Can ChatGPT emulate humans in software engineering surveys?

Igor Steinmacher
Northern Arizona University
United States
Igor.Steinmacher@nau.edu

Jacob Mcauley Penney
Northern Arizona University
United States
jmp458@nau.edu

Katia Romero Felizardo*
UTFPR-CP
Brazil
Northern Arizona University
USA
karfelizardo@gmail.com

Alessandro F. Garcia*
PUC-Rio
Brazil
Northern Arizona University
USA
afgarcia@inf.puc-rio.br

Marco A. Gerosa
Northern Arizona University
United States
marco.gerosa@nau.edu

Abstract

Context: There is a growing belief in the literature that large language models (LLMs), such as ChatGPT, can mimic human behavior in surveys. **Gap:** While the literature has shown promising results in social sciences and market research, there is scant evidence of its effectiveness in technical fields like software engineering. **Objective:** Inspired by previous work, this paper explores ChatGPT's ability to replicate findings from prior software engineering research. Given the frequent use of surveys in this field, if LLMs can accurately emulate human responses, this technique could address common methodological challenges like recruitment difficulties, representational shortcomings, and respondent fatigue. **Method:** We prompted ChatGPT to reflect the behavior of a 'mega-persona' representing the demographic distribution of interest. We replicated surveys from 2019 to 2023 from leading SE conferences, examining ChatGPT's proficiency in mimicking responses from diverse demographics. **Results:** Our findings reveal that ChatGPT can successfully replicate the outcomes of some studies, but in others, the results were not significantly better than a random baseline. **Conclusions:** This paper reports our results so far and discusses the challenges and potential research opportunities in leveraging LLMs for representing humans in software engineering surveys.

CCS Concepts

• **General and reference** → **Empirical studies**.

Keywords

Generative AI, Replication Study, Mega-Personas, Survey

ACM Reference Format:
Igor Steinmacher, Jacob Mcauley Penney, Katia Romero Felizardo, Alessandro F. Garcia, and Marco A. Gerosa. 2024. Can ChatGPT emulate humans in software engineering surveys?. In *Proceedings of the 18th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '24), October 24–25, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3674805.3690744

*Also with Northern Arizona University, Flagstaff, AZ, USA

ESEM '24, October 24–25, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-1047-6/24/10
https://doi.org/10.1145/3674805.3690744

1 Introduction

Surveys have multiple applications in Software Engineering (SE), including confirming trends, understanding how developers receive new technologies or tools, exploring social phenomena in software development, and understanding population-wide trends and perspectives. However, traditional data collection methods in survey research often face challenges such as participant recruitment, scaling, and the representation of diverse perspectives. These long-lasting problems have been widely discussed in multiple papers in SE research [12, 18, 26, 27, 29, 33, 39].

With the recent advances in large language models, such as those employed by ChatGPT, one might envision whether they might help to address these challenges. There is growing evidence to support this vision in other fields. Jiang and colleagues [22] suggest that LLMs might replace human participants in psychological science. Similarly, other researchers [2, 20, 25, 30, 34] conclude that LLMs can generate credible responses in nationally representative surveys, political polling, management, and gaming. Hutson [21] is enthusiastic, stating that it is plausible that we will have a system within a few years that can just be placed into any study and produce behavior indistinguishable from human behavior.

In this context, we investigate the vision of whether ChatGPT can emulate humans in SE surveys. Based on previous work [17], we explore ChatGPT's ability to replicate findings from prior software engineering research. We use the concept of "mega-personas", as proposed in the work by Gerosa et al. [17], to emulate human participants in surveys. In this approach, large language models (LLMs) are prompted to act like a population that follows the distribution of the desired demographics [8, 43], and the LLM emulates human responses to the questionnaire.

This paper reports the emerging results of our study. We employed ChatGPT [31] to replicate survey responses from studies published in notable SE conferences such as FSE, ASE, and ICSE between 2019 and 2023. Our findings reveal that ChatGPT can successfully replicate the outcomes of some studies, but in others, the results were not significantly better than a random baseline. Even

though we do not perform an exhaustive validation nor define guidelines, our results contribute to the discourse on the evolving role of AI in enhancing the breadth and depth of SE research, aligning with the growing interest in integrating AI into research methodologies [37, 40, 42].

2 Method

The vision investigated in this paper is that large language models, as used by ChatGPT, can emulate humans and help to address some of the challenge in SE surveys. We perform a study by performing a , using ChatGPT 4 [31] to generate the distribution of responses for previous survey questions and compare them to the actual results. All artifacts and scripts are available in our replication package [7].[1]

2.1 Data Collection

We started by (i) finding studies to replicate and (ii) collecting data from these studies.

2.1.1 Studies selection. As illustrated in Figure 1, to select our studies of interest, we sampled papers from three premier conferences in SE: the International Conference on Foundations in Software Engineering (FSE), the International Conference on Automated Software Engineering (ASE), and the International Conference on Software Engineering (ICSE). Since we are interested in generating contemporary responses from the LLM, we focused on full papers published on the main track of the five latest editions of the three conferences (2019–2023). We collected 3,617 papers via digital libraries, such as ACM DL, IEEE Xplore, and Google Scholar. Then, we filtered and sampled the papers according to the steps below:

IC_1. Selecting studies that conducted surveys. We performed a textual search to find papers that conducted surveys, looking for survey-identifying terms: "survey" and "questionnaire". After this step, we kept 190 candidate studies.

IC_2. Excluding studies from collocated events and other tracks. We manually inspected the 190 papers to filter out those not on the main track. After this step, we ended up with 124 studies.

IC_3. Checking for replication package availability. As we needed the full instrument to replicate the study, we analyzed each of the 124 studies to look for those that had explicitly provided a replication package. We skimmed through the studies for the presence of the instrument or replication package and searched for a set of word steams to help us find the packages: "replic," "package," "supplem," "artifact," "http,"'zenodo," "figshare," "OSF," "bit.ly," "github." This step reduced our candidates to 75.

IC_4. Analyzing the type of study. We analyzed the replication packages and the paper to understand the type of research and the applicability of the survey. We dismissed literature surveys and cases where intervention or pre-survey activity was required (e.g., participants had to use an experimental tool before the survey). After this step, we kept 52 candidate studies.

Randomly selecting studies. We randomly sampled 14 papers to extract data and emulate the survey. We limited the number of analyzed surveys due to the manual work necessary to replicate the studies.

Further analysis on data. During the data extraction, we analyzed each study to ensure the feasibility of using their data. We discarded 4 studies because it was not possible to collect the responses to the survey. This happened because (i) the data was presented only in graphics, with no way to extract the exact values/percentages; or (ii) the results were presented in aggregated form or omitting some of the results (X% strongly agreed or agreed, or even Y% were positive about Item 1), without presenting the details about the other options. It is worth mentioning that the data was not available on the paper nor the supplementary. We ended up analyzing 10 studies.

2.1.2 Data Extraction. For each paper in our sample, one of the authors went through the replication package and the method and results sections of the paper, and two others reviewed the data extracted. The goal was to extract the following information:

- Context of the study: a short description of the goal of the survey;
- Population of interest: definition of the population that was surveyed (e.g., C++ developers from Germany; newcomers to the software industry; DevOps from Company_A)
- Demographic collection: for example, gender, age, role, level of experience, team size, organization size, country, etc. This information is important to characterize the population to emulate since we expect the LLM to consider the demographics.
- Questions collection: we collected the close-ended questions/items from each instrument, the options provided, and what kind of answer they would require (multiple option/-multiple choice).
- Results: lastly, we collected the distribution of answers for each question.

We manually analyzed CSVs, PDFs, figures, tables, and text to collect the needed information and populate a spreadsheet.

2.2 Emulation Process

In the emulation process, we prompted ChatGPT 4 [31], to generate responses to the collected questions. For each study, we created a new conversation (to guarantee that the conversation context was reset).

We started by prompting the information about the study, including the *population of interest*, the *context of study*, and the *demographic distributions* reported. Key to this process was creating prompts that described the characteristics of the specific populations from the original surveys so that the model could behave as that specific population. For each study, we transformed the context, population of interest, and demographics into a prompt using the following template:

```
"You are a population of <N> <CONTEXTUAL POPULATION>.
I want you to respond to survey questions reflecting
how this population would answer the questions. The
survey is about <CONTEXT>. I am providing background
information and demographics so that you can pretend
that each of the contributors is a distinct individual.
<DEMOGRAPHIC INFORMATION>"
```

After prompting the context, we input each question of the survey, one at a time. For each question, we also followed a specific template to guarantee consistency. The template included the options available, the type of question, and optional details like "you

[1] https://zenodo.org/records/10962878

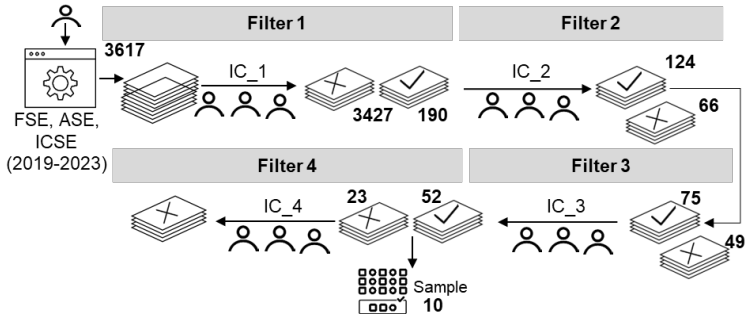

Figure 1: We followed two key phases: (1) searching for studies in major SE conferences: 3617 candidate studies; and (2) filtering studies through different inclusion criteria (IC_): IC_1: 190, IC_2: 124, IC_3: 75, and IC_4: 52 included studies, respectively. We analyzed 10 studies from this set.

may provide multiple answers" for the multiple-choice questions. Here is the template:

```
"<QUESTION X>
This is a <TYPE OF QUESTION> question <MAYBE DETAILS
LIKE, 0 to X answers are possible>
The options are: <OPTIONS>"
```

We collected the answers and added them to a CSV file. All the artifacts are available in our replication package [7].

2.3 Data Analysis

The responses generated by ChatGPT were compared against the original survey results. This comparison focused on assessing the extent to which the AI-generated responses matched the patterns and distributions of the human participants. We analyzed the delta between the answers for each survey item.

We also created a baseline by spreading evenly the responses of multiple-choice questions that allowed a single answer—for example, for a 5-point Likert scale question, we defined that each response would receive 20% of the answers, while for a boolean (Yes/No question), each response would receive 50% of the answers. We calculated the delta between this baseline and the original and ChatGPT answers.

Finally, we verified the characteristics of the survey, seeking patterns correlated with the results to inform our discussions.

3 Results

To investigate whether LLMs can be prompted to simulate how a population would answer software engineering-related survey questions, we prompted ChatGPT using the demographic data obtained from 10 published surveys [3, 6, 9–11, 19, 23, 36, 38, 41]. The model was instructed to emulate the responses based on the demographics of each survey.

Figure 2 shows the delta between the answers from ChatGPT and the original answers from the studies. For example, if 30% of the respondents chose option X for a given question and the outcome of ChatGPT for the same question was 32%, the delta is 2. It is important to notice that the deltas are absolute (or *modulus*) values, i.e., we were interested in the difference regardless of its direction.

ChatGPT shows effective performance in some contexts. In most studies (7 out of 10), the third quartile falls below 10%. Notably, these studies (except 141), exhibit minimal differences and low variance (IQR≤11.5). Conversely, for studies 9 and 44 more than half of the deltas approach or exceed 30% (for Study 9, only 1 answer out of 10 has a delta smaller than 10%, and the minimum for Study 44 is 13.4%). These two studies are composed only of Yes/No and Multiple Answer questions, which may be indicative of bad performance.

Motivated by these results, we compared the delta of the percentages (GPT − Original) according to the type of the questions (Likert, Multiple Option/Multiple Answer, Multiple Option/Single Answer, Boolean—like Yes/No). We conducted a Kruskal-Wallis test, which indicated differences across the groups ($p < 0.0001$). Dunn's posthoc test showed that Likert and Multiple Option/Single Answer items are smaller than Multiple Options/Multiple Answers and Boolean questions ($p < 0.05$).

We also compared the results produced by ChatGPT and the baseline we created (responses distributed evenly across the possible answers). As it is possible to see in Figure 3, ChatGPT outperforms the baseline in 6 of the studies (114, 3, 31, and 50). These differences are confirmed by a Mann-U statistical test (all of them with $p<0.05$ with Bonferroni correction). For the other 5 studies, we could not observe differences. Study 7 was not compared because it was composed only of multiple-choice questions that allowed multiple answers (thus, we could not create the baseline).

To better understand our results, we analyzed other characteristics of the surveys, aiming to find a correlation of our results with the type of demographic information available, context, population size, number of questions. We could not find any patterns, so further investigation is required to understand what would make the LLMs appropriate (or not) in SE.

4 Lessons Learned

In this paper, we bring evidence that emulating survey responses with an LLM may not provide accurate answers for all SE studies. Nevertheless, exploring LLM as alternates for human data sources in qualitative research opens many opportunities and reveals several challenges that deserve further investigation. In the following, we reflect upon open problems and research opportunities.

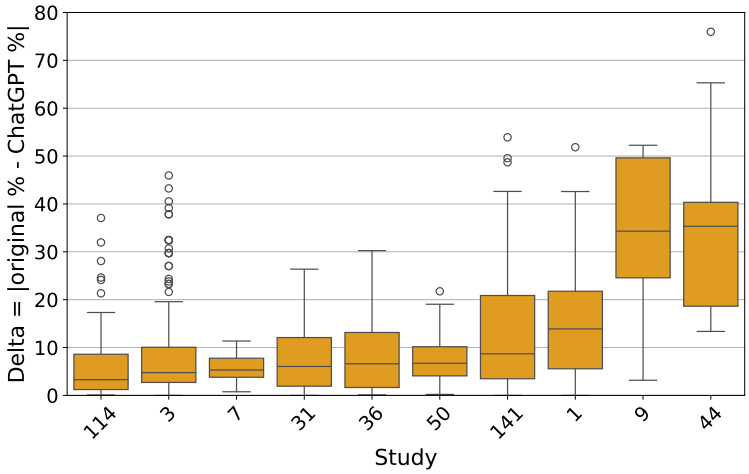

Figure 2: Comparing the delta between original results and ChatGPT outcomes (each question/item is one observation).

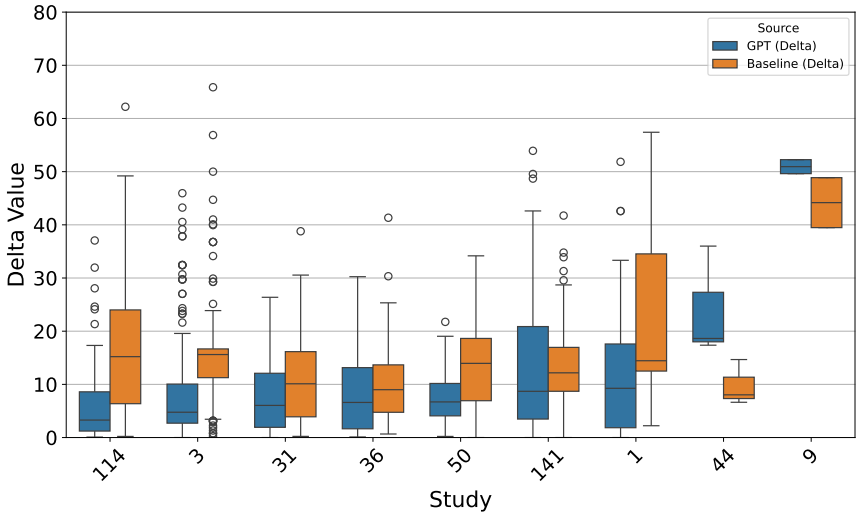

Figure 3: Comparing the delta to ChatGPT outcomes delta and the delta to the baseline (sorted by the ChatGPT delta).

LLMs will improve over time: While LLMs may not currently emulate all SE surveys accurately, we found impressive results in some analyzed studies. We can expect that the results may improve as the area evolves. The SE community needs to engage in a dialogue about human-based research in the context of the emerging capabilities of LLMs. This discussion should focus on the current state and anticipate future developments. By doing so, we can better understand how to integrate these advanced tools into our research methodologies, ensuring that they complement rather than replace human input and insight.

Demographics aspects representation: A central challenge of adopting LLMs in surveys is discerning and managing the demographics aspects. Capturing a population's perspective is crucial so that the LLMs accurately emulate specific personas. The surveys used in our study adopted from just 1 aspect to define the population [23] up to 7 [11]. The aspects also ranged from generic, such as role (student or professional), to specific, for example, robotics experience. We need a deeper understanding of the most pertinent aspects of persona creation. Creating prompts that represent populations depends on a detailed understanding of the demographic aspects of the population surveyed. Our experience extracting demographic aspects from papers revealed that SE studies need to present more detailed demographic information and standardize the way to present it. Future studies could investigate which demographic aspects are key (and how to present them in papers to facilitate extraction) to develop more accurate prompts. Still, it would be interesting to investigate how LLM answers change based on the granularity and breadth of demographics provided. If the models could accurately represent changes in the population, researchers can investigate whether the models can be used to study under-represented populations and intersectionalities.

Prompt sensibility: ChatGPT was sensitive to subtle changes in prompt formatting. Given the prompt's fragility, further research should target to understand better factors that impact prompts on LLMs (order sensitivity, context length, etc.) and find techniques that make them more accurate and reliable.

Rigorous assessment of LLM responses: One fear of using LLMs is that they could provide generic or erroneous answers. Researchers advertise that if LLMs do not "know" the answer, they will "fabricate" one [4, 22], making it difficult to distinguish between truth and falsehood answers (hallucinations). Therefore, before researchers can use LLM responses to replace (or extend) human data sources, the fidelity of the responses must be profoundly and widely validated. As we did in the present study, future research could replicate published studies to analyze human responses with those emulated by the models and analyze the reliability of the results. We adopted ChatGPT as LLM; moreover, our dataset [7] can be used to analyze some of the aspects of the prompting and assess the performance of other models.

Specialized domains: The adoption of LLMs in SE could be enhanced by training models on domain-specific data. For example, the surveys used in our research covered areas such as continuous integration [41], OSS [23], mentoring [15], and code review [9], DevOps [36], bots [6], robotics [16], among others. There is an opportunity to build datasets representing practices and behaviors of specialized areas and their population, increasing the relevance and accuracy of AI-generated data within those areas.

Ethical Considerations: Ethical considerations certainly emerge when LLMs emulate human responses. There is a need for research to regulate the use of AI as an alternative source of qualitative data to replicate human responses and behaviors in research settings to safeguard the interests of both the research community and the population represented by the models.

5 Limitations and threats to validity

Limited surveys and demographic diversity bias: Limited surveys and demographic diversity may have impacted our results. Even though we adopted a systematic process to select studies to replicate, our sample is restricted to recent editions of top-tier SE conferences. We plan to conduct a more comprehensive study, including more surveys and a diverse demographic pool.

Data extraction: Data extraction is one of the most prone to errors because raw data were manually transferred from the studies into the spreadsheet template. To mitigate this bias, we created a data extraction template and involved more than one researcher. We conducted this manually because the papers did not follow a standard way of presenting the data.

Data generalization: The limited number of studies means that results may not represent the full spectrum of the SE domain. Moreover, even though ChatGPT was the state-of-the-art model at the time of this writing, other LLMs may show different performance.

Training data overlap: A potential bias stems from the overlap between our study materials and the training data of the LLMs. Given that our study relies on academic papers from top-tier SE conferences, which may be publicly available, it is possible that they were included in training the LLM. Nevertheless, we observe that the LLM could not accurately predict the responses for some papers. We invite researchers to emulate responses from future surveys before writing about them to enable an independent comparison.

6 Related Work

Recent studies have highlighted the potential of LLMs in replicating human-like responses and personalities in various research settings. Dillion et al. [14] explore whether and when LLMs might replace human participants in psychological science. Jiang et al. [22] show that LLMs can generate content that aligns with their assigned personality profiles when adequately prompted. Hämäläinen et al. [20] investigated how LLMs can generate open-ended questionnaire responses about experiencing video games as art. Kim and Lee [25], Sanders et al. [34], and Motoki et al. [30] conclude that LLMs can generate credible responses in nationally representative surveys, political polling, management, and gaming.

Conversely, Kaddour et al. [24] suggest caution in treating LLM survey responses as equivalent to those from human populations. Agnew et al. [1] reviewed previous studies and identified and critiqued the motivation to use LLM to replace human data, highlighting the loss of values like participation, representation, and inclusion in AI-driven qualitative research. Demszky et al. [13] argue that although LLMs have the potential to advance psychological measurement, experimentation, and practice, they are not yet ready for many of the most transformative psychological applications. Simmons and Hare [35] also present a review of using LLMs as subpopulation representative models. Argyle et al. [5] propose a systematic way to evaluate and measure how closely LLM results align with real human subpopulations.

Like the works previously mentioned, our research also analyzes the use of LLMs to support research, contributing to the ongoing discourse about the role of AI in data collection and its potential applications in SE research. The use of LLMs in research is an emerging area; most of the related work is not peer-reviewed and available only in preprint repositories like Arxiv [1, 22, 24, 25, 28, 30, 32, 34, 35]. There is limited evidence in SE [17], and our community must devote substantial effort to thoroughly evaluating the use of AI in surveys and its potential effects on the field.

7 Conclusion

Generating survey data for SE with sufficient accuracy is an unprecedented alternative for scaling research efforts. However, this alternative also raises concerns about managing some demographic aspects of data gathering and ethical implications. Therefore, although recent advances in generative AI have attracted growing interest in replacing human participants in surveys, research is still needed for its effective use in SE research.

Acknowledgments

This work was partially funded by National Science Foundation (NSF) grants #2236198, #2247929, and #2303042 and the Brazilian National Council for Scientific and Technological Development (CNPq), Grant #3023392022-1. All authors have contributed to all activities related to the submission.

References

[1] W. Agnew, A. S. Bergman, J. Chien, M. Diaz, S. El-Sayed, J. Pittman, S. Mohamed, and K. R. McKee. 2024. The illusion of artificial inclusion. arXiv:2401.08572 [cs.CY]

[2] G. V. Aher, R. I. Arriaga, and A. T. Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *40th International Conference on Machine Learning (ICML'23)*. PMLR, Honolulu, Hawaii, USA, 337–371.

[3] R. S. Alsuhaibani, C. D. Newman, M. J. Decker, M. L. Collard, and J. I. Maletic. 2021. On the Naming of Methods: A Survey of Professional Developers. In *ICSE '21*. IEEE Press, Madrid, Spain, 587–599. https://doi.org/10.1109/ICSE43902.2021.00061

[4] A. Anghelescu, F. C. Firan, G. Onose, C. Munteanu, A. Trandafir, I. Ciobanu, S. Gheorghita, and V. Ciobanu. 2023. PRISMA Systematic Literature Review, including with Meta-Analysis vs. ChatbotGPT (AI) regarding Current Scientific Data on the Main Effects of the Calf Blood Deproteinized Hemoderivative Medicine (Actovegin) in Ischemic Stroke. *Biomedicines* 6, 11 (2023), 1–13.

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

[6] S. Asthana, H. Sajnani, E. Voyloshnikova, B. Acharya, and K. Herzig. 2023. A Case Study of Developer Bots: Motivations, Perceptions, and Challenges. In *ESEC/FSE'23* (San Francisco, CA, USA). Association for Computing Machinery, New York, NY, USA, 1268–1280. https://doi.org/10.1145/3611643.3616248

[7] Anonymous Author(s). 2024. Replication Package: Can ChatGPT emulate humans in software engineering surveys? https://doi.org/10.5281/zenodo.10578334

[8] N. Bodani, A. Lal, A. Maqsood, S. Altamash, N. Ahmed, and A. Heboyan. 2023. Knowledge, Attitude, and Practices of General Population Toward Utilizing ChatGPT: A Cross-sectional Study. *SAGE Open* 13 (2023), 1–9. https://doi.org/10.1177/21582440231211079

[9] L. Braz and A. Bacchelli. 2022. Software security during modern code review: the developer's perspective. In *ESEC/FSE'22* (Singapore). Association for Computing Machinery, New York, NY, USA, 810–821. https://doi.org/10.1145/3540250.3549135

[10] A. Chatterjee, M. Guizani, C. Stevens, J. Emard, M. E. May, M. Burnett, I. Ahmed, and A. Sarma. 2021. AID: An automated detector for gender-inclusivity bugs in OSS project pages. In *ICSE '21*. IEEE Press, Madrid, Spain, 1423–1435. https://doi.org/10.1109/ICSE43902.2021.00128

[11] B. Chen, L. Chen, C. Zhang, and X. Peng. 2021. BuildFast: history-aware build outcome prediction for fast feedback and reduced cost in continuous integration. In *ASE '20* (Virtual Event, Australia). Association for Computing Machinery, New York, NY, USA, 42–53. https://doi.org/10.1145/3324884.3416616

[12] S. Cruz, F. Q. B. da Silva, and L. F. Capretz. 2015. Forty years of research on personality in software engineering: A mapping study. *Computers in Human Behavior* 46 (2015), 94–113. https://doi.org/10.1016/j.chb.2014.12.008

[13] D. Demszky, D. Yang, D. S. Yeager, C. J. Bryan, M. Clapper, S. Chandhok, J. C. Eichstaedt, C. Hecht, J. Jamieson, M. Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology* 2, 1 (2023), 1–14.

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

[15] Z. Feng, A. Chatterjee, A. Sarma, and I. Ahmed. 2022. A case study of implicit mentoring, its prevalence, and impact in Apache. In *ESEC/FSE '22* (Singapore). Association for Computing Machinery, New York, NY, USA, 797–809. https://doi.org/10.1145/3540250.3549167

[16] S. García, D. Strüber, D. Brugali, T. Berger, and P. Pelliccione. 2020. Robotics software engineering: a perspective from the service robotics domain. In *28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Virtual Event, USA) *(ESEC/FSE '20)*. Association for Computing Machinery, New York, NY, USA, 593–604. https://doi.org/10.1145/3368089.3409743

[17] M. Gerosa, B. Trinkenreich, I. Steinmacher, and A. Sarma. 2023. Can AI serve as a substitute for human subjects in software engineering research? *Automated Software Engineering* 31, 13 (2023), 1–25. https://doi.org/10.1007/s10515-023-00409-6

[18] A. N. Ghazi, K. Petersen, S. S. V. R. Reddy, and H. Nekkanti. 2019. Survey Research in Software Engineering: Problems and Mitigation Strategies. *IEEE Access* 7, 1 (2019), 24703–24718. https://doi.org/10.1109/ACCESS.2018.2881041

[19] F. Grund, S. Chowdhury, N. C. Bradley, B. Hall, and R. Holmes. 2021. CodeShovel: Constructing Method-Level Source Code Histories. In *ICSE '21*. IEEE Press, Madrid, Spain, 1510–1522. https://doi.org/10.1109/ICSE43902.2021.00135

[20] P. Hämäläinen, M. Tavast, and A. Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. In *CHI'23* (Hamburg, Germany). Association for Computing Machinery, New York, NY, USA, Article 433, 19 pages. https://doi.org/10.1145/3544548.3580688

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

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

[23] A. Ju, H. Sajnani, S. Kelly, and K. Herzig. 2021. A Case Study of Onboarding in Software Teams: Tasks and Strategies. In *ICSE '21*. IEEE Press, Madrid, Spain, 613–623. https://doi.org/10.1109/ICSE43902.2021.00063

[24] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. 2023. Challenges and applications of large language models.

[25] J. Kim and B. Lee. 2023. AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys.

[26] B. A. Kitchenham and S. L. Pfleeger. 2002. Principles of survey research part 2: designing a survey. *SIGSOFT Softw. Eng. Notes* 27, 1 (jan 2002), 18––20. https://doi.org/10.1145/566493.566495

[27] E. Kokinda, M. Moster, J. Dominic, and P. Rodeghero. 2023. Under the Bridge: Trolling and the Challenges of Recruiting Software Developers for Empirical Research Studies. In *ICSE-NIER '23*. Association for Computing Machinery, Melbourne, Australia, 55–59. https://doi.org/10.1109/ICSE-NIER58687.2023.00016

[28] S. Lee, T.-Q. Peng, M. H. Goldberg, S. A. Rosenthal, J. E. Kotcher, E. W. Maibach, and A. Leiserowitz. 2023. Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias.

[29] J. S. Molléri, K. Petersen, and E. Mendes. 2016. Survey Guidelines in Software Engineering: An Annotated Review. In *ESEM '16*. Association for Computing Machinery, New York, NY, USA, Article 58, 6 pages. https://doi.org/10.1145/2961111.2962619

[30] S. Motoki, F. Yoshio, J. Monteiro, R. Malagueño, and V. Rodrigues. 2023. From Data Scarcity to Data Abundance: Crafting Synthetic Survey Data in Management Accounting using ChatGPT.

[31] OpenAI. 2024. ChatGPT (4) [Large language model]. https://chat.openai.com

[32] P. Petrak, T. T. Tran, and I. Gurevych. 2024. Learning from Emotions, Demographic Information and Implicit User Feedback in Task-Oriented Document-Grounded Dialogues. arXiv:2401.09248 [cs.CL]

[33] S. L. Pfleeger and B. A. Kitchenham. 2001. Principles of survey research: part 1: turning lemons into lemonade. *SIGSOFT Softw. Eng. Notes* 26, 6 (nov 2001), 16––18. https://doi.org/10.1145/505532.505535

[34] N. E Sanders, A. Ulinich, and B. Schneier. 2023. Demonstrations of the potential of AI-based political issue polling.

[35] G. Simmons and C. Hare. 2023. Large Language Models as Subpopulation Representative Models: A Review.

[36] D. Sokolowski, P. Weisenburger, and G. Salvaneschi. 2021. Automating serverless deployments for DevOps organizations. In *ESEC/FSE'21* (Athens, Greece). Association for Computing Machinery, New York, NY, USA, 57–69. https://doi.org/10.1145/3468264.3468575

[37] E. A. M. van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting. 2023. ChatGPT: five priorities for research. *Nature* 614, 7947 (2023), 224–226. https://doi.org/10.1038/d41586-023-00288-7

[38] E. L. Vargas, M. Aniche, C. Treude, M. Bruntink, and G. Gousios. 2020. Selecting third-party libraries: the practitioners' perspective. In *ESEC/FSE '20* (Virtual Event, USA). Association for Computing Machinery, New York, NY, USA, 245–256. https://doi.org/10.1145/3368089.3409711

[39] S. Wagner, D. Mendez, M. Felderer, D. Graziotin, and M. Kalinowski. 2020. *Challenges in Survey Research*. Springer International Publishing, Cham, 93–125. https://doi.org/10.1007/978-3-030-32489-6_4

[40] J. Wen and W. Wang. 2023. The future of ChatGPT in academic research and publishing: A commentary for clinical and translational medicine. *Clinical and Translational Medicine* 13, 3 (2023), e1207. https://doi.org/10.1002/ctm2.1207

[41] D. G. Widder, M. Hilton, C. Kästner, and B. Vasilescu. 2019. A conceptual replication of continuous integration pain points in the context of Travis CI. In *27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Tallinn, Estonia) *(ESEC/FSE '19)*. Association for Computing Machinery, New York, NY, USA, 647–658. https://doi.org/10.1145/3338906.3338922

[42] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu, J. Qiu, K. Hua, W. Su, J. Wu, H. Xu, Y. Han, Ch. Fu, Z. Yin, M. Liu, R. Roepman, S. Dietmann, M. Virta, F. Kengara, Z. Zhang, L. Zhang, T. Zhao, J. Dai, J. Yang, L. Lan, M. Luo, Z. Liu, T. An, B. Zhang, X. He, S. Cong, X. Liu, W. Zhang, J. P. Lewis, J. M. Tiedje, Q. Wang, Z. An, F. Wang, L. Zhang, T. Huang, C. Lu, Z. Cai, F. Wang, and J. Zhang. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* 2, 4 (2021), 100179. https://doi.org/10.1016/j.xinn.2021.100179

[43] T. Zack, E. Lehman, M. Suzgun, J. A. Rodriguez, L. A. Celi, J. Gichoya, D. Jurafsky, P. Szolovits, D. W. Bates, R.-E. E. Abdulnour, A. J. Butte, and E. Alsentzer. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Nature* 6, 17 (2024), E12–E22.