Keeping Users Engaged During Repeated Interviews by a Virtual Agent: Using Large Language Models to Reliably Diversify Questions

Hye Sun Yun
yun.hy@northeastern.edu
Northeastern University
Boston, MA, USA

Mehdi Arjmand
arjmand.me@northeastern.edu
Northeastern University
Boston, MA, USA

Phillip Sherlock
phillip.sherlock@ufl.edu
University of Florida
Gainesville, FL, USA

Michael K. Paasche-Orlow
mpo@tufts.edu
Tufts University
Boston, MA, USA

James W. Griffith
jamesgriffith@uchicago.edu
University of Chicago
Chicago, IL, USA

Timothy Bickmore
t.bickmore@northeastern.edu
Northeastern University
Boston, MA, USA

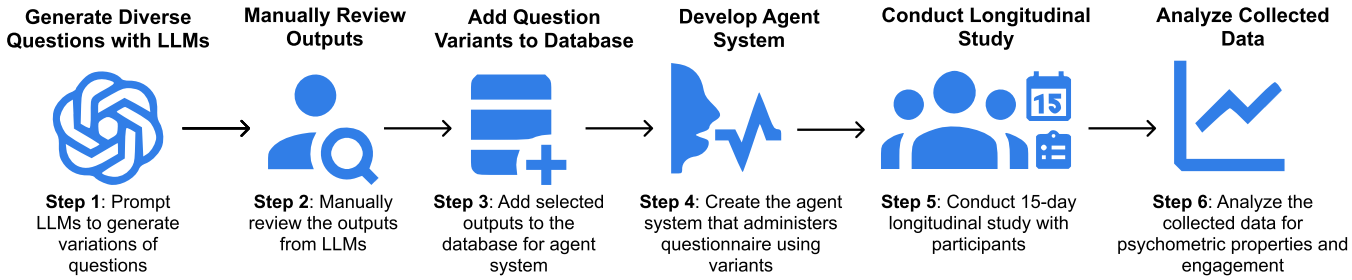

Figure 1: A workflow diagram of the longitudinal validation study which evaluated the validity, reliability, and user engagement of utilizing large language model-generated variants of a standardized depression questionnaire.

ABSTRACT

Standardized, validated questionnaires are vital tools in research and healthcare, offering dependable self-report data. Prior work has revealed that virtual agent-administered questionnaires are almost equivalent to self-administered ones in an electronic form. Despite being an engaging method, repeated use of virtual agent-administered questionnaires in longitudinal or pre-post studies can induce respondent fatigue, impacting data quality via response biases and decreased response rates. We propose using large language models (LLMs) to generate diverse questionnaire versions while retaining good psychometric properties. In a longitudinal study, participants interacted with our agent system and responded daily for two weeks to one of the following questionnaires: a standardized depression questionnaire, question variants generated by LLMs, or question variants accompanied by LLM-generated small talk. The responses were compared to a validated depression questionnaire. Psychometric testing revealed consistent covariation between the external criterion and focal measure administered across the three conditions, demonstrating the reliability and validity of the LLM-generated variants. Participants found that the variants were significantly less repetitive than repeated administrations of the same standardized questionnaire. Our findings highlight the potential of LLM-generated variants to invigorate agent-administered questionnaires and foster engagement and interest, without compromising their validity.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Natural language generation*; **Intelligent agents**.

KEYWORDS

questionnaires, engagement, large language models, virtual agents, health, longitudinal research

IVA '24, September 16–19, 2024, GLASGOW, United Kingdom

ACM ISBN 979-8-4007-0625-7/24/09
https://doi.org/10.1145/3652988.3673929

ACM Reference Format:
Hye Sun Yun, Mehdi Arjmand, Phillip Sherlock, Michael K. Paasche-Orlow, James W. Griffith, and Timothy Bickmore. 2024. Keeping Users Engaged During Repeated Interviews by a Virtual Agent: Using Large Language Models to Reliably Diversify Questions. In *ACM International Conference on Intelligent Virtual Agents (IVA '24), September 16–19, 2024, GLASGOW, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3652988.3673929

1 INTRODUCTION

Self-report questionnaires are a type of self-report method that includes a set of questions in a highly structured, standardized written form. Validated questionnaires are widely used in research and healthcare as an assessment strategy as they offer dependable self-report data. Prior work has revealed that human and virtual agent-administered questionnaires are nearly equivalent to self-administered questionnaires in the electronic form [7, 31]. These studies have shown the feasibility and reliability of using virtual agents (VAs) to administer questionnaires simulating interviews for a single session. However, many repeated-measures evaluation studies and longitudinal interventions require the same self-report questionnaire to be administered to the same individual multiple times. In healthcare, patient-reported outcomes (PROs) are used to obtain self-reports of a patient's condition at home, typically involving the repeated administration of surveys to capture symptoms or quality of life [23].

However, response rates to repeated surveys tend to decline over time, as respondents become fatigued by repeatedly filling out the same questionnaires [20, 44, 50]. Even in healthcare, where PROs can be used as the basis for treatment decisions, longitudinal survey completion rates can be as low as 48% [20, 30, 44]. Dwindling response rates can lead to a nonresponse measurement bias [25] and limit the ability to evaluate important changes over time.

Several strategies have been proposed to increase repeated measure response rates, including incentives [62], more frequent contact and engagement with respondents [15, 51], and providing survey responses back to the individuals being surveyed [65]. In addition, a variety of approaches have been studied to increase the usage rates for repeated interactions with VAs, including the use of syntactic and visual variability in the interface [8] and humor [47]. Other strategies for automated systems include reminders [29] and social support and reinforcement [60].

In this work, we explore two strategies to increase response rates to a PRO administered daily for two weeks by a virtual agent (VA) that simulates a face-to-face interview with a healthcare professional. The first strategy involved using survey questions that vary in every administration so that the survey administrations sound different. We went beyond straightforward syntactic variation of questions to variants generated by large language models (LLMs) to capture the latent construct we are interested in. Syntactic variations primarily entail the reordering of words within a sentence, whereas LLM-generated variations we employed had slightly different words or phrases that convey comparable meanings. Second, we explore the use of small talk, humor, and empathy generated by LLMs to make daily interactions with the VA more conversational, entertaining, and engaging.

Many questionnaires, including most PROs, have been validated using laborious methods involving testing with dozens, if not thousands, of respondents to establish reliability and validity [24]. An important question raised when validated questionnaires are modified is whether the new derivative versions retain the reliability and validity of the original form. We report the results of a longitudinal study involving a validated PRO for depression, in which participants engaged with our virtual agent system daily for two weeks. Participants were randomized to either a repeated, standardized depression questionnaire or one of the two interventions with LLM-generated questionnaire variants. All participants completed an additional standardized, validated depression questionnaire which was a criterion for comparison. Our hypotheses are:

- **H1**: VA administration of LLM-generated questionnaire variants will retain similar validity and reliability to the VA administration of the original questionnaire.
- **H2**: Questionnaires delivered in a different form using LLM-generated variants daily will be more engaging for participants, based on the number of questionnaires completed and feedback from participants.
- **H3**: Questionnaires delivered with LLM-generated conversational small talk, humor, and empathy will be more engaging compared to those delivered as strictly question-and-response interviews by a VA.

Our primary contributions include the introduction of an innovative approach using LLMs to reliably generate diverse versions of validated questionnaires for a VA system. Furthermore, we demonstrate the feasibility of employing these questionnaire variants to enhance user engagement and mitigate repetitiveness.

2 RELATED WORK

Our work draws on previous research on alternative delivery methods for questionnaires, user engagement methods for longitudinal research, and applications of LLMs to agents and surveys. Traditionally, paper or mail surveys have been used to administer questionnaires. However, web-based or online surveys have been more frequently employed, as response rates to paper surveys have declined over time [19, 22, 39, 54, 58]. However, web-based surveys are not immune to dwindling response rates [42, 64]. Particularly for long or repetitive administrations of surveys in research, respondents can experience fatigue, which can lower the completion rates and data quality of the responses [10, 38, 50, 59]. Increasing the quality of self-report data and completion rates through surveys remains an important challenge for researchers to overcome.

2.1 Computers & Agents for Quality Self-Report

Several past studies have shown that using computers to administer surveys and interviews can lead to greater self-disclosure, especially for sensitive information in the context of healthcare, as the pressure to respond in socially desirable ways is reduced [21, 34, 40, 46, 61, 68]. Several studies have expanded this approach by incorporating VAs, as research has indicated that using conversational interviews for surveys can effectively reduce errors. [55]. In health-screening interviews with VAs, Lucas et al. [40] discovered that individuals who perceived the VA as automated showed reduced fear of self-disclosure and expressed sadness more intensely than those who perceived the VA as human-operated. A similar study by Schuetzler et al. [56] demonstrated that people disclose more about their sensitive behavior to a conversational agent than to a human; however, people disclose less when the conversational agent appears to understand. In addition, Kocielnik [36] used a chatbot called HarborBot to test a conversational approach for social needs screening in emergency departments and compared it with a traditional survey tool. The results revealed that the conversational approach was perceived to be more engaging,

caring, understandable, and accessible among low health literacy users than the traditional approach.

Prior work has demonstrated that medical questionnaires or PROs administered by VAs are valid and statistically equivalent to human or self-administered questionnaires [4, 7]. For example, Jaiswal et al. [31] conducted two sets of studies using mental health questionnaires: one comparing VA administration to standard self-administration, and the second comparing VA administration to an actual human. The results showed that the questionnaires administered by the VAs were statistically equivalent to human or self-administered questionnaires. Additionally, Mancone et al. [41] showed that voice assistants, such as Alexa, can be used to administer psychological assessment questionnaires as a powerful way to capture attention and engage users emotionally without compromising validity.

With LLMs becoming increasingly capable and prevalent, researchers have begun investigating how LLMs can be employed for self-reporting and survey research. Jansen et al. [32] highlighted how LLMs can help overcome some of the challenges in survey research by generating responses to survey items or question-wording. Despite this promising direction, the authors warned about the risks of harmful and inaccurate outputs when using LLMs. Similarly, Kjell et al. [35] provided a narrative review of how LLMs can potentially be used for psychological assessments using natural language instead of rating scales. Furthermore, one study employed GPT-3 to power a chatbot to collect self-reported data, such as food intake, exercise, sleep, and work productivity [66]. The authors found that LLMs also provided the ability to maintain context, state tracking, and provide off-topic suggestions.

2.2 Maintaining Longitudinal Self-Report with Agents

Longitudinal studies that require multiple self-reports often have low completion rates [1]. Although VAs increase engagement when administering questionnaires, they still suffer from user disengagement. The length of the first interaction with a VA has been shown to be the primary predictor of the number of healthcare questionnaires completed by a participant [63]. The findings showed that longer first interactions can result in fewer completed questionnaires. Prior work on maintaining engagement in long-term health interventions with VAs by Bickmore et al. [8] shows that increased variability in agent behavior and giving the agent a human backstory can also lead to increased engagement.

2.3 LLMs for Agents

Recently, several studies have explored the use of LLMs in agent design and implementation. Antunes et al. [2] prompted LLMs to assist in creating scenarios for socially intelligent agents often used in education and entertainment. They created a pipeline to generate an agent's beliefs, desires, intentions, plans of action, and emotions. Furthermore, other studies have examined how LLMs can generate dialogue utterances for embodied conversational agents such as social robots [28, 57] to mitigate boredom and increase engagement. Olafsson et al. [48] explored how LLMs can be used as part of VAs for health applications by incorporating GPT-2 in a hybrid dialog system for a virtual alcohol misuse counselor. GPT-2 generated responses were combined with a rule-based approach to transition through structured counseling sessions. Similarly, our study takes a rule-based approach but incorporates diverse messages generated by LLMs. Due to the sensitive nature of mental health questionnaires, we decided on a human-in-the-loop approach due to LLMs' potential harms in the healthcare context [9, 26, 70]

Figure 2: A screenshot of the agent waiting for the user to respond after asking a depression questionnaire question. The dialogue response options are displayed at the top right corner of the screen.

2.4 LLMs for Generating Diverse Text

In addition to utilizing LLMs for agents, LLMs have been used to generate diverse texts or paraphrases [69]. Cegin et al. [12] conducted a study comparing the quality of crowd-sourced and LLM-generated paraphrases for their diversity and robustness in intent classification. The authors found that ChatGPT is a viable alternative to human paraphrasing. Furthermore, one study showed that while GPT-4 might not necessarily outperform humans in generating diverse motivational messages, it took only 6 seconds to generate one message compared to an average of 73 seconds for humans [17]. Although LLMs may not always provide the most diverse generated text, they are significantly faster and more grammatically correct than humans. To mitigate the challenges of LLMs, Pehlivanoğlu et al. [49] demonstrated how prompt engineering can enhance lexical diversity, phrasal variations, fluency, relevance, and syntactical differences while preserving the original meaning.

3 SYSTEM DESIGN

To evaluate our study hypotheses, we created a VA system deployed over the web for participants to interact daily (Figure 2). Our agent is a 3D animated character that converses with users using synthetic speech, conversational behavior, and multiple-choice menu inputs for user responses. The agent's synchronized nonverbal conversational behavior, such as hand gestures, head nods, eyebrow raises, and posture shifts, was automatically generated using the Behavior Expression Animation Toolkit [11]. Agent utterances were generated using template-based text generation. The agent's dialogue is driven by a hierarchical task network-based dialogue engine. The

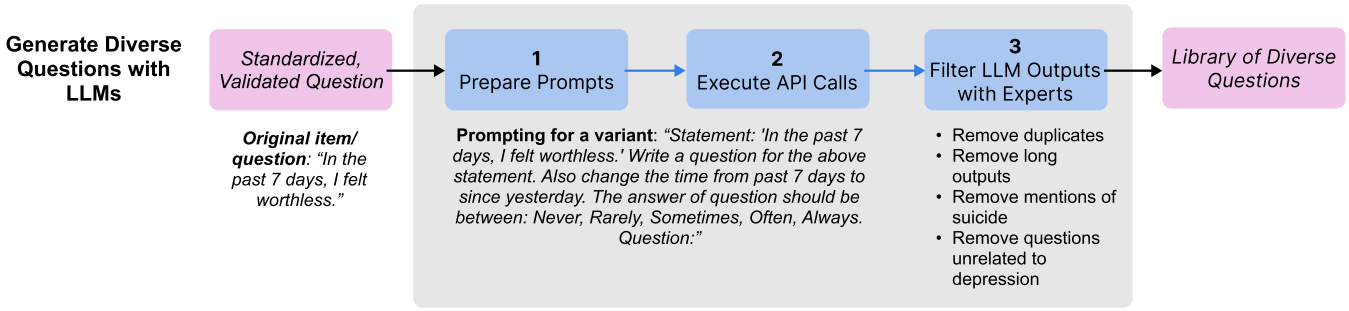

Figure 3: A workflow diagram of how LLMs were used to generate diverse questions. A simple example is provided.

VA system was implemented using the Unity3D game engine and CereProc speech synthesizer.

Our agent, Marie, interacts with participants daily by verbally administering an eight-question questionnaire in dialogue. For our prototype, we focused on one self-report PRO questionnaire using the eight-item PROMIS® short form depression questionnaire (version 8a) [13]. This questionnaire was developed to assess a respondent's level of emotional distress caused by depressed mood where each statement is rated on a five-point scale from 1 being "Never" to 5 being "Always."

3.1 Generating Question Variants Using LLMs

By using LLMs, we can yield a greater range of variations for each question faster and with fewer human resources, potentially increasing user engagement [17]. However, it is important to note that such variations may include harmful language or language that deviates from the main concept of the original questionnaire items. Particularly in the domain of mental health, unconstrained outputs from LLMs may not be suitable for measuring various aspects of mental health issues, as LLMs are known to provide dangerous advice or misinformation [9, 45, 48, 52].

To address the potential issue of harmful outputs (hallucinations or misinformation) from LLMs, we used ChatGPT (March 2023 version) and GPT-3 to generate different variants of each question and manually filtered them before using them in our VA system (Figure 3). For prompting, we provided the original item and response scale and asked the LLM to paraphrase the main item into a new question. These models were prompted to generate variants that fit the response scale. These variations often contained words, phrases, or concepts closely related to the latent construct we wanted to measure. A total of 178 unique variants of the eight questions were generated. A psychologist then ranked and filtered the variants to create a final list that matched the meaning and purpose of the original question. We also removed potentially dangerous questions that might involve suicide or thoughts about killing oneself. We obtained 67 variants of the questions, with a few samples available in Table 1. All implementation details of generating the item variants and conversational small talk including jokes and empathetic responses and the full list of LLM-generated content used for the study can be found in supplementary materials[1].

[1] https://github.com/hyesunyun/va-item-variants

4 LONGITUDINAL EVALUATION STUDY

From April to May 2023, we conducted a 15-day longitudinal online study to evaluate the psychometric properties of our questionnaire variants, user engagement, and user perception of the VA questionnaire system. For the first 14 days, participants were asked to talk with the VA once per day, which lasted a few minutes, and answer a short online survey after each interaction. On the 15th day, participants completed a final online survey. Figure 1 provides the entire study workflow.

The experiment followed a randomized between-subject design with three study conditions. In one condition — CONTROL, we had the VA administer the standardized eight-item PROMIS® depression questionnaire in question format for a more conversational experience. In the two intervention conditions — ITEM VARIANTS ONLY and ITEM VARIANTS PLUS, the agent randomly chooses question variants described in subsection 3.1. In addition, ITEM VARIANTS PLUS includes additional social and conversational content for dialogue, such as anecdotes, jokes, empathetic responses, inspiring or hopeful messages, and farewells generated by LLMs (Table 2). In a typical session for the ITEM VARIANTS PLUS condition, the agent first shared a short personal anecdote and then a randomly selected joke before administering the questionnaire. The agent provides empathetic responses based on user responses to the questions and ends with a randomly chosen motivational message and farewell statements.

4.1 Measures & Data Collection

We collected the following four items with a 7-point scale response after each interaction with the agent: "How satisfied are you with the agent?" (1="not at all" and 7="very satisfied"), "How much would you like to continue talking with the agent?" (1="not at all" and 7="very much"), "How natural was your conversation with the agent?" (1="not at all" and 7="very natural"), and "Did the agent feel repetitive?" (1="not at all" and 7="very repetitive"). User engagement is assessed as the number of completed interactions with the agent.

After the two-week study period, we administered the final survey on the 15th day, which included the eight-item Patient Health Questionnaire depression scale (PHQ-8) [37, 53], the system usability scale (SUS) [5], overall system satisfaction measures, agent satisfaction measures, and measures related to user perception of

Table 1: Depression Questionnaire with Sample Item Variants Generated by LLMs. This table provides the wording variations of the eight-item PROMIS® short form depression questionnaire [13]. All items are rated on a five-point scale from 1="Never" to 5="Always". *# of Variants* refers to the number of variants we used for the longitudinal evaluation study.

ID	Original	Control	Sample Variants	# of Variants
1	In the past 7 days, I felt worthless.	In the past 7 days, how often have you felt worthless?	Since we last spoke, have you ever felt like you were a burden to others?; Have you felt like you were not good enough recently?; Since the last time we talked, have you felt like you're not important to anyone?	8
2	In the past 7 days, I felt helpless.	In the past 7 days, how often have you felt helpless?	How often have you felt like you were unable to control a situation in the past day?; How often do you feel like you're stuck in a cycle of negativity when faced with challenges?; Have you felt powerless or helpless when dealing with a problem in the past day? How often?	7
3	In the past 7 days, I felt depressed.	In the past 7 days, how often have you felt depressed?	Have you been feeling like you can't escape negative thoughts or feelings?; How often have you been feeling empty or numb?; Have you been experiencing changes in your sleep patterns?	8
4	In the past 7 days, I felt hopeless.	In the past 7 days, how often have you felt hopeless?	How frequently have you felt like you're drowning in negativity since the last time we talked?; Have you ever felt like everything is pointless, even if things are going well? If so, how often?; Have you felt like you're stuck in a rut or in a situation that's beyond your control? If so, how often?	7
5	In the past 7 days, I felt like a failure.	In the past 7 days, how often have you felt like a failure?	How often do you feel like you're not making the most of your talents and abilities?; How often do you feel like you're not contributing enough to society?; How often do you feel like you've fallen short of your own expectations?	8
6	In the past 7 days, I felt unhappy.	In the past 7 days, how often have you felt unhappy?	How often do you experience feelings of unhappiness?; Do you tend to dwell on negative thoughts and feelings?; Have you noticed yourself feeling unhappy more frequently than usual?	7
7	In the past 7 days, I felt that I had nothing to look forward to.	In the past 7 days, how often have you felt that you had nothing to look forward to?	Do you frequently feel like your life lacks purpose or direction?; How often do you feel like there's nothing to look forward to in the coming days or weeks?; Have you been struggling to find joy in your daily activities?	11
8	In the past 7 days, I felt that nothing could cheer me up.	In the past 7 days, how often have you felt that nothing could cheer you up?	Do you rarely feel happy or uplifted when you're feeling low?; Do you ever feel like you just can't shake off a negative mood?; Have you found it hard to see the positive side of things lately?	11

questions asked by the agent (Table 3). We also asked four open-ended questions about their experiences.

4.2 Participants

Participants were recruited via an online research platform (www.prolific.com). They were required to be 18 years old or older, able to read and write English, located in the USA, have working audio for their computer, and have a browser that supports WebGL 2.0. Participants were told to interact daily with the system at least seven times during the two weeks and complete a final survey for compensation. Each interaction consisted of a conversation with the VA and a short survey. The minimum interaction requirement was to ensure each participant was provided with the questionnaire several times. Due to the sensitive nature of asking about depression symptoms, participants were told that the system is for assessment only and were provided with a list of mental health resources in the USA. The study was approved by our institutional review board.

5 RESULTS

A total of 105 participants began the longitudinal evaluation study, with 35 participants assigned to each study condition. In total, 93 participants met the compensation requirements and completed the study successfully. All participants were on average 39 (*SD*=12, *Mdn*=37, *Range*=21~73) years old. The gender breakdown was 49.5% women, 46.7% men, 2.9% non-binary, and 1.0% others. Participants were 75.2% white, 7.6% multiracial, 6.7% Black/African-American, 3.8% Asian/Asian American, 2.9% Hispanic/Latinx, 1.9 % American Indian/Alaska Native, and 1.9% other. Participants had at least a high school degree or equivalent (43.8% with a bachelor's degree, 24.8% with some college, 10.5% with a master's degree, 10.5% with an associate degree, and 1.9% with a doctoral/professional degree). When asked if they are currently in therapy or taking medication for depression, 80.0% of participants said "no", 19.1% said "yes", and 1.0% preferred not to answer.

5.1 Psychometric Properties of LLM-generated Item Variants

We calculated Cronbach's alpha [18] to measure the internal consistency or reliability of the eight depression questions, or how closely related the eight questions are as a group. This involved looking at the responses of each participant for each administration. Cronbach's alpha for the CONTROL condition was α=.76 whereas the item variants were α=.65. Although α for the version with variants was lower than that of the CONTROL, it showed acceptable internal consistency. Supplementary materials provide the percentages of responses for each question.

Table 2: Examples of Conversational Content Generated by LLMs.

Category	Example Content	Count
Personal Anecdotes	• I love going for hikes in the beautiful outdoors! This morning, I took a hike around a nearby lake. The fresh air and peaceful atmosphere made it the perfect way to start the day! • I just finished reading this amazing book I stumbled upon! I couldn't put it down. It was a captivating journey that kept me on the edge of my seat and I can't wait to recommend it to all my friends. • This past weekend I decided to try a new restaurant in town. The atmosphere was cozy and the food was delicious! I'm already looking forward to my next visit so I can try something else off the menu.	37
Jokes	• Why don't scientists trust atoms? Because they make up everything! • Why did the smartphone need glasses? Because it lost all its contacts! • What do you call a bear with no teeth? A gummy bear!	24
Empathetic Responses	• I understand how overwhelming helplessness can be, and I'm here to support you. • I'm sorry to hear that you feel this way. Please remember that you are valuable and that your feelings are valid. • I understand how you're feeling. It's normal to feel overwhelmed at times and it's ok to take a step back and take care of yourself.	35
Inspiring or Hopeful Messages	• You are not alone in your struggles. Reach out to others for support and comfort. • Shiv Khera once said, Your positive action combined with positive thinking results in success. • Today is your day to shine! Believe in yourself and make it happen.	23
Farewells or Ending Conversations	• Well, I should get going. It was nice talking to you! • It was great catching up with you. I hope we can chat again soon! • I enjoyed our conversation. It was nice talking with you. Have a great day!	42

We also investigated whether the psychometric properties of items (questions) were consistent across the three groups. To investigate the validity of the PROMIS® depression questionnaire across the three study conditions, we conducted a measurement alignment analysis using the `sirt` package in `R`, which allows the assessment of how the properties of individual items differ across groups [27]. In this model, each item is related to a single latent factor (depression) by a linear relationship described by a slope (i.e., factor loading) and intercept parameter. This analysis examined the degree to which groups can be "aligned" on the same scale. First, a confirmatory factor analysis (CFA) model allowed item slopes and intercepts to vary across groups. An R^2 was used to express how much variance in group differences was captured by true mean differences in the groups, rather than by different item properties across groups [3]. Our results, using the alignment procedure, indicated that 99% of the between-group variation associated with slopes and 98% of the between-group variation associated with intercepts could be attributed to factor mean and variance differences across the groups. Thus, the properties of the items were very consistent across groups. In a simulation, Asparouhov and Muthen [3] found that R^2 values of at least .98 were required to procure reliable factor rankings and that in general, R^2 values greater than .75 (i.e., up to 25% non-invariance) were needed to produce trustworthy alignment results. Therefore, based on having achieved R^2 values greater than .98 for both the aligned item intercepts and loadings, we demonstrated the consistency of the PROMIS® questionnaire administered across the three study conditions and concluded that only 2% and 1% of the variance could be attributed to differences in the item slopes and intercepts across the three study conditions.

To test the validity of each of the three administrations of the PROMIS® questionnaire, we compared the three administrations against an external criterion—the PHQ-8. Specifically, the PHQ-8 was added to the CFA, mentioned above. We found that the correlations between the PROMIS® questionnaire and the PHQ-8 were greater than or equal to .80 across all study conditions, demonstrating convergent validity of the LLM-generated items.

5.2 Engagement

We analyzed differences in the number of completed interactions by study conditions, including participants who did not complete the study. Participants in the CONTROL group had an average of 9.9 (SD=3.8) interactions with the agent while ITEM VARIANT ONLY and ITEM VARIANT PLUS groups had an average of 11.3 (SD=3.0) and 10.8 (SD=3.1) interactions, respectively, with no significant differences between conditions, ($F(2, 102)=1.81$, $p=.17$). Looking at the number of participants who met the minimum interaction requirement, we found a trending difference among the three groups ($X^2(2, N=105)=5.1$, $p=.08$). CONTROL condition had 80% of participants who met the requirement while ITEM VARIANTS ONLY condition had 97% and ITEM VARIANTS PLUS condition had 89%. We found no significant differences between participants who received treatment for depression and those who did not.

5.3 Perception of System & Agent

At the end of the two-week study period, participants rated the overall system as usable with a mean SUS score of 76.3 (SD=15.1). They also reported an above neutral rating (Mdn=4, IQR=2) on a 7-point scale for overall satisfaction with the system. Participants rated their satisfaction with the agent with a median of 3.5, which was significantly higher than a neutral score of 3, $Z=1.9$, $p=.03$, $r=.25$. In addition, they reported the agent's repetitiveness at 4.5, significantly greater than a neutral score of 3, $Z=6.6$, $p<.001$, $r=.71$. There were no significant differences among the three conditions for any of these measures (Table 3). From the content analysis of open-ended responses, we found that those in CONTROL were significantly more likely to mention "repetitiveness" compared to those in the two variant groups, $X^2(1, N=93)=5$, $p=.029$.

Table 3: User perceptions of the system, agent, and the questions. System-related items are on 7-point scales (from "not at all" to "very much"), with all other items on 5-point scales, with medians per group reported.

Category	Item	CONTROL	ITEM VARIANTS ONLY	ITEM VARIANTS PLUS
System	Mean system usability scale (0-100)	78.6 ± 12.9	75.2 ± 14.8	75.3 ± 17.3
	How satisfied are you with the system?	4.0	4.5	5.0
	How much would you like to continue using the system?	3.0	4.0	3.0
	Would you recommend the system to your friends and family?	4.0	4.0	3.0
	Mean of composite score	3.6 ± 1.7	4.0 ± 1.8	3.9 ± 1.9
Agent	How satisfied are you with the agent?	3.0	4.0	4.0
	How much would you like to continue talking with the agent?	3.0	4.0	3.0
	How much do you trust the agent?	3.0	3.0	3.0
	How much do you like the agent?	3.0	4.0	4.0
	How knowledgeable was the agent?	3.0	3.0	3.0
	How natural was your conversation with the agent?	2.0	2.5	2.0
	Did the agent feel repetitive?	5.0	4.0	4.0
	How would you characterize your relationship with the agent? (complete stranger - close friend)	2.5	3.0	2.0
	Mean of composite scores	3.0 ± 0.85	3.2 ± 0.92	3.1 ± 1.03
Questions	How coherent were the questions asked by the agent?	4.0	4.0	4.0
	How natural were the questions asked by the agent?	4.0	3.0	4.0
	Were the questions asked by the agent easy to understand?	4.0	4.5	5.0
	How often were the questions asked by the agent related to the topic of mental health? (never - almost constantly)	5.0	5.0	4.0
	Mean of composite score	4.2 ± 0.57	4.1 ± 0.53	4.1 ± 0.68

Participants reported higher overall satisfaction with the agent's questions, based on the median composite scores (Mdn=4.1) being greater than a neutral of 3, Z=8.2, p<.001, r=.86. Across all conditions, participants reported responses significantly above neutral of 3 for coherence (Mdn = 4.5, Z = 7, p<.001, r=.86), naturalness (Mdn=4, Z=4.1, p<.001, r=.43), how easy the questions were to understand (Mdn=4.5, Z=8.2, p<.001, r=.87), and relevance (Mdn=4.5, Z=8.5, p<.001, r=.88). No significant differences among study conditions were found (Table 3).

For the repeated measures collected after each interaction with the agent, we did not find any significant differences across the study conditions. Although not significantly different, participants in the CONTROL group reported a mean score of 5.1 (SD=1.3) for agent repetitiveness over 14 days while ITEM VARIANTS ONLY and ITEM VARIANTS PLUS conditions had means of 4.9 (SD=1.4) and 4.7 (SD=1.4), respectively.

5.4 Qualitative Results

We conducted a deductive thematic analysis of the open-ended responses (3,003 words), guided by sensitizing concepts that focused on participant satisfaction and feedback on additional features [14]. We used elements of the grounded theory method, including open, axial, and selective coding [16].

Comforting vs Uncanny Agents. Some participants expressed positive sentiments about talking to the agent and mentioned their willingness to interact daily: *"I like how someone was checking in with me daily to make sure I was alright."* [P43 - ITEM VARIANTS PLUS] and *"I liked the character, she felt like a safe person to talk to."* [P67 - ITEM VARIANTS ONLY]. One participant mentioned that their least favorite part of the system was that they were not able to have more interactions with the agent, *"I can't really give the answers I want or talk with her as long as I want"* [P13 - ITEM VARIANTS PLUS]. Another participant mentioned their desire to have deeper interaction with the agent on sharing their feelings, *"Maybe an option to expand on questions if I'm feeling down, like a deeper dive into my feelings, but still utilizing the multiple-choice selections"* [P3 - CONTROL]. Conversely, some found the interaction with the VA to be uncanny and unnatural. For instance, P80 [CONTROL] found the interaction with the agent strange, *"The attempt to make the robot AI feel human looking—it was uncanny valley to the max."*

Various Reasons for Repetitiveness. Most participants, especially in the CONTROL group, mentioned the repetitiveness of the system and agent. P82 [CONTROL] said, *"The repetition, being asked the same questions every single day, was a chore even though it wasn't very difficult. It lost its charm after the first few days."* P89 [CONTROL] also commented on the repetitiveness of questions, *"same questions over and over"*. Some participants, across all conditions, talked about how the user response options were repetitive. P88 [CONTROL] expressed that their least favorite part of the system was *"How repetitive the responses were"*. Others, even in the intervention groups, expressed how the agent's responses felt repetitive. For instance, P66 [ITEM VARIANTS ONLY] said, *"the feedback was repetitive"*.

Humor and Small Talk Does Not Always Work. Some participants mentioned *"hearing the jokes she had"* [P85] as their favorite part of the system, while others said that they would like to skip *"the bad dad jokes"* [P11]. Furthermore, P36 found the anecdotes and jokes to be forced, saying that they would like *"No forced stories and jokes in the beginning of the session."* P87 did not like the agent telling stories from her daily life saying, *"Probably the 'let me tell*

you about myself' stupidity. It was ridiculously patronizing that I was expected to take that seriously. A toddler would know an AI isn't getting sore throats and going to the movies". While some participants appreciated the humor and small talk in their interaction with the agent, others felt that the VA's small talk detracted from their ability to focus on answering the questionnaire.

6 DISCUSSION

We demonstrated the reliability and validity of LLM-generated question variants in a two-week validation study. The measurement alignment analysis and Cronbach's alpha showed the reliability of the questions administered in all three study conditions. In addition, the three different administrations of the PROMIS® depression questionnaires demonstrated good validity when compared to an external criterion of PHQ-8 which is another validated, standardized questionnaire for screening depression. These findings support **H1** by showing that the LLM-generated questionnaire variants do retain reliability and validity when compared to an external criterion. Furthermore, participants given the item variants found the questions coherent, natural, easy to understand, and relevant to the conversational topic based on the above neutral median self-reported data.

In total, 105 participants started the study and 93 participants met the minimum interaction requirement. The CONTROL group had the lowest percentage of participants (80%) who met the minimum interaction requirement. We saw a trending difference in the number of participants who met the minimum interaction requirement of seven interactions among the three study conditions. In addition, participants in the CONTROL group found the agent's questions more repetitive compared to participants with the LLM-generated questionnaire variants based on our content analysis reported in subsection 5.3. However, we did not find any other significant results from the post-study survey results. These findings partially support **H2** as we observed that questionnaires delivered in a different form daily did show trending differences in the number of participants who met the minimum interaction requirement.

Furthermore, we did not find any differences in engagement, satisfaction, or usability between ITEM VARIANTS ONLY and ITEM VARIANTS PLUS conditions. Therefore, our findings do not support **H3**, in which questionnaires delivered with humor and small talk will increase engagement compared to those without them in an interview with a VA. Qualitative findings showed that some participants found the jokes and stories entertaining and interesting while others found them to be forced. This demonstrates how simple jokes and small talk might not always be a reliable mechanism to increase engagement or satisfaction for repeated interviews.

6.1 Limitations & Future Work

There are several limitations to our study beyond the small convenience sample used. We conducted our evaluation study using only one standardized questionnaire, so it is unclear whether our results hold for other questionnaires. Our compensation structure with a strict minimum interaction requirement may have affected the results of engagement and interaction with the agent and could be seen as a limitation of our study design. In addition, the technical limitations of the prototype could have affected participant satisfaction.

Future work should consider ways of adding more variations in the dialogue structure, VA responses to users, and user question response options (scale anchor response options) to further reduce repetitiveness. Future work could also study the effects of letting participants interact with the agent using an unstructured input. This approach could further reduce perceptions of repetitiveness. However, finding a balance between personalization and standardization would need further examination. In addition, varying questions could increase the cognitive effort to process and respond. Future studies to understand the trade-offs between respondent fatigue/cognitive load and variations are needed.

Further research on incorporating humor and anecdotes generated by LLMs for longitudinal VA research should be considered. Our LLM-generated jokes were similar to how Jentzsch and Kersting [33] found ChatGPT to only produce limited joke patterns. Having more diverse jokes and stories (backstories of the agent or even stories of real people) and adapting to user conversation responses can be interesting future directions.

A previous study [31] has compared using VA and form-based questionnaires, and for our future studies, we will compare the effect of VA-based interactions with item variants with form-based questionnaires. Furthermore, future studies could examine utilizing more novel approaches, such as logical control [6], chain-of-thought prompting [67], or augmentation to use external tools [43], to create a safeguard for utilizing LLMs more directly to conversational agents.

7 CONCLUSION

We demonstrated that LLM-generated item variants for a depression questionnaire maintain good psychometric properties when delivered by a virtual agent. The LLM-generated item variants demonstrated validity and reliability and were seen to be coherent, natural, easy to understand, and relevant to the topic at hand. Additionally, participants who received these LLM-generated item variants generally found the agent less repetitive over a two-week study period compared to the CONTROL group. However, we found that including conversational humor and small talk in questionnaire administration interviews by an agent did not result in higher satisfaction or engagement. Striking a balance between personalization and standardization will be crucial for maintaining high-quality data collection and boosting response rates in delivering longitudinal self-report questionnaires. While using LLMs for producing questionnaire variants necessitates meticulous prompt preparation and manual output review, it offers the advantage of efficiently scaling and expediting the generation of diverse content. We view this study as a step forward in integrating LLMs into VAs to diversify and enhance questionnaire administration while maintaining validity and reliability.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute of Cancer of the National Institutes of Health under award number R01CA271145.

REFERENCES

[1] Jacob Anhøj, Lene Nielsen, et al. 2004. Quantitative and qualitative usage data of an Internet-based asthma monitoring tool. *Journal of Medical Internet Research* 6, 3 (2004), e57.

[2] Ana Antunes, Joana Campos, Manuel Guimarães, João Dias, and Pedro A. Santos. 2023. Prompting for Socially Intelligent Agents with ChatGPT. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (, Würzburg, Germany,) *(IVA '23)*. Association for Computing Machinery, New York, NY, USA, Article 20, 9 pages. https://doi.org/10.1145/3570945.3607303

[3] Tihomir Asparouhov and Bengt Muthén. 2014. Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural equation modeling: A multidisciplinary Journal* 21, 3 (2014), 329–341.

[4] Marc Auriacombe, Sarah Moriceau, Fuschia Serre, Cecile Denis, Jean-Arthur Micoulaud-Franchi, Etienne de Sevin, Emilien Bonhomme, Stephanie Bioulac, Melina Fatseas, and Pierre Philip. 2018. Development and validation of a virtual agent to screen tobacco and alcohol use disorders. *Drug and alcohol dependence* 193 (2018), 1–6.

[5] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[6] Erkan Basar, Divyaa Balaji, Linwei He, Iris Hendrickx, Emiel Krahmer, Gert-Jan de Bruijn, and Tibor Bosse. 2023. HyLECA: A Framework for Developing Hybrid Long-term Engaging Controlled Conversational Agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. 1–5.

[7] Timothy Bickmore, Amy Rubin, and Steven Simon. 2020. Substance use screening using virtual agents: towards automated Screening, Brief Intervention, and Referral to Treatment (SBIRT). In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–7.

[8] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2010. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence* 24, 6 (2010), 648–666.

[9] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Journal of medical Internet research* 20, 9 (2018), e11510.

[10] Ann Bowling. 2005. Mode of questionnaire administration can have serious effects on data quality. *Journal of public health* 27, 3 (2005), 281–291.

[11] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. BEAT: the Behavior Expression Animation Toolkit. In *Life-Like Characters: Tools, Affective Functions, and Applications*, Helmut Prendinger and Mitsuru Ishizuka (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 163–185. https://doi.org/10.1007/978-3-662-08373-4_8

[12] Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness. *arXiv preprint arXiv:2305.12947* (2023).

[13] David Cella, William Riley, Arthur Stone, Nan Rothrock, Bryce Reeve, Susan Yount, Dagmar Amtmann, Rita Bode, Daniel Buysse, Seung Choi, et al. 2010. Initial adult health item banks and first wave testing of the patient-reported outcomes measurement information system (PROMIS™) network: 2005–2008. *Journal of clinical epidemiology* 63, 11 (2010), 1179.

[14] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.

[15] Andrew Cleary and Nigel Balmer. 2015. The impact of between-wave engagement strategies on response to a longitudinal survey. *International Journal of Market Research* 57, 4 (2015), 533–554.

[16] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[17] Samuel Rhys Cox, Ashraf Abdul, and Wei Tsang Ooi. 2023. Prompting a Large Language Model to Generate Diverse Motivational Messages: A Comparison with Human-Written Messages. In *Proceedings of the 11th International Conference on Human-Agent Interaction*. 378–380.

[18] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[19] W De Heer and E De Leeuw. 2002. Trends in household survey nonresponse: A longitudinal and international comparison. *Survey nonresponse* 41 (2002), 41–54.

[20] Nicola R Dean and Tamara Crittenden. 2016. A five year experience of measuring clinical effectiveness in a breast reconstruction service using the BREAST-Q patient reported outcomes measure: a cohort study. *Journal of Plastic, Reconstructive & Aesthetic Surgery* 69, 11 (2016), 1469–1477.

[21] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1061–1068.

[22] Joel R Evans and Anil Mathur. 2005. The value of online surveys. *Internet research* 15, 2 (2005), 195–219.

[23] Oluwadamilola M Fayanju, Tinisha L Mayo, Tracy E Spinks, Seohyun Lee, Carlos H Barcenas, Benjamin D Smith, Sharon H Giordano, Rosa F Hwang, Richard A Ehlers, and Jesse C Selber. 2016. Value-based breast cancer care: a multidisciplinary approach for defining patient-centered outcomes. *Annals of surgical oncology* 23 (2016), 2385–2390.

[24] R Michael Furr. 2021. *Psychometrics: an introduction*. SAGE publications.

[25] Robert M Groves and Emilia Peytcheva. 2008. The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly* 72, 2 (2008), 167–189.

[26] Joschka Haltaufderheide and Robert Ranisch. 2024. The Ethics of ChatGPT in Medicine and Healthcare: A Systematic Review on Large Language Models (LLMs). *arXiv preprint arXiv:2403.14473* (2024).

[27] Hyemin Han. 2024. Using measurement alignment in research on adolescence involving multiple groups: A brief tutorial with R. *Journal of Research on Adolescence* 34, 1 (2024), 235–242.

[28] Leon Hanschmann, Ulrich Gnewuch, and Alexander Maedche. 2023. Saleshat: A LLM-Based Social Robot for Human-Like Sales Conversations. In *International Workshop on Chatbot Research and Design*. Springer, 61–76.

[29] Alberto Hernández-Reyes, Fernando Cámara-Martos, Guillermo Molina Recio, Rafael Molina-Luque, Manuel Romero-Saldaña, and Rafael Moreno Rojas. 2020. Push notifications from a mobile app to improve the body composition of overweight or obese women: randomized controlled trial. *JMIR mHealth and uHealth* 8, 2 (2020), e13747.

[30] Victoria Huynh, Kathryn Colborn, Shelby Smith, Levi N Bonnell, Gretchen Ahrendt, Nicole Christian, Simon Kim, Dan D Matlock, Clara Lee, and Sarah E Tevis. 2021. Early trajectories of patient reported outcomes in breast cancer patients undergoing lumpectomy versus mastectomy. *Annals of Surgical Oncology* 28 (2021), 5677–5685.

[31] Shashank Jaiswal, Michel Valstar, Keerthy Kusumam, and Chris Greenhalgh. 2019. Virtual human questionnaire for analysis of depression, anxiety and personality. In *Proceedings of the 19th ACM international conference on intelligent virtual agents*. 81–87.

[32] Bernard J Jansen, Soon-gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal* 4 (2023), 100020.

[33] Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. *arXiv preprint arXiv:2306.04563* (2023).

[34] Patricia Kissinger, Janet Rice, Thomas Farley, Shelly Trim, Kayla Jewitt, Victor Margavio, and David H Martin. 1999. Application of computer-assisted interviews to sexual behavior research. *American journal of epidemiology* 149, 10 (1999), 950–954.

[35] Oscar NE Kjell, Katarina Kjell, and H Andrew Schwartz. 2023. Beyond Rating Scales: With Targeted Evaluation, Language Models are Poised for Psychological Assessment. *Psychiatry Research* (2023), 115667.

[36] Rafal Dariusz Kocielnik. 2021. *Designing engaging conversational interactions for health & behavior change*. Ph. D. Dissertation.

[37] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders* 114, 1-3 (2009), 163–173.

[38] Austin Le, Benjamin H Han, and Joseph J Palamar. 2021. When national drug surveys "take too long": An examination of who is at risk for survey fatigue. *Drug and alcohol dependence* 225 (2021), 108769.

[39] Michael W Link and Ali H Mokdad. 2005. Alternative modes for health surveillance surveys: an experiment with web, mail, and telephone. *Epidemiology* (2005), 701–704.

[40] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.

[41] Stefania Mancone, Pierluigi Diotaiuti, Giuseppe Valente, Stefano Corrado, Fernando Bellizzi, Guilherme Torres Vilarino, and Alexandro Andrade. 2023. The Use of Voice Assistant for Psychological Assessment Elicits Empathy and Engagement While Maintaining Good Psychometric Properties. *Behavioral Sciences* 13, 7 (2023), 550.

[42] Katja Lozar Manfreda, Michael Bosnjak, Jernej Berzelak, Iris Haas, and Vasja Vehovar. 2008. Web surveys versus other survey modes: A meta-analysis comparing response rates. *International journal of market research* 50, 1 (2008), 79–104.

[43] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842* (2023).

[44] Yul Ha Min, Jong Won Lee, Yong-Wook Shin, Min-Woo Jo, Guiyun Sohn, Jae-Ho Lee, Guna Lee, Kyung Hae Jung, Joohon Sung, and Beom Seok Ko. 2014. Daily collection of self-reporting sleep disturbance data via a smartphone app in breast cancer patients receiving chemotherapy: a feasibility study. *Journal of medical Internet research* 16, 5 (2014), e135.

[45] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical

health. *JAMA internal medicine* 176, 5 (2016), 619–625.
[46] Jessica Clark Newman, Don C Des Jarlais, Charles F Turner, Jay Gribble, Phillip Cooley, and Denise Paone. 2002. The differential effects of face-to-face and computer interview modes. *American journal of public health* 92, 2 (2002), 294–297.
[47] Stefan Olafsson, Teresa K. O'Leary, and Timothy W. Bickmore. 2020. Motivating Health Behavior Change with Humorous Virtual Agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland, UK) *(IVA '20)*. Association for Computing Machinery, New York, NY, USA, Article 42, 8 pages. https://doi.org/10.1145/3383652.3423915
[48] Stefan Olafsson, Paola Pedrelli, Byron C. Wallace, and Timothy Bickmore. 2023. Accomodating User Expressivity While Maintaining Safety for a Virtual Alcohol Misuse Counselor. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (Würzburg, Germany) *(IVA '23)*. Association for Computing Machinery, New York, NY, USA, Article 3, 9 pages. https://doi.org/10.1145/3570945.3607361
[49] Meltem Kurt Pehlivanoğlu, Muhammad Abdan Syakura, and Nevcihan Duru. 2023. Enhancing Paraphrasing in Chatbots Through Prompt Engineering: A Comparative Study on ChatGPT, Bing, and Bard. In *2023 8th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 432–437.
[50] Stephen R Porter, Michael E Whitcomb, and William H Weitzer. 2004. Multiple surveys of students and survey fatigue. *New directions for institutional research* 2004, 121 (2004), 63–73.
[51] Yvette Pronk, Peter Pilot, Justus M Brinkman, Ronald J van Heerwaarden, and Walter van der Weegen. 2019. Response rate and costs for automated patient-reported outcomes collection alone compared to combined automated and manual collection. *Journal of patient-reported outcomes* 3 (2019), 1–8.
[52] Katyanna Quach. 2020. Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves. *The Register* (2020).
[53] Ilya Razykov, Roy C Ziegelstein, Mary A Whooley, and Brett D Thombs. 2012. The PHQ-9 versus the PHQ-8—is item 9 useful for assessing suicide risk in coronary artery disease patients? Data from the Heart and Soul Study. *Journal of psychosomatic research* 73, 3 (2012), 163–168.
[54] Catherine A Roster, Robert D Rogers, Gerald Albaum, and Darin Klein. 2004. A comparison of response characteristics from web and telephone surveys. *International Journal of Market Research* 46, 3 (2004), 359–373.
[55] Michael F Schober and Frederick G Conrad. 1997. Does conversational interviewing reduce survey measurement error? *Public opinion quarterly* (1997), 576–602.
[56] Ryan M Schuetzler, Justin Scott Giboney, G Mark Grimes, and Jay F Nunamaker Jr. 2018. The influence of conversational agent embodiment and conversational relevance on socially desirable responding. *Decision Support Systems* 114 (2018), 94–102.
[57] Javier Sevilla-Salcedo, Enrique Fernández-Rodicio, Laura Martín-Galván, Álvaro Castro-González, José C Castillo, and Miguel A Salichs. 2023. Using Large Language Models to Shape Social Robots' Speech. (2023).
[58] David M Shannon and Carol C Bradshaw. 2002. A comparison of response rate, response time, and costs of mail and electronic surveys. *The Journal of Experimental Education* 70, 2 (2002), 179–192.
[59] Angela Sinickas. 2007. Finding a cure for survey fatigue. *Strategic Communication Management* 11, 2 (2007), 11.
[60] Kirsten P Smith and Nicholas A Christakis. 2008. Social networks and health. *Annu. Rev. Sociol* 34 (2008), 405–429.
[61] Charles F Turner, Leighton Ku, Susan M Rogers, Laura D Lindberg, Joseph H Pleck, and Freya L Sonenstein. 1998. Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science* 280, 5365 (1998), 867–873.
[62] Jonathan B VanGeest, Timothy P Johnson, and Verna L Welch. 2007. Methodologies for improving response rates in surveys of physicians: a systematic review. *Evaluation & the health professions* 30, 4 (2007), 303–321.
[63] Laura Vardoulakis. 2013. Social desirability bias and engagement in systems designed for long-term health tracking. (2013).
[64] Vasja Vehovar, Zenel Batagelj, Katja Lozar Manfreda, and Metka Zaletel. 2002. Nonresponse in web surveys. *Survey nonresponse* (2002), 229–242.
[65] Sudheer Vemuru, Shelby Smith, Kathryn Colborn, Victoria Huynh, Laura Leonard, Levi Bonnell, Laura Scherer, Dan Matlock, Clara Lee, and Simon Kim. 2023. Access to Results of Patient Reported Outcome Surveys Does Not Improve Survey Response Rates. *Journal of Surgical Research* 283 (2023), 945–952.
[66] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2023. Leveraging large language models to power chatbots for collecting user self-reported data. *arXiv preprint arXiv:2301.05843* (2023).
[67] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.
[68] Suzanne Weisband and Sara Kiesler. 1996. Self disclosure on computer forms: Meta-analysis and implications. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3–10.
[69] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *arXiv preprint arXiv:2306.15895* (2023).
[70] Hye Sun Yun, Iain Marshall, Thomas Trikalinos, and Byron Wallace. 2023. Appraising the Potential Uses and Harms of LLMs for Medical Systematic Reviews. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10122–10139. https://doi.org/10.18653/v1/2023.emnlp-main.626