collective viewpoints, shared experiences, and areas where opinions diverge significantly. Language models can be prompted to create an interaction among multiple personas. By generating distinct voices and perspectives within a single narrative, these models can effectively mimic a diverse dialog in a focus group among various characters. Table 2 exemplifies this capacity, presenting the initial segment of a conversation generated by the model. In the full dialog (available at https://chat.openai.com/share/7c9d8697-fff9-4061-ba63-fcf1e07f9545), we observe characters engaging in a rich exchange, with one persona notably developing ideas introduced by others to articulate their viewpoint.

2.3 Surveys: mega-persona prompting

Researchers can also prompt language models to emulate how a population would answer survey questions. To exemplify this idea, we prompted an LLM using the demographic data obtained from an actual survey [7]. The responses to the five first questions of the survey are aggregated in Table 3. Check the full dialog at https://chat.openai.com/share/4e3cc425-6550-481d-be6c-8fd9ed1a794b. The model

Prompt:	**Response:**
You are the coordinator of a multi-round collaboration process. Run multiple rounds of conversation until a final solution is reached. The participants will not participate in the future development of the solution but will provide critical comments and detailed suggestions whenever necessary and can consider each other's comments in their responses. The participants are: Pat, a project maintainer with many years of experience; John, a male newcomer who is also a novice in computer science; Mary, a female newcomer who has experience in other OSS projects; Sam, a web designer worried about accessibility, diversity, equity, and inclusion; and Sandra, a software architect. Your task is to design a web portal to support the onboarding of newcomers to an open source project.	**Coordinator:** Welcome, everyone. We're here to design a web portal that will support the onboarding of newcomers to our open source project. We need to ensure that the portal is welcoming, informative, and easy to navigate for all users, regardless of their experience level or background. Let's start by gathering initial ideas from each of you on what you think are the most important features or aspects we should include. Pat, could you please start us off? **Pat:** Thanks. Based on my experience, the portal should have a clear and concise "Getting Started" guide that includes setting up the development environment, a tutorial for the first contribution, and a list of beginner-friendly issues. We also need a clear code of conduct and contribution guidelines to set expectations right from the start. **John:** I think it would be helpful to have a glossary of terms. As a novice, I often get lost in jargon. Also, maybe some interactive elements like quizzes or checklists to keep track of learning (...)

Table 2 Using prompt engineering to simulate a focus group. Response generated by GPT4.

responses exhibit an impressive resemblance with the actual data, with an average deviation of just **4%**.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree	Not sure
Q1: I contribute to OSS because I have fun writing programs.						
Real Survey	63%	26%	8%	3%	0%	1%
LLM	60%	20%	10%	5%	1%	5%
Difference (abs)	3%	6%	2%	2%	1%	4%
Q2: I contribute to OSS because I need the software for my work.						
Real Survey	33%	33%	14%	13%	8%	0%
LLM	35%	30%	15%	10%	5%	5%
Difference (abs)	2%	3%	1%	3%	3%	5%
Q3: I contribute to OSS because I need the software for non-work purposes.						
Real Survey	22%	34%	19%	12%	12%	2%
LLM	25%	20%	20%	15%	10%	10%
Difference (abs)	3%	14%	1%	3%	2%	8%
Q4: I contribute to OSS because I can solve a problem that couldn't be solved by proprietary software.						
Real Survey	30%	31%	18%	8%	9%	4%
LLM	40%	30%	15%	10%	2%	3%
Difference (abs)	10%	1%	3%	2%	7%	1%
Q5: I contribute to OSS because I want to develop and improve my skills.						
Real Survey	61%	31%	6%	2%	1%	0%
LLM	50%	25%	10%	5%	2%	8%
Difference (abs)	11%	6%	4%	3%	1%	8%

Table 3 Differences between real survey data and LLM-generated responses.

An alternative approach to implementing surveys on a large language model is to create a virtual population of personas that follow the distribution of the desired demographics. In this one-at-a-time approach, each persona within this population would be prompted to respond to survey questions individually.

2.4 Observation and user experiments: Multimodal foundational models

Qualitative research often relies on inputs beyond text. Currently, there are no foundational models that can replicate the nuanced spectrum of human behavior within a given software engineering environment. However, the concept of training such a model is not beyond the realm of possibility, should the necessary datasets become available.

A sophisticated AI model could be trained on a comprehensive video dataset that captures software engineering professionals at work, meticulously annotated with the tasks they undertake. A model trained on this dataset could, theoretically, be prompted to project the sequence of actions a developer might take to fulfill a given new task in their work environment. Such a model would extend the capacity of AI beyond verbal interaction, simulating physical behavior. It could generate a variety of scenarios: a developer collaborating with end users to elicit requirements, the team engaging in a sprint retrospective, an individual brainstorming at a whiteboard, etc.

Additionally, software engineering research frequently leads to the development of new tools, with their efficacy typically evaluated through case studies, lab studies, or field deployment. Building upon the hypothetical scenario outlined above, a foundational model could be trained upon a huge dataset of software professionals interacting with tools. The model could then be used to predict the interaction of software professionals with new tools, enabling simulated evaluations under controlled conditions. By tagging the training dataset with the demographic details of software professionals, researchers could utilize persona-based prompts, akin to the methods discussed earlier, to approximate the engagement of distinct demographic groups with a given tool. For example, considering the established research indicating gender-based differences in technology adoption and interaction (e.g., [9]), one could expect that the foundational model, cognizant of these disparities, might forecast the unique engagement patterns of men and women with new tools, allowing designers to anticipate issues and fix inclusivity bugs. While these scenarios are currently speculative propositions, they invite us to ponder the future possibilities of AI in qualitative research and the potential impacts on the field of software engineering.

3 Open Problems and Research Opportunities

Ethical Considerations. As foundational models take on a more significant role in emulating human responses, ethical considerations come to the forefront. We need governance frameworks that regulate the use of AI in this capacity. This includes developing transparent methodologies, ensuring stakeholders are informed of their operational mechanics and consequences for the simulated individuals.

Enhancing Fairness. Foundational models, proficient at pattern recognition, mirror societal biases from the training datasets [10]. In certain contexts, biases are crucial to capture an authentic perspective—eliminating all biases could undermine the capability of language models to accurately emulate specific personas. However, researchers should be aware that the model may be tainted by external societal perceptions, which may not necessarily match the population's behavior. Moreover, using AI as a replacement for humans risks further marginalizing underrepresented groups since stereotypes and the perspective of the majority will be prevalent, exacerbating an already widespread issue. Therefore, researchers need to actively work to prevent their perpetuation in some cases. The issue of bias and fairness is not unique to AI, but the scalability and automation potential of foundational models amplify the risk of widespread propagation.

Rigorous Evaluation. A critical step in advancing AI-based foundational models as alternatives for human data sources is the rigorous evaluation of their outputs' fidelity. Researchers should consider replicating qualitative studies to gauge the extent to which AI can yield comparable results, with the caveat that prior research could have been incorporated into the models' training datasets. New research that collects data simultaneously from AI and humans can provide a more realistic picture. Yet, any evaluation may quickly become dated since the models are rapidly evolving and becoming better at generating human-like text. Moreover, the accuracy and applicability of these models are likely to vary across different domains and research questions. Understanding and delineating the contexts in which these models provide high-quality data versus those where they may introduce biases or inaccuracies are fundamental questions that must be systematically explored. Future research can also create benchmarks and standards that compare AI-generated data with human-generated data, helping to ascertain the reliability and validity of prompting approaches.

Feedback Loops in AI Training. Soon, a large portion of text available will be generated or enhanced by AI. When an AI-generated text is used as part of the training corpus for future AI iterations, there is a potential for the model to become increasingly insular and detached from genuine human input. This self-referential cycle could lead to amplification of inherent model biases, reduction in the diversity of generated responses, and potentially the emergence of new, unintended biases within the data. The prospect of AI 'echo chambers' necessitates careful research into methodologies for detecting and mitigating feedback loops.

Setting the right level of randomness. Just as human respondents may occasionally deviate from their typical patterns, a language model should also introduce an element of randomness to simulate this aspect of human behavior. This can be achieved by fine-tuning the model's hyperparameters. However, there is a delicate equilibrium to be maintained: increasing randomness may lead to a rise in hallucinations—instances where the model generates content that is not grounded in its training data. Further research is necessary to establish the optimal parameters.

Detecting hallucinations. Ensuring the reliability of AI outputs is critical and may involve tools for validating generated content by, for example, cross-referencing with online sources. Researchers can also generate results using multiple models to assess consistency across various configurations and training data sources. Finally, researchers could conduct sanity checks with a small sample of humans and compare it to the automatic results. Nevertheless, humans may resist questions that are illogical, biased, offensive, or unsuitable. In contrast, LLMs tend to generate responses regardless of the question's nature, only restricted by their pre-programmed guardrails.

Multimodal models. While current foundational models primarily handle textual data, qualitative data also encompasses visual, auditory, and behavioral information. Developing multimodal foundational models that can synthesize and interpret such data would greatly expand the horizons of AI in qualitative research.

Persona Specification. Differentiating between personas with subtle or complex characteristics remains a challenge. Current models may struggle to consistently capture the intricacies of human behavior and societal nuances that influence individual experiences. Research opportunities include developing methods to enhance the sensitivity of models to such nuances and the ability to handle intersectionality more adeptly. Further work could also examine how personas evolve over time and how models might simulate this progression. Future research can also investigate the optimal number of ‘interviews’ for each persona.

Context Specification. Research in linguistics has long established that humans intuitively perceive and adjust their discourse to a specific interactional context [11, 12]. For language models to effectively mirror human-like data collection, these interactional contexts must be precisely encoded within the prompts. However, it is challenging to represent the whole context in textual form in a prompt. Future research can determine which situational parameters are most influential and how they can be intricately woven into prompt designs.

Precision in Demographic Representation. For simulating surveys (Section 2.3), without a precise representation of the sub-populations’ characteristics, such as age, region, gender, experience, education, etc., it becomes challenging to construct a representative prompt that generates synthetic responses akin to those of the intended demographic. To address this issue, conducting studies to gather detailed demographic information is crucial. We also need a deeper understanding of the attributes most pertinent for persona creation. Some prior work has identified a few characteristics that influence how people interact with software [9] and can be used as baseline. Complementary, large language models can help refine persona definitions [13–16].

Complex Group Dynamics in Multi-Persona Dialogues. In focus groups (Section 2.2), ensuring the authenticity and naturalness of multi-persona dialogues is a key concern. Research can investigate how to better simulate the spontaneity of human interactions, including managing and reflecting the nuances of group dynamics, power imbalances, and conversational flow. Research could explore algorithms that manage turn-taking, conflict resolution, and the emergence of group leaders or influencers.

4 Related Work

Researchers in the social sciences and beyond have begun discussing the use of AI to generate qualitative data. As stated in an article in the prestigious Science journal [17], *”it is plausible that we will have a system within a few years that can just be placed into any experiment and will produce behavior indistinguishable from human behavior.”* Argyle et al. [18] suggest that models with sufficient algorithmic fidelity constitute a novel tool to advance the understanding of humans and society. Other works discuss when language models might replace human participants in psychological science [19, 20] and how models can embody assigned personality traits in user personas [21]. A few studies have compared human surveys with the model output [22–26] and found impressive results. Conversely, other researchers suggest caution in treating survey responses from language models as equivalent to those of human populations at the present time [27]. A specific study [28] found that current LLMs fail to represent the perceptions of some subpopulations. Aher et al. [29] present the Turing Experiment test to evaluate how AI can simulate a group of participants and demonstrate that language models accurately replicate findings from classic experiments across several disciplines. Finally, Simmon and Hare [30] present a review of using LLMs as subpopulation representatives.

Given the nascent nature of this subject area, much of the related work currently resides in preprint repositories like Arxiv, which are not peer-reviewed. We invite the software engineering community to closely follow the area's development and undertake dedicated research efforts to thoroughly examine and substantiate this idea within the context of software engineering.

5 Conclusion

Can AI serve as a substitute for human subjects in software engineering research? This provocative question challenges traditional paradigms and opens a Pandora's box of ethical, methodological, and practical considerations. If we could simulate conversations, generate narratives, and model complex human behaviors with sufficient accuracy, the potential for scaling research efforts could be unprecedented. However, this scenario also raises critical concerns about the authenticity of synthetic qualitative data, the loss of nuanced human insights, and the ethical implications of reducing the human experience to algorithmically generated data. While these models could manage some aspects of data gathering and analysis, they lack the innate human ability to contextualize and empathize with the subjective complexities of human stories. It is, therefore, crucial to approach such a possibility with caution, ensuring that the human element remains at the forefront of qualitative inquiry.

This paper does not propose to replace human subjects, but rather to explore the boundaries of AI's capabilities and to discuss the implications of its use as a tool in qualitative research. Innovations often cause the feeling that they will replace the previous status quo. Yet, what eventually happens is a symbiotic relationship where both old and new coexist and enhance one another. In the specific context of qualitative data in software engineering research, we are strongly convinced that a balanced, integrated approach is more likely to yield the most effective solutions. Synthetic and human-generated data are both poised to have their unique applications—the boundaries will become clearer through ongoing experimentation, scholarly debate, and the iterative process of community consensus. For example, AI can be instrumental in the early stages of research, such as pilot studies or experimental design, offering substantial savings in time and resources. They may also prove invaluable in scenarios where human involvement would be impractical, unethical, or unsafe [17]. Anyway, reviewers in top-tier venues may become more critical of research that relies on simplistic data collection methodologies that could be easily generated by AI. As the field progresses, the expectation is that studies will demonstrate more sophisticated and nuanced use of data to provide insights beyond the reach of AI's generative abilities. The collective wisdom and discernment of the community will determine the most effective and ethical applications of these diverse data collection approaches in research.

6 Acknowledgments

Partial support of the NSF grants 2236198, 2235601, 2247929, 2303043, and 2303042. ChatGPT v4 was used to copy-edit this article.

References

[1] Storey, M.-A., Ernst, N.A., Williams, C., Kalliamvakou, E.: The who, what, how of software engineering research: a socio-technical framework. Empirical Software Engineering **25**, 4097–4129 (2020)

[2] Kokinda, E., Moster, M., Dominic, J., Rodeghero, P.: Under the bridge: Trolling and the challenges of recruiting software developers for empirical research studies. In: 2023 IEEE/ACM 45th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER), pp. 55–59 (2023). https://doi.org/10.1109/ICSE-NIER58687.2023.00016

[3] Smith, M., Danilova, A., Naiakshina, A.: A meta-research agenda for recruitment and study design for developer studies. In: 1st International Workshop on Recruiting Participants for Empirical Software Engineering (RoPES'22), 2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE) (2022)

[4] Xiao, Z., Yuan, X., Liao, Q.V., Abdelghani, R., Oudeyer, P.-Y.: Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 75–78 (2023)

[5] Chew, R., Bollenbacher, J., Wenger, M., Speer, J., Kim, A.: LLM-assisted content analysis: Using large language models to support deductive coding. arXiv preprint arXiv:2306.14924 (2023)

[6] Dai, S.-C., Xiong, A., Ku, L.-W.: LLM-in-the-loop: Leveraging large language model for thematic analysis. arXiv preprint arXiv:2310.15100 (2023)

[7] Gerosa, M., Wiese, I., Trinkenreich, B., Link, G., Robles, G., Treude, C., Steinmacher, I., Sarma, A.: The shifting sands of motivation: Revisiting what drives contributors in open source. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), pp. 1046–1058 (2021). IEEE

[8] Trinkenreich, B., Wiese, I., Sarma, A., Gerosa, M., Steinmacher, I.: Women's participation in open source software: A survey of the literature. ACM Transactions on Software Engineering and Methodology (TOSEM) **31**(4), 1–37 (2022)

[9] Burnett, M., Stumpf, S., Macbeth, J., Makri, S., Beckwith, L., Kwan, I., Peters, A., Jernigan, W.: Gendermag: A method for evaluating software's gender inclusiveness. Interacting with Computers **28**(6), 760–787 (2016)

[10] Treude, C., Hata, H.: She elicits requirements and he tests: Software engineering gender bias in large language models. arXiv preprint arXiv:2303.10131 (2023)

[11] Biber, D.: Text-linguistic approaches to register variation. Register Studies **1**(1),

42–75 (2019)

[12] Chaves, A.P., Egbert, J., Hocking, T., Doerry, E., Gerosa, M.A.: Chatbots language design: The influence of language variation on user experience with tourist assistant chatbots. ACM Transactions on Computer-Human Interaction **29**(2), 1–38 (2022)

[13] Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., Ji, H.: Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. arXiv preprint arXiv:2307.05300 (2023)

[14] Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J.: Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910 (2022)

[15] Jung, S.-g., Salminen, J., Kwak, H., An, J., Jansen, B.J.: Automatic persona generation (APG) a rationale and demonstration. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, pp. 321–324 (2018)

[16] De Paoli, S.: Improved prompting and process for writing user personas with LLMs, using qualitative interviews: Capturing behaviour and personality traits of users. arXiv preprint arXiv:2310.06391 (2023)

[17] Hutson, M., Mastin, A.: Guinea pigbots. Science (New York, NY) **381**(6654), 121–123 (2023)

[18] Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: Using language models to simulate human samples. Political Analysis **31**(3), 337–351 (2023)

[19] Dillion, D., Tandon, N., Gu, Y., Gray, K.: Can AI language models replace human participants? Trends in Cognitive Sciences **27**(7), 597–600 (2023) https://doi.org/10.1016/j.tics.2023.04.008

[20] Demszky, D., Yang, D., Yeager, D.S., Bryan, C.J., Clapper, M., Chandhok, S., Eichstaedt, J.C., Hecht, C., Jamieson, J., Johnson, M., et al.: Using large language models in psychology. Nature Reviews Psychology, 1–14 (2023)

[21] Jiang, H., Zhang, X., Cao, X., Kabbara, J., Roy, D.: PersonaLLM: Investigating the ability of GPT-3.5 to express personality traits and gender differences. arXiv preprint arXiv:2305.02547 (2023)

[22] Hämäläinen, P., Tavast, M., Kunnari, A.: Evaluating large language models in generating synthetic HCI research data: A case study. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3580688 . https://doi.org/10.1145/3544548.3580688

[23] Kim, J., Lee, B.: AI-augmented surveys: Leveraging large language models for opinion prediction in nationally representative surveys. arXiv preprint arXiv:2305.09620 (2023)

[24] Sanders, N.E., Ulinich, A., Schneier, B.: Demonstrations of the potential of AI-based political issue polling. arXiv preprint arXiv:2307.04781 (2023)

[25] Suguri Motoki, F.Y., Monteiro, J., Malagueño, R., Rodrigues, V.: From data scarcity to data abundance: Crafting synthetic survey data in management accounting using ChatGPT. Available at SSRN (2023)

[26] Eliot, L.: The bold promise Of mega-personas as a new shake-up for prompt engineering generative AI techniques. Accessed: 2023-11-08 (2023). https://www.forbes.com/sites/lanceeliot/2023/08/15/the-bold-promise-of-mega-personas-as-a-new-shake-up-for-prompt-engineering-generative-ai-techniques/?sh=2be155065552

[27] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., McHardy, R.: Challenges and applications of large language models. arXiv preprint arXiv:2307.10169 (2023)

[28] Lee, S., Peng, T.-Q., Goldberg, M.H., Rosenthal, S.A., Kotcher, J.E., Maibach, E.W., Leiserowitz, A.: Can large language models capture public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. arXiv preprint arXiv:2311.00217 (2023)

[29] Aher, G.V., Arriaga, R.I., Kalai, A.T.: Using large language models to simulate multiple humans and replicate human subject studies. In: International Conference on Machine Learning, pp. 337–371 (2023). PMLR

[30] Simmons, G., Hare, C.: Large language models as subpopulation representative models: A review. arXiv preprint arXiv:2310.17888 (2023)