

NOTES AND INSIGHTS

Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPTMohammad S. Jalali, PhD^{a,b*}  and Ali Akhavan, PhD^a *Abstract*

The recent advent of artificial intelligence (AI) language tools like ChatGPT has opened up new opportunities in qualitative research. We revisited a previous project on obesity prevention interventions, where we developed a causal loop diagram through in-depth interview data analysis. Utilizing ChatGPT in our replication process, we compared its results against our original approach. We discuss that ChatGPT contributes to improved efficiency and unbiased data processing; however, it also reveals limitations in context understanding. Our findings suggest that AI language tools currently have great potential to serve as an augmentative tool rather than a replacement for the intricate analytical tasks performed by humans. With ongoing advancements, AI technologies may soon offer more substantial support in enhancing qualitative research capabilities, an area that deserves more investigation.

Copyright © 2024 The Authors. *System Dynamics Review* published by John Wiley & Sons Ltd on behalf of System Dynamics Society.

Syst. Dyn. Rev. (2024)

Introduction

Qualitative analysis is a crucial methodology for interpreting text data to extract meaningful patterns and insights from non-numeric data, often to emphasize understanding human behavior, experiences, and perceptions. The analysis of interview transcripts, a common form of qualitative data, involves a careful process of coding, aiming to distill complex, often nuanced narratives into comprehensible themes and categories (Burnard, 1991; Castleberry & Nolen, 2018; Leech & Onwuegbuzie, 2007). These methods have evolved with technological advancements, encompassing software tools—for example, ATLAS.ti, NVivo, and MAXQDA, which assist researchers in managing and interpreting textual data (Rädiker & Kuckartz, 2020; Woods *et al.*, 2016). The recent integration of AI language models in qualitative research opens new possibilities, enhancing data analysis efficiency and depth while potentially reducing human error and bias (De Paoli, 2023; Lee, 2024; Tai *et al.*, 2024).

Artificial intelligence (AI) language tools, such as ChatGPT and Google Bard, among other recent platforms, have become increasingly relevant in qualitative research. Platforms like ChatGPT are generative transformer models, part of the larger family of large language models, that utilize deep learning techniques and are adept at processing, understanding, and generating human language. Their capabilities extend from simple text generation to complex tasks such as

^a MGH Institute for Technology Assessment, Harvard Medical School, Boston, Massachusetts, USA

^b Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

* Correspondence to: Mohammad S. Jalali, MGH Institute for Technology Assessment, 125 Nashua St., Boston, MA 02114, USA.

E-mail: msjalali@mg.harvard.edu

Accepted by Andreas Größler, Received 2 February 2024; Revised 2 April 2024; Accepted 23 April 2024

sentiment analysis, thematic categorization, and contextual understanding (Xiao *et al.*, 2023), all of which are integral to qualitative research. Researchers have begun to explore the use of AI in various stages of qualitative analysis, from initial data coding to the extraction of nuanced insights. Studies highlight the potential of these tools not only to automate time-consuming tasks but also to bring a new level of depth to the analysis (Zhang *et al.*, 2023), thanks to their ability to process large datasets with consistency and lower cognitive biases that might affect researchers.

However, AI language tools are also prone to various biases because they are heavily trained on human-generated data, and there have been ample examples of such cases (Ashwin *et al.*, 2023; Ray, 2023). Recent advancements in AI models have mitigated such biases, fostering optimism for their application in research (Hagendorff *et al.*, 2023). Given that AI language models surpass human limitations in both data processing and recall, they emerge as powerful allies in the qualitative analysis process. For example, researchers have used ChatGPT as an assistant for idea generation and data identification in finance research (Dowling & Lucey, 2023). As another example, ChatGPT has been used as a virtual colleague for developing postgraduate courses (Meron & Araci, 2023).

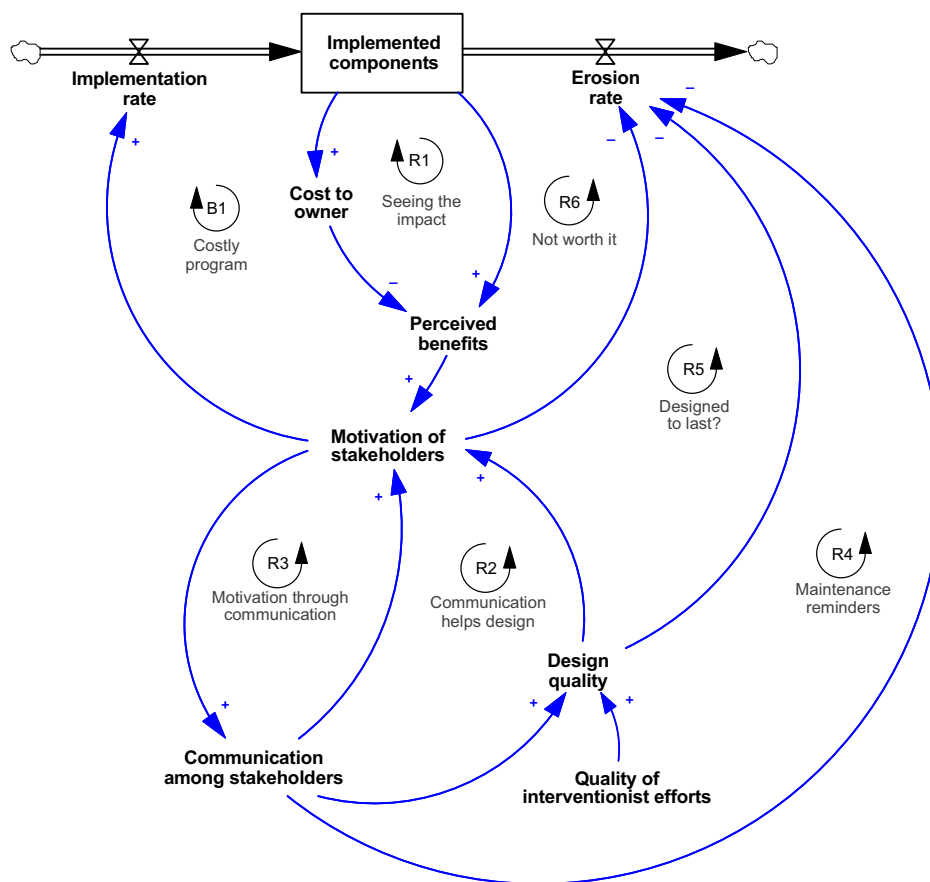
In specific research areas, the depth of qualitative text analysis extends beyond theme and pattern identification. For instance, in system dynamics, researchers engage in rigorous coding of textual data to discern model variables, causal links, and feedback loops (Kim & Andersen, 2012; Newberry & Carhart, 2023; Tomoia-Cotisel *et al.*, 2022). We revisited one of our prior studies, where interview data were analyzed to develop a causal loop diagram (CLD) (Jalali *et al.*, 2019). We aim to reassess that analysis, now employing ChatGPT, to draw comparisons and gauge its effectiveness.

Replication analysis

The original study (Jalali *et al.*, 2019) presented a CLD, as illustrated in Figure 1. The study reported the results of a detailed analysis of over 40 semi-structured interviews and focused on understanding the dynamics of adoption, implementation, and maintenance of obesity prevention interventions in various organizations, such as hospitals, daycares, and carry-out restaurants. It presented how small changes in intervention implementation can significantly affect the long-term success of the interventions. The research particularly emphasized the role of stakeholder communication and motivation in intervention sustainability, highlighting the impact of intervention design quality and resource allocation on the effectiveness of these public health interventions (Jalali *et al.*, 2014, 2017, 2019).

The process of the original study included coding interview transcripts to identify potential variables, links, and feedback loops. Here, we apply the same method, while adapting it to incorporate ChatGPT (particularly, version GPT-4). We used the interview transcripts from the original study and asked ChatGPT the following prompt:

Fig. 1. Causal loop diagram representing the dynamics of obesity prevention implementation, adapted from Jalali *et al.* (2019).



I want to create a causal loop diagram (CLD) from the text I provide. The text is the transcript of multiple interviews. In the first step, I just want to identify the key variables that can be used in the CLD. Could you please go through these interview data and extract the key variables of interest? Read the text thoroughly.

Next, we asked ChatGPT to identify the causal links between variables and propose as many feedback loops as possible using the following prompt:

This is good. Thank you. Now, in the next step, I want to identify the causal links between variables. Note that you can consider all the variables you identified above, and it is okay if the relationship between the two variables comes from different parts of the interviews. Also, note that you can always go back and read the interview data and find more variables if you think something is missing or you need more context and variables to establish causal relationships. Then, please identify and list the causal links between variables. After identifying the causal links, identify and list as many feedback loops as possible.

We repeated these prompts for all the interview data within the same chat session. Keeping the interview text analysis in the same sessions allows ChatGPT to

have a better understanding of the shared contextual information (Ashwin *et al.*, 2023). We repeated the whole process three times (in three separate chat sessions) and rotated the order of the interview documents that ChatGPT received.

Overall, ChatGPT identified 31 feedback loops, many of which overlap because of our multiple attempts in separate chat sessions as well as similar information from different interviews. To make the results from ChatGPT comparable with the CLD from the original study, we focused on content around Motivation of stakeholders. For example, mechanisms similar to feedback loop R1 in Figure 1 were identified by ChatGPT:

“As the hospital administration increases support for the wellness program, the effectiveness of the program improves, which could lead to further support from the administration.”

However, ChatGPT identified feedback loops that were not reported in the original study. Table 1 presents such feedback loops. For a visual representation, we drew loops based on ChatGPT's outcomes.

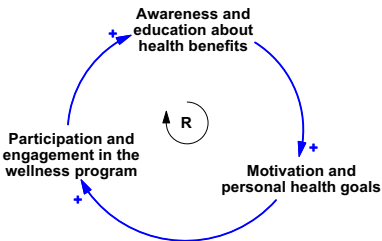
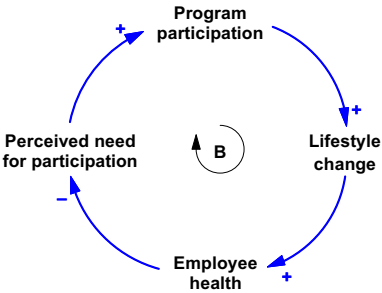
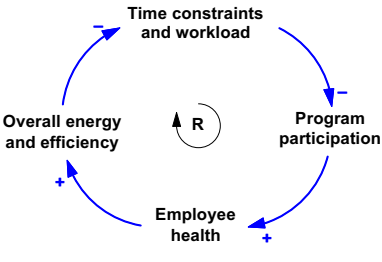
Table 1 includes examples of balancing and reinforcing feedback loops. One of the main differences between the loops identified by ChatGPT and the original study is stakeholder heterogeneity. Since the original study's focus is on high-level and aggregate concepts, its CLD considers only the managerial perspectives and their corresponding factors around motivation, such as costs to the owner. However, the additional loops identified by ChatGPT consider the employees and staff. In these feedback loops, the ‘program participation’ by employees is one of the variables that ChatGPT identifies but was not included in the original study. This detailed consideration enriches the analysis, offering a more nuanced view of the dynamics at play, beyond the managerial focus of the original model. Although the original study's model is versatile enough to potentially include variables like program participation, ChatGPT's explicit recognition of such factors presents a more detailed reflection of the data.

We also observed that ChatGPT does not capture all the nuances shown in Figure 1. For example, ChatGPT does not identify the erosion of implemented components and its related feedback loops affecting the maintenance of the intervention. Additionally, ChatGPT analyzes the relationships between variables independent of loop descriptions, and it does not trace the causal chains in the same direction. For example, in the last loop of Table 1, while ChatGPT notes that an increase in time constraints and workload would decrease program participation, it reverses the direction of variable change by saying that improvements in employee health would lead to higher overall energy and efficiency—in contrast, it should have been: increase in time constraints and workload would lead to lower program participation, and lower program participation would lead to lower employee health.

Discussion

AI language tools introduce notable advancements in qualitative research. In our analysis, ChatGPT's ability to identify feedback loops not seen in the original study highlights its capacity for direct interpretation of data. On the other hand,

Table 1. Identified feedback loops based on interview data by ChatGPT.

Feedback loops identified by ChatGPT	Feedback loops drawn by authors based on ChatGPT's outputs
<p>“Increased awareness and education about health benefits lead to higher motivation and personal health goals, which in turn encourage further participation and engagement in the wellness program, leading to more awareness.”</p>	
<p>“As employee health improves due to lifestyle changes and program participation, there might be a decrease in the perceived need for intense participation, balancing out the engagement levels.”</p>	
<p>“Time constraints and workload impact program participation negatively, but as health improves through participation, there might be an increase in overall energy and efficiency, potentially easing time constraints.”^a</p>	

^aNote the directional change in the elaboration of this feedback loop. We discuss this below.

researchers’ findings are often shaped by their understanding, intuition, mental models, and the influence of their literature of interest. In our replication exercise, the feedback loops identified by the original authors, while aligned with their research focus, may include subjective elements. The authors interpreted the data from an organizational perspective for their CLD. However, ChatGPT’s objective and straightforward analysis added a layer of purity to the research process. This direct approach potentially reduces the risk of introducing biases that may arise from the researchers’ own mental models during interpretation. This

10991727, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sdr.1772 by Massachusetts General Hospital, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

suggests a future approach where researchers may still apply their analytical perspectives but also include a more direct analysis of interview data as provided by tools like ChatGPT. This dual approach could offer a more thorough and balanced understanding of the data. It could also help improve transparency in qualitative research and CLD reporting—an area that needs major enhancements (Jalali & Beaulieu, 2023).

Despite the benefits gained through AI language tools, their application is not without challenges. Unlike human researchers, these tools may lack the capacity for nuanced understanding and integration of data with broader academic discourses and theories. One can argue that AI language tools are also aware of such broad knowledge; however, while they can access a broad spectrum of literature, leveraging this knowledge effectively for nuanced analysis requires extensive training, experimentation, and navigation through inherent ambiguities. This process can be hampered by a lack of transparency, making it challenging to achieve the level of contextual and theoretical integration often seen in human-led research.

Ethical and data ownership concerns also arise, especially when sharing human-based data (e.g., interview transcripts) with these privately owned platforms. While there are compliance claims with legal frameworks such as the General Data Protection Regulation, there is no clear guideline on the use of generative models for handling sensitive research data, which could include identifiable personal information, personal stories, or proprietary data (Wu *et al.*, 2024). Some institutions and universities have recently initiated their in-house AI language tools to address the privacy and ownership challenges.

Additionally, one should note that AI language tools are frequently updated; thus, mentioning the specific version used is crucial. Each version may function differently, affecting the study's reproducibility. However, even a model like GPT-4 has different sub-versions, which may include updates in training data, fine-tunings, or variations tailored for specific tasks or performance improvements. Such variations can cause different responses (Bender *et al.*, 2021; Holtzman *et al.*, 2020). Importantly, the stochastic nature of these tools can result in scenarios where the identical version of a tool, prompted in the same manner, still fails to yield consistent output across replications. There have been attempts to increase the replicability of outcomes (e.g., by controlling the model's settings; Davis *et al.*, 2024), yet there is need for more research and investigation on reproducibility.

Another challenge includes potential biases in AI models due to their training data, which could affect analysis results. In our brief assessment, we limited ChatGPT to analyzing only provided interview transcripts, but the influence of its underlying training remains an area of ambiguity. Overall, the black box nature of these AI tools raises concerns about the objectivity of outcomes and provokes questions as to researchers' reliance on such tools without proper knowledge of their training and development phases. These tools are trained on a vast amount of data, encompassing a wide variety of internet text. However, they may only absorb hegemonic worldviews from their training data that may be biased, inaccurate, or irrelevant to the research context (e.g., interviewees), leading to various racial, gender, and socio-political framing biases (see, e.g., Holtzman *et al.*, 2020; Sap *et al.*, 2020). Therefore, researchers should avoid being overly reliant on the

outcomes of these tools. Such concerns are not limited to the applications of AI language tools, as we have utilized in our analysis, and can be observed in other areas where researchers apply an advanced AI algorithm without having a full understanding of its limitations (e.g., see Obermeyer *et al.*, 2019, for the case of racial disparities in health outcomes as a result of using commercial prediction algorithms). Exploring the examples of challenges and biases, as mentioned above, is beyond the scope of this paper, but it has been explored and discussed elsewhere (e.g., Ray, 2023). Further research is required to systematically explore the limitations and challenges imposed by using AI language tools that raise concerns about objectivity, especially in the context of qualitative research in system dynamics.

Looking ahead, the potential developments in AI appear promising in supporting qualitative research despite its challenges. Future innovations may improve AI's capability to understand the context more deeply, handle subtleties in data with greater precision, minimize biases, and improve reproducibility. We encourage researchers to test and compare the capabilities of current and future versions of various AI language tools, while being cautious of their limitations.

Finally, as we recently discussed elsewhere (Akhavan & Jalali, 2023), it is important to consider that AI is not to replace the critical analytical thinking inherent to human researchers. Relying on the outcomes of these tools without understanding their limitations and without cross-checking the results can present risks to the quality of research. Instead, AI's role would be that of an assistant, augmenting human capabilities to enhance the efficiency, thoroughness, and depth of qualitative research.

Funding information

MSJ was supported in part by grant R01CE003358 from the US Centers for Disease Control and Prevention.

References

- Akhavan A, Jalali MS. 2023. Generative AI and simulation modeling: how should you (not) use large language models like ChatGPT. *System Dynamics Review*. <https://doi.org/10.1002/sdr.1773>.
- Ashwin J, Chhabra A, Rao V. 2023. *Using large language models for qualitative analysis can introduce serious bias*. <http://arxiv.org/abs/2309.17147>
- Bender EM, Gebru T, McMillan-Major A, Shmitchell S. 2021. On the dangers of stochastic parrots: can language models Be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. Association for Computing Machinery: New York, NY, USA; 610–623.
- Burnard P. 1991. A method of analysing interview transcripts in qualitative research. *Nurse Education Today* 11(6): 461–466. [https://doi.org/10.1016/0260-6917\(91\)90009-Y](https://doi.org/10.1016/0260-6917(91)90009-Y).
- Castleberry A, Nolen A. 2018. Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning* 10(6): 807–815. <https://doi.org/10.1016/j.cptl.2018.03.019>.

- Davis J, Bulck LV, Durieux BN, Lindvall C. 2024. The temperature feature of ChatGPT: Modifying creativity for clinical research. *JMIR Human Factors* **11**(1): e53559. <https://doi.org/10.2196/53559>.
- De Paoli S. 2023. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*. <https://doi.org/10.1177/08944393231220483>.
- Dowling M, Lucey B. 2023. ChatGPT for (finance) research: the Bananarama conjecture. *Finance Research Letters* **53**: 103662. <https://doi.org/10.1016/j.frl.2023.103662>.
- Hagendorff T, Fabi S, Kosinski M. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* **3**(10): 833–838. <https://doi.org/10.1038/s43588-023-00527-x>.
- Holtzman A, Buys J, Du L, Forbes M, Choi Y. 2020. *The curious case of neural text degeneration*. <http://arxiv.org/abs/1904.09751>.
- Jalali M, Rahmandad H, Bullock S, Ammerman A. 2017. Dynamics of implementation and maintenance of organizational health interventions. *International Journal of Environmental Research and Public Health* **14**(8): 917. <https://doi.org/10.3390/ijerph14080917>.
- Jalali MS, Beaulieu E. 2023. Strengthening a weak link: transparency of causal loop diagrams—current state and recommendations. *System Dynamics Review*. <https://doi.org/10.1002/sdr.1753>.
- Jalali MS, Rahmandad H, Bullock S, Ammerman A. 2014. Dynamics of obesity interventions inside organizations. In *The 32nd International Conference of the System Dynamics Society*, Delft, Netherlands.
- Jalali MS, Rahmandad H, Bullock SL, Lee-Kwan SH, Gittelsohn J, Ammerman A. 2019. Dynamics of intervention adoption, implementation, and maintenance inside organizations: the case of an obesity prevention initiative. *Social Science & Medicine* **224**: 67–76. <https://doi.org/10.1016/j.socscimed.2018.12.021>.
- Kim H, Andersen DF. 2012. Building confidence in causal maps generated from purposive text data: mapping transcripts of the Federal Reserve. *System Dynamics Review* **28**(4): 311–328. <https://doi.org/10.1002/sdr.1480>.
- Lee EA. 2024. Deep neural networks, explanations, and rationality. In *Bridging the Gap between AI and Reality*, Steffen B (ed). Springer Nature Switzerland: Cham; 11–21.
- Leech NL, Onwuegbuzie AJ. 2007. An array of qualitative data analysis tools: a call for data analysis triangulation. *School Psychology Quarterly* **22**(4): 557–584. <https://doi.org/10.1037/1045-3830.22.4.557>.
- Meron Y, Araci YT. 2023. Artificial intelligence in design education: Evaluating ChatGPT as a virtual colleague for post-graduate course development. *Design Science* **9**: e30. <https://doi.org/10.1017/dsj.2023.28>.
- Newberry P, Carhart N. 2023. Constructing causal loop diagrams from large interview data sets. *System Dynamics Review* **40**(1): 1745. <https://doi.org/10.1002/sdr.1745>.
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**(6464): 447–453. <https://doi.org/10.1126/science.aax2342>.
- Rädiker S, Kuckartz U. 2020. *Focused analysis of qualitative interviews with MAXQDA*. MAXQDA Press: DE.
- Ray PP. 2023. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* **3**: 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>.
- Sap M, Gabriel S, Qin L, Jurafsky D, Smith NA, Choi Y. 2020. Social bias frames: reasoning about social and power implications of language.
- Tai RH, Bentley LR, Xia X, Sitt JM, Fankhauser SC, Chicas-Mosier AM, Monteith BG. 2024. An examination of the use of large language models to aid analysis of textual data.

-
- International Journal of Qualitative Methods*, 23. <https://doi.org/10.1177/16094069241231168>.
- Tomoaia-Cotisel A, Allen SD, Kim H, Andersen D, Chalabi Z. 2022. Rigorously interpreted quotation analysis for evaluating causal loop diagrams in late-stage conceptualization. *System Dynamics Review* 38(1): 41–80. <https://doi.org/10.1002/sdr.1701>.
- Woods M, Paulus T, Atkins DP, Macklin R. 2016. Advancing qualitative research using qualitative data analysis software (QDAS)? reviewing potential versus practice in published studies using ATLAS.ti and NVivo, 1994–2013. *Social Science Computer Review* 34(5): 597–617. <https://doi.org/10.1177/0894439315596311>.
- Wu X, Duan R, Ni J. 2024. Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence* 2(2): 102–115. <https://doi.org/10.1016/j.jiixd.2023.10.007>.
- Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer P-Y. 2023. Supporting qualitative analysis with large language models: combining codebook with GPT-3 for deductive coding. In *28th International Conference on Intelligent User Interfaces*. ACM: Sydney NSW Australia; 75–78.
- Zhang H, Wu C, Xie J, Kim C, Carroll JM 2023. QualiGPT: GPT as an easy-to-use tool for qualitative coding.