Keeping Users Engaged During Repeated Administration of the Same Questionnaire: Using Large Language Models to Reliably Diversify Questions

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 3922

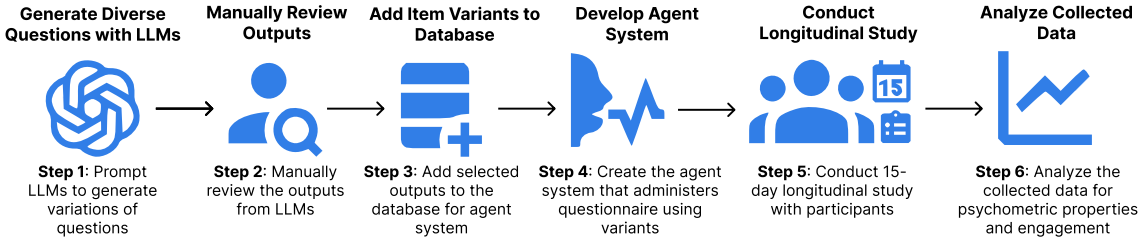

Fig. 1. A workflow diagram of the longitudinal validation study which evaluated the validity, reliability, and user engagement of utilizing large language model-generated variants of a standardized depression questionnaire.

Standardized, validated questionnaires are vital tools in HCI research and healthcare, offering dependable self-report data. However, their repeated use in longitudinal or pre-post studies can induce respondent fatigue, impacting data quality via response biases and decreased response rates. We propose utilizing large language models (LLMs) to generate diverse questionnaire versions while retaining good psychometric properties. In a longitudinal study, participants engaged with our agent system and responded daily for two weeks to either a standardized depression questionnaire or one of two LLM-generated questionnaire variants, alongside a validated depression questionnaire. Psychometric testing revealed consistent covariation between the external criterion and the focal measure administered across the three conditions, demonstrating the reliability and validity of the LLM-generated variants. Participants found the repeated administration of the standardized questionnaire significantly more repetitive compared to the variants. Our findings highlight the potential of LLM-generated variants to invigorate questionnaires, fostering engagement and interest without compromising validity.

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

Additional Key Words and Phrases: questionnaires, engagement, large language models, virtual agents, health, longitudinal research

ACM Reference Format:
Anonymous Author(s). 2018. Keeping Users Engaged During Repeated Administration of the Same Questionnaire: Using Large Language Models to Reliably Diversify Questions. In *CHI '24: ACM CHI Conference on Human Factors in Computing Systems, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Many repeated-measures evaluation studies or longitudinal interventions require that the same self-report questionnaire be administered to the same individual multiple times. For example, in HCI, dozens of studies have investigated changes

Manuscript submitted to ACM

in usability or related measures over time spans ranging from days to years [29]. In medicine, patient-reported outcomes (PROs) are used to obtain self-reports of a patient's condition at home, typically involving repeated administration of surveys to capture symptoms or quality of life [21].

However, response rates to repeated surveys tend to decline over time as respondents become fatigued by repeatedly filling out the same questionnaires [18, 42, 47]. Even in medicine, where PROs can be used as the basis for treatment decisions, the longitudinal survey completion rates can be as low as 48% [18, 26, 42]. Dwindling response rates can lead to a non-response measurement bias [23] and limit the ability to evaluate important changes over time.

Several strategies have been proposed to increase repeated-measure response rates, including incentives [60], more frequent contact and engagement with respondents [14, 49], and providing survey responses back to the individuals being surveyed [63]. In addition, a variety of approaches have been studied to increase the usage rates for repeated interactions with automated systems in general, including the use of syntactic and visual variability in the interface [8], humor and other forms of entertainment [46], reminders [25], and social support and reinforcement [58].

In this work, we explore the use of three strategies to increase response rates to a PRO administered daily for two weeks. Our primary focus is on the use of survey questions that vary in every administration so that the surveys appear different. We explore going beyond straightforward syntactic variation of questions to semantically similar variants generated by large language models (LLMs). Syntactic variations primarily entail reordering words within a sentence, whereas semantically similar variations entail employing slightly different words or phrases that convey comparable meanings. We also explore the administration of questionnaires by a virtual agent (VA) that simulates a face-to-face interview with a researcher rather than simply filling out a text-based form. Finally, we explore the use of humor and small talk to make daily interactions with the VA more conversational, entertaining, and engaging.

Many questionnaires, including most PROs, are validated using laborious methods involving testing with dozens, if not thousands, of respondents to establish reliability and validity [22]. An important question raised when validated questionnaires are modified is whether the new derivative versions retain the reliability and validity of the original validated form. We report the results of a longitudinal validation 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 LLM-generated questionnaire variants. All participants filled out an additional standardized, validated depression questionnaire which was a criterion for comparison. Our hypotheses are:

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

Our primary contributions encompass the introduction of an innovative approach, using LLMs, to generate diversified versions of validated questionnaires reliably for a VA system. Furthermore, we demonstrate the feasibility of employing these questionnaire variants to enhance user engagement and mitigate repetitiveness. Additionally, our work offers guidelines for effectively prompting LLMs in generating questionnaire variants and outlines avenues for future research aimed at maintaining longitudinal self-reports through VAs. In this article, we first review related work and then describe the VA system we used for conducting our experimentation. We then report our methodology for generating questionnaire variants using LLMs and finally report the results of our validation study.

2 RELATED WORK

Our work draws on previous research on alternative delivery methods for questionnaires, and user engagement methods for longitudinal research. 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 [17, 20, 36, 53, 56]. However, even web-based surveys are not immune to dwindling response rates [40, 62]. Especially 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, 31, 48, 57]. Increasing the quality of self-report data and completion rates through surveys remains an important challenge for researchers to overcome. In the following sections, we provide an overview of how computers and agent systems can be used for quality human self-reports and for maintaining longitudinal self-reports to better situate our contribution.

2.1 Computers & Agents for Quality Self-Report

Several past studies have shown that utilizing 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 [28, 45, 59, 66]. Several studies have expanded upon this approach by incorporating conversational agents and virtual agents (VAs), as research has indicated that using conversational interviews for surveys can effectively reduce errors. [54]. Lucas et al. [38] conducted health-screening interviews with VAs and had participants believe that the VA was controlled by either humans or automation. The results revealed that participants who thought the VA to be automated reported a lower fear of self-disclosure and displayed their sadness more intensely, which suggests that automated VAs can assist in obtaining more honest patient self-report data. A similar study by Schuetzler et al. [55] demonstrated that people disclose more about their sensitive behavior to a conversational agent than to a human, but people disclose less when the conversational agent appears to understand.

Furthermore, prior work has demonstrated how medical questionnaires or PROs administered by VAs are valid and statistically equivalent to human or self-administered questionnaires [3, 7]. For example, Jaiswal et al. [27] 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 questionnaires administered by VAs were statistically equivalent to human or self-administered questionnaires. Additionally, Mancone et al. [39] 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.

Existing research shows great potential for using VAs to administer healthcare questionnaires because they can elicit user engagement without compromising the validity and reliability of the questionnaire responses. Our work extends existing research as we also have a VA administering the questionnaires but incorporate additional variability by using questionnaire variants and small talk generated by LLMs.

2.2 Maintaining Longitudinal Self-Report with Agents

Longitudinal research studies that require self-reporting often have low completion rates and frequencies over time [1]. Although VAs increase engagement when administering questionnaires, they still suffer from user disengagement over time. It has been shown that the length of the first interaction with a VA is the primary predictor of the number of healthcare questionnaires completed by a participant [61]. The findings show that longer first interactions can result in fewer completed questionnaires.

Fig. 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.

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 lead to increased engagement. Similarly, our study incorporates variability by generating multiple variations to the items in a standardized questionnaire and includes additional conversational content, such as agent personal anecdotes and jokes, to further increase engagement. Instead of using manually written stories, jokes, and question variants, we demonstrate how LLMs can assist in creating short variable content for automated VAs to increase user engagement and interest.

3 SYSTEM DESIGN

To evaluate our study hypotheses, we created a prototype VA system deployed over the web that participants could interact with every day for an extended period of time (Figure 2). Our agent is a 3D animated character who 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, is automatically generated using the Behavior Expression Animation Toolkit [11]. Agent utterances are generated using template-based text generation. The agent's dialogue is driven by a hierarchical task network-based dialogue engine. The VA system is implemented using the Unity3D game engine and the CereProc speech synthesizer.

Our agent, named 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) [12] as our baseline. 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".

To support our study, we created three versions of the VA system:

- **CONTROL**: The VA administers the PROMIS® questionnaire in question format every day, in addition to minimal greeting and farewell dialogue.

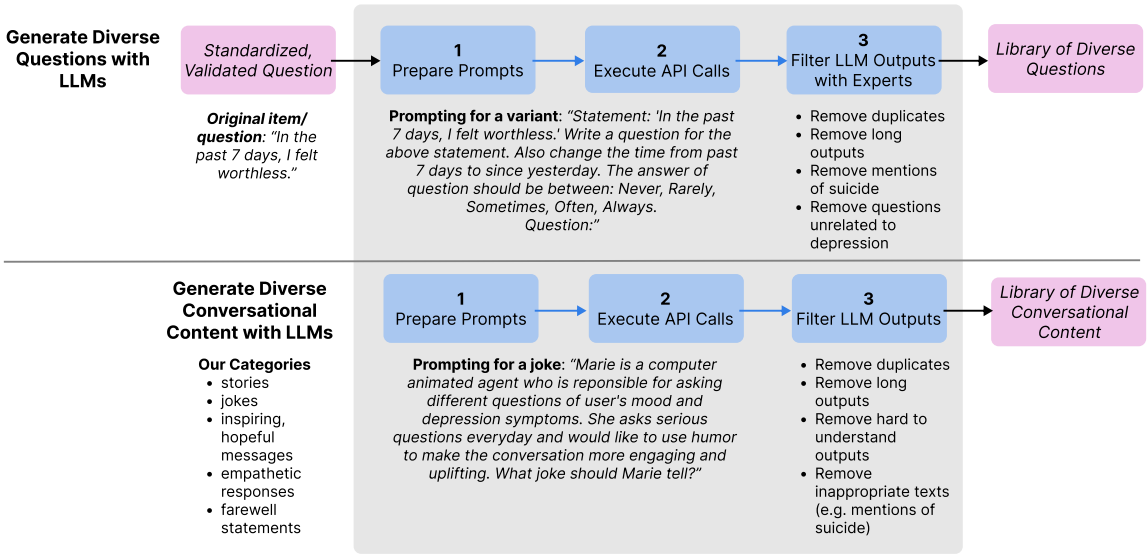

Fig. 3. A workflow diagram of how LLMs were used to generate diverse questions and conversational content such as stories and jokes. A simple example is provided for each process.

- **ITEM VARIANTS ONLY**: The VA administers different versions of each question every day, in addition to minimal greeting and farewell dialogue.
- **ITEM VARIANTS PLUS**: The VA administers different versions of each question every day, in addition to social chat, jokes, anecdotes, and empathetic feedback.

The original wording of each PROMIS® questionnaire item (question), followed by the question format and their LLM-based variants, are available in Table 1. The detailed process for generating variants using LLMs is described next.

3.1 Generating Questionnaire Variants Using LLMs

By using LLMs, we can yield a greater range of semantic variations for each question, potentially increasing user engagement. However, it is important to note that such variations may not always align with the goal or purpose of the original questionnaire items (i.e., the model-generated output may include harmful language or language that deviates from the main concept of measurement). 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, 43, 50].

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. We used OpenAI API access for GPT-3 and a web-based user interface for accessing ChatGPT since API access was not available at the time of the study. For prompting, we provided the original item and the response scale and asked the LLM to paraphrase the main item into a new question. These models were prompted to generate new variants that fit the response scale. The exact prompts used for both LLMs are available in Appendix A. In total, 178 new variants of the eight questions of the questionnaire were generated. A psychologist then ranked and

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 [12]. All the items are rated on a five-point scale from 1 = "Never" to 5 = "Always". The total number 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

filtered the new variants to create a final list that matched the original question's meaning and purpose. We ended up with 67 new variants of the questionnaire questions, with a few samples available in Table 1.

3.2 Generating Conversational and Empathetic Language Using LLMs

4 LONGITUDINAL EVALUATION STUDY

From April to May 2023, we conducted an online longitudinal user evaluation study to evaluate the psychometric properties of LLM-generated questionnaire variants, user engagement, and user perceptions of the VA survey system. The duration of the study was 15 days. For the first 14 days of the longitudinal study, participants were asked to talk with the VA once a day, which lasted a few minutes and answer a short online survey after each interaction. On the 15th

Table 2. Conversational Content Generated by LLMs. This table provides some samples of conversational small talk and humor generated by LLMs that were added to the ITEM VARIANTS PLUS study condition. The number of content for each category that we ended up using for the longitudinal evaluation study is also provided.

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

day, the participants completed a final online survey. The experiment followed a between-subjects design, in which a user was randomly assigned to one of the three study conditions.

In one condition — CONTROL, we had the VA administer daily the standardized eight-item PROMIS® Short Form Depression questionnaire in question format to allow for the VA to conversationally administer the questionnaire. In the two intervention conditions — ITEM VARIANTS ONLY and ITEM VARIANTS PLUS, we had the agent randomly choose a variant generated by LLMs for each item of the PROMIS® Short Form Depression questionnaire from a library of pre-generated, pre-examined variations. These item variants were generated according to subsection 3.1. In addition to LLM-generated item variants, the ITEM VARIANTS PLUS includes additional conversational content for dialogue, such as the sharing of anecdotes, jokes, empathetic responses, inspiring or hopeful messages, and farewells generated by LLMs. In a typical session for the ITEM VARIANTS PLUS condition, the agent would first share 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 selected motivational message and farewell statements. All three conditions required the agent to deliver the same baseline questions for the first interaction. However, after the initial interaction, the two intervention conditions only administered a questionnaire using LLM-generated variants.

4.1 Measures & Data Collection

Prior to the first interaction with the agent, we collected sociodemographic information from each participant. Subsequently, we collected the following 4 items with a 7-point Likert 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").

During the study, we collected system usage logs. We assessed user engagement (the number of completed interactions) using these system-logged usage metrics. All responses to the agent's administration of the depression questionnaire were automatically collected into the database of the system and used later for analysis.

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) [30, 51], the system usability scale (SUS) [5], overall system satisfaction measures, agent satisfaction measures, and measures related to user perception of questions asked by the agent.

Measures related to overall system satisfaction, agent satisfaction, and user perception of the questions asked by the agent were developed internally for this project. The overall system satisfaction measures consist of seven items which include three items with 7-point Likert scale response (ranging from 1 = "Not at all" to 7 = "Very much"): 1) "How satisfied are you with the system?", 2) "How much would you like to continue using the system?", and 3) "Would you recommend the system to your friends and family?". The other four items were open-ended questions: 1) "Did the system work the way you wanted it to? If no, why not?", 2) "What was your favorite part of the system?", 3) "What was your least favorite part of the system?", and 4) "Are there any additional features that would have improved your experience with the agent?" For the agent-related measures, we asked participants about their satisfaction, desire to continue talking to the agent, trust, likability, knowledgeability, naturalness, and repetitiveness on a 5-point Likert scale ranging from 1 = "Not at all" to 5 = "Very much". To determine user perception of their relationship with the agent, we ask participants to respond to "How would you characterize your relationship with the agent?" on a 5-point scale ranging from 1 = "Complete Stranger" to 5 = "Close Friend". In terms of user perceptions of questions asked by the agent, we asked how coherent, natural, easy to understand, and relevant questions were on a 5-point Likert scale (ranging from 1 = "Not at all" to 5 = "Very much" or "Always").

4.2 Participants

Participants of the study were recruited via an online research platform (www.prolific.co). 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 with the agent once per day and answer a short survey at the end of each interaction. In addition, they were told to fill out pre- and post-study surveys and interact with the agent a minimum of seven times during the two-week period to fully complete the study and be compensated. The minimum interaction requirement was to ensure that each participant was given the questionnaire several times. The study was approved by our institution's institutional review board, and participants were compensated for their time ($13.00, for a total of 60 minutes spanning 15 days).

5 RESULTS

A total of 105 participants started the longitudinal evaluation study, where there were 35 participants assigned to each study condition. However, only 93 participants met the compensation requirements and successfully completed the study. Participants who successfully completed the study were on average 40 (*SD*=12, *Mdn*=38, *Range*=21~73) years old. The gender breakdown was 48.4% men, 47.3% women, 3.3% non-binary, and 1.1% other. Participants were 75.3% white, 9.9% multiracial, 7.5% Black/African-American, 4.3% Asian/Asian American, 2.2% Hispanic/Latinx, and

1.1 % American Indian/Alaska Native. All participants had at least a high school degree or equivalent (44.1% with a bachelor's degree, 24.7% with some college, 10.8% with a master's degree, 9.7% with an associate degree, and 2.2% with a doctoral/professional degree). When asked if they are currently in therapy or taking medication for depression, 78.5% of participants said "no", 20.4% said "yes", and 1.1% preferred not to answer.

5.1 Psychometric Properties of LLM-generated Item Variants

First, we calculated Cronbach's Alpha [16] to measure the internal consistency or reliability of the eight depression questions. Cronbach's Alpha for the CONTROL condition was α = .76 while the item variants were α = .65. Although the α for the version with variants was lower compared to the CONTROL, it showed acceptable internal consistency.

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 one to assess how the properties of individual items differ across groups. 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 allowed examination of the degree to which the groups can be "aligned" on the same scale. To begin, 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 is captured by true mean differences in the groups rather than different item properties across groups [2]. 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 can be attributed to factor mean and variance differences across the groups. In other words, the properties of the items were very consistent across groups. In a simulation, Asparouhov and Muthen 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® depression questionnaire administered across the three study conditions, concluding 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® depression, 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® depression questionnaire and the PHQ-8 were greater than or equal to 0.80 across all of the study conditions, demonstrating convergent validity of the LLM-generated items.

5.2 Engagement

To understand the longitudinal effects of our method for administering depression questionnaires on engagement, we investigated user engagement via the number of completed interactions by participants. In our analysis, we also included participants who did not fully complete the study to get a more complete picture. This results in each study condition having 35 participants. 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. A one-way ANOVA revealed that there was no statistically significant difference in mean number of interactions among the three groups ($F(2, 102) = 1.81$, $p = 0.17$). However, when we look at the number of participants who met the minimum requirements to successfully complete the study or not, we do find a trending

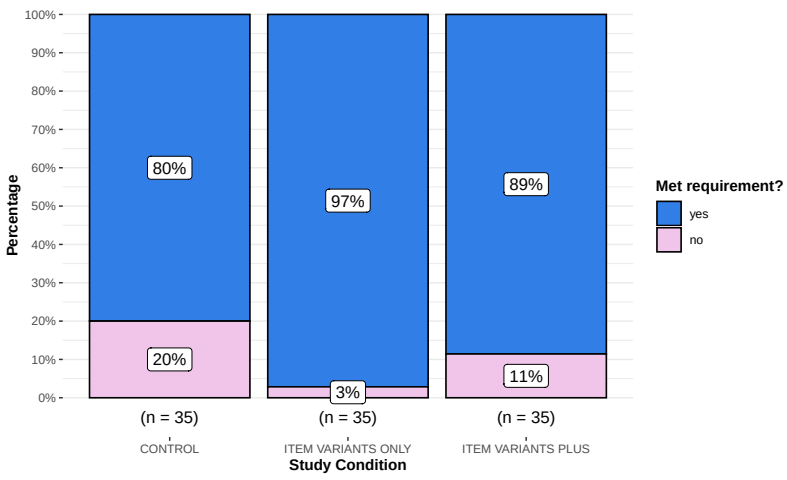

Fig. 4. A plot that shows the distribution of participants who met the requirement of interacting with the agent at least seven times during the 14 days.

difference among the three groups ($X^2(2, N = 105) = 5.1$, $p = .08$). Figure 4 shows the distribution of participants who met or did not meet the minimum interaction requirements to successfully complete the full study.

During each interaction with the agent, participants responded to the depression questionnaire with various responses. The majority of the responses for all the questions comprised of "Never" or "Rarely." Figure 5 shows the percentages of the responses to each item for all study conditions. We did not observe any significant differences in responses across study conditions.

5.3 Satisfaction & Perception of System and Agent

Based on the analysis of self-reported data, participants found the overall system usable with the mean SUS score of 76.3 ($SD = 15.1$). There were no significant differences among the 3 study conditions. Table 3 provides the mean SUS scores for each condition. In terms of general satisfaction with the system after the two-week study period, participants reported a median of 4 ($IQR = 2$) which is neutral. Overall, the numbers across all study conditions were about neutral or slightly below (Table 3). There were no significant differences among the study conditions.

In terms of user satisfaction with the agent, participants reported lower overall satisfaction with the agent based on the mean composite scores ($M = 3.1$, $SD = .93$) than neutral of 4, $t(92) = -9$, $p < 0.001$. However, we observed that participants reported higher satisfaction with agent on the single-item with a median of 3.5 than neutral of 3, $Z = 1.9$, $p = .03$, $r = .25$. In addition, participants reported that the agent is more repetitive with a median of 4.5 than neutral of 3, $Z = 6.6$, $p < .001$, $r = .71$. The median responses on user perceptions of the agent from each study condition are all provided in Table 3. In our analysis, there were no significant differences among the three conditions.

After conducting a content analysis on the open-ended responses from participants about their thoughts on the system and the agent, we found various mentions of "repetitiveness" across all conditions. A chi-square test of independence was performed to examine the relation between the version of questions administered and the mentioning of "repetitiveness." The relation between these variables was significant, $X^2(1, N = 93) = 5$, $p = .029$. Participants who interacted with the

Percentages of Response Values

Factors in Missing Days of Interaction

PROMIS Depression Items	Response options (1-5): Higher numbers mean higher severity						
	Never	Rarely	Sometimes	Often	Always	Skip	Missing
1 - worthless	47.3%	14.6%	10.7%	5.1%	3.4%	1.3%	17.7%
2 - helpless	34.3%	20.1%	17.3%	6.1%	3.3%	1.2%	17.7%
3 - depressed	34.6%	21.6%	13.9%	5.0%	6.1%	1.2%	17.7%
4 - hopeless	39.4%	18.1%	14.7%	5.5%	3.8%	0.8%	17.7%
5 - a failure	31.2%	19.0%	16.7%	6.6%	7.6%	1.2%	17.7%
6 - unhappy	28.8%	23.4%	16.5%	8.5%	4.1%	0.8%	17.7%
7 - nothing to look forward to	30.8%	17.4%	16.9%	9.3%	6.5%	1.4%	17.7%
8 - nothing could cheer me up	35.9%	21.0%	15.7%	5.5%	3.3%	0.8%	17.7%

Legend

0% 20% 40% 60% 80% 100%

Data from all 3 study conditions. N = 10416

Fig. 5. A table showing the percentages of responses for each item from all three study conditions. Both skipped and missing responses are also included.

VA administering the standardized questionnaire were more likely to mention any repetitiveness of the system than those who were given LLM-generated item variants.

In terms of user perception of the questions asked by the agent, participants were generally positive. Participants reported higher overall satisfaction with the questions based on the median composite scores (Mdn = 4.1) 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.

For the repeated measures that we collected after each interaction with the agent, we did not find any significant differences across the study conditions. Figure 6 shows the changes in the mean user satisfaction and perception of the agent over the two-week study period for each study condition. 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

As mentioned in subsection 4.1, we asked participants open-ended questions in the final survey, asking them about different aspects of their experience using the system. We conducted a deductive thematic analysis guided by sensitizing concepts that focused on participant satisfaction and feedback on additional features [13]. In our analysis, the first and

Table 3. User perceptions of the system, agent, and the questions asked by the agent. All the system-related single-item questions were on a 7-point Likert scale (from "not at all" to "very much") while the agent- and question-related ones were on a 5-point Likert scale. The median for each study group is reported. We also report the mean system usability scale (0-100) and the individual composite scores were calculated by averaging all the single-item responses. The responses from "Did the agent feel repetitive?" were reversed before calculating the agent composite score.

Category	Item	CONTROL	ITEM VARIANTS ONLY	ITEM VARIANTS PLUS
System	Mean system usability scale	78.6 ± 12.9	75.2 ± 14.8	75.3 ± 17.3
	Satisfaction with the system	4.0	4.5	5.0
	Like to continue using the system	3.0	4.0	3.0
	Recommend to 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

second authors coded the responses independently and then collectively grouped the codes to come up with categories and general themes iteratively. We used elements of the grounded theory method, including open, axial, and selective coding [15]. We coded 3,003 words of open-ended responses labeling emerging concepts to arrive at a code book using NVivo 12.7.0 qualitative analysis software.

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], *"I liked the character, she felt like a safe person to talk to."* [P67 - ITEM VARIANTS ONLY], and *"My favorite part of the system was interacting with the character and answering questions"* [P74 - ITEM VARIANTS ONLY]. One participant even 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*

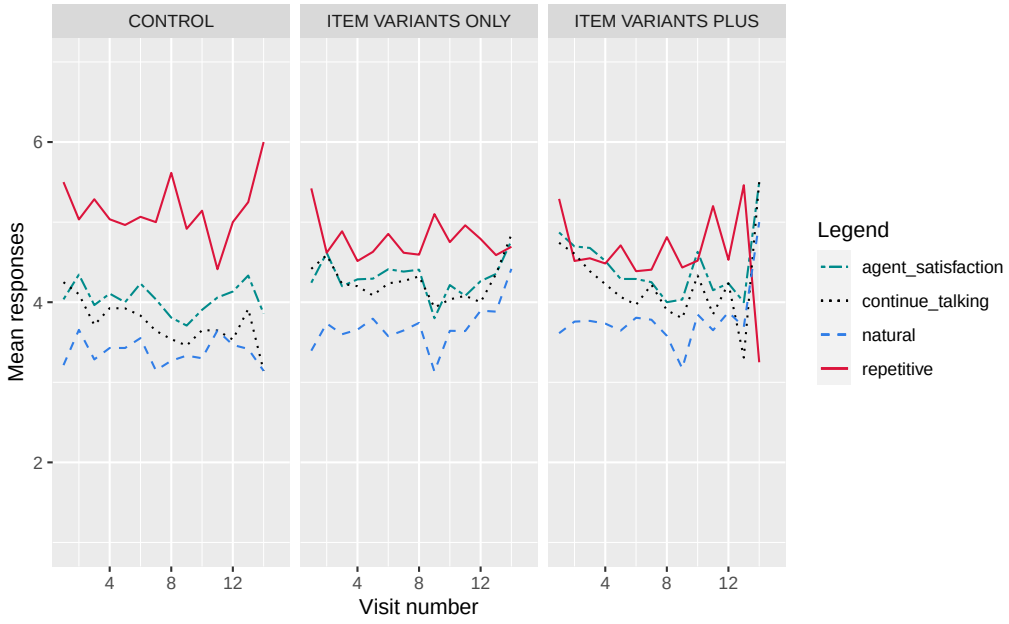

Fig. 6. Changes in satisfaction and perception of the agent throughout the 14-day study interaction period by study condition.

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]. This shows that some participants wanted longer and deeper interactions with the agent to discuss how they were feeling in more detail. In conclusion, several participants expressed positive sentiments about interacting with the agent to respond to the depression questionnaire, highlighting its role as a comforting presence.

Moreover, there were varied opinions regarding the experiences of conversing with a VA. Some participants mentioned how they liked talking to a VA rather than a real human, *"There was no interaction with an actual person"* [P94 - CONTROL]. Conversely, some found the interaction with the VA 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."* Moreover, P64 [CONTROL] said, *"The interaction was super unnatural and felt awkward."* This can help explain the low median scores observed across all study conditions for agent naturalness from the post-study survey (Table 3). The diverse range of opinions on conversing with the VA highlights the complexity of using VAs to administer questionnaires.

Furthermore, as the participants were asked to interact with the system in a daily manner, most of the participants, especially in the CONTROL group, mentioned the repetitiveness of the system and the 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."* Having to interact with the VA in the same manner daily made the experience repetitive. Some participants 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 expressed other aspects of repetitiveness. For instance, P66 [ITEM VARIANTS ONLY] said, *"the feedback was repetitive"* while P89 [CONTROL] commented on

the repetitiveness of questions, *"same questions over and over"*. Overall, participants found various aspects of the system and the study to be repetitive regardless of the condition they were assigned.

In regards to the additional conversational content found in the ITEM VARIANTS PLUS condition, participants expressed differing opinions. 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]. On the other hand, 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"*. Although some participants found the humor and small talk a nice addition to the interaction with the agent, others found the VA's small talk to be ingenuine and took away from focusing on the task of answering the questionnaire. This shows that more care needs to be taken when incorporating humor and personal anecdotes in VA systems.

6 DISCUSSION

By conducting a two-week validation study, we were able to evaluate the psychometric properties and user engagement with the LLM-generated depression questionnaire item variants. A total of 105 participants completed one of the three study conditions (CONTROL, ITEM VARIANTS ONLY, and ITEM VARIANTS PLUS). Participants were tasked to interact with the agent daily and answer questions related to depression asked by the VA. 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 who were given the item variants found the questions coherent, natural, easy to understand, and relevant to the conversational topic based on median self-reported data being above neutral.

Although 105 participants started the study, only 93 participants met the minimum requirements for compensation. The CONTROL group had the largest percentage of participants (20%) who did not meet the minimum number of interactions. We saw a trending difference in the number of participants who met the minimum requirement of seven interactions among the three study conditions. In addition, participants in the CONTROL group found the system and the agent to be more repetitive compared to participants with the LLM-generate questionnaire variants based on our content analysis. However, we were not able to 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 requirements. 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 conversational humor and small talk will increase engagement compared to those without them in an interview with a VA.

Based on our qualitative findings, we discovered that several participants enjoyed interacting with the VA and responding to the questionnaire. Some participants even wanted to talk with the VA longer and with more follow-up questions while others expressed negative sentiments as they found the VA to be uncanny or unnatural. Furthermore, we observed that participants found their experience to be repetitive which can be due to various reasons such as the questions being the same, the dialogue flow being the same, the feedback from the VA being the same, or the limited

response options being the same. Participants who were given the humor and small talk expressed divergent opinions about the jokes and the stories. Some participants found them entertaining and interesting while others found them to be forced and inappropriate. This demonstrates how conversational humor and small talk might not always be a reliable mechanism to increase engagement or satisfaction in the longitudinal administration of questionnaires. The very simplistic nature of our LLM-generated jokes and humor could have affected our negative results, and further research is needed on better prompting strategies to generate higher-quality stories and jokes using LLMs.

6.1 Guidelines for Prompting LLMs to Generate Diverse Questions

When using LLMs to generate diverse questions, it is important to perform prompt engineering to obtain good prompts that will provide the best results. Different prompts can lead to substantially different results [37, 52, 67]. In this section, we provide some guidelines based on our learnings on how to use LLMs to generate diverse questions from established questionnaires.

Provide Clear, Detailed Instructions. When prompting, it is important to include clear and detailed instructions of what we want as output to the model. In our case, we had to explicitly prompt the model to output a given questionnaire item in question format. Furthermore, including adjectives such as "conversational" or "engaging" can produce very different outputs compared to not providing these adjectives when asking the model to generate variations of a given question. In addition, instructions including user context can be used to generate different types of variants to provide more personalized variations to questionnaires. For example, adding demographic information of the intended user in the prompt can provide more personalized variants. Appendix A provides actual prompts we used to generate the depression questionnaire variants and other conversational content for the VA.

Align with Response Scale. Standardized and validated questionnaires often have specific response scales. When creating item variants for these types of questionnaires, it is important to factor the wording of the response scale into the prompt. For the depression questionnaire we worked with, we provided the LLM with the 5-point scale to provide more guidance (e.g., "The question should be formatted for the following answers only: Never, Rarely, Sometimes, Often, Always"). Sometimes, we saw question variants that were negatively phrased and did not work with the fixed response scale that we had. For example, ChatGPT generated "How often do you feel excited about the future?" when prompted to generate a variant for "In the past 7 days, I felt that I had nothing to look forward to." Although we have not experimented with varying the response scale or options, we expect this task to be easily performed, given the right prompt.

Expect Duplicates. When asking LLMs to generate variations, it is common to find many duplicate outputs when using the same prompt multiple times to generate more outputs. Both `temperature` and `top_p` sampling are powerful tools for controlling the behavior of GPT models. These hyperparameters can be used independently or together when making API calls. By adjusting these parameters, you can achieve different levels of creativity and control. Increasing the `temperature` which can range from 0 to 2 generally provides more diverse outputs as this parameter controls the "creativity" or randomness of text generated. Increasing `top_p` which ranges from 0 to 1 can be another way of increasing text diversity. From our experience, we recommend using `temperature` to increase diversity and creativity. In our study, we used `temperature` = 0.7 and `top_p` = 1 for GPT-3 model as we saw the best results despite the duplicate outputs.

Refer to Experts. Although LLMs may offer benefits in various settings, including healthcare, they also pose many risks. LLMs can cause material harm by disseminating poor or false information through hallucinations [4, 33, 35]. When working in high-risk settings, such as healthcare, it is important to consult domain experts when filtering LLM-generated outputs. Prior work has shown that LLMs and conversational AI can cause physical harm by disseminating poor or false medical information [9, 19, 34, 43]. In our study, psychologists in our team assisted in filtering out potentially inappropriate or irrelevant item variants of the depression questionnaire. We recommend researchers to refer to domain experts when working with sensitive content and in high-risk settings.

In addition to hallucinations, LLMs can produce biased language and output text that can proliferate social stereotyping which may skew survey responses [65]. General purpose LLMs have access to vast generic vocabulary but often struggle with domain-specific terms [24, 32, 44]. This can result in inappropriate or confusing language for surveys in specific domains such as healthcare. It is important to take these limitations of LLMs into consideration and do due diligence when reviewing LLM-generated questionnaire items.

6.2 Limitations

There are several limitations to our study beyond the small convenience sample used. We conducted our evaluation study using only one standardized questionnaire (PROMIS depression), so it is unclear whether our results hold for other standardized questionnaires used in healthcare or HCI research. Furthermore, our study was conducted with a virtual agent administering the questionnaires, and it is unclear whether our results hold for other methods of questionnaire administration such as paper or online form-based surveys.

In this study, we only collected data for two weeks of use. Longer-term patterns of use with questionnaire item variants generated by LLMs remain to be explored. We do recognize that our compensation structure may have affected the results of engagement and interaction with the agent and can be seen as a limitation of our study design. In addition, the technical limitations of the prototype such as unnatural-sounding voice and glitches could have affected the satisfaction of participants.

6.3 Future Work

Future works should consider ways of adding more variations in dialogue structure, VA feedback to users, and user response options which would be beneficial and reduce repetitiveness. Based on participants' suggestions, some participants would like to have a more in-depth conversation with the agent in terms of both follow-up questions related to their feelings or daily chatting. Likewise, future work could study the effects of generating more response options for the answers or letting the participants interact with the agent using an unstructured input. However, finding a balance between various personalizations and standardizations would need further examination.

Further research on how to effectively incorporate humor and anecdotes generated by LLMs for longitudinal VA research should be considered. In our study, we only included very simple jokes and stories from the agent's perspective that did not resonate with everyone. Having more diverse jokes and stories (backstories of the agent or even stories of real people) can be an interesting future direction.

Previous studies [27, 39] have compared using VA and paper-based questionnaires, and for our future studies, we will compare the effect of VA-based interactions with item variants with paper-based questionnaires. Furthermore, future studies could examine utilizing more novel approaches, such as logical control [6], chain-of-thought prompting [64], or augmentation to use external tools [41], to create a safeguard for hooking up LLMs more directly to conversational agents.

7 CONCLUSION

Our results show that LLM-generated item variants of 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 interacted with the agent more frequently over the two-week study period compared to the CONTROL group. Furthermore, qualitative responses indicated that participants found this experience to be less repetitive. Nevertheless, there were no significant differences between the two ITEM VARIANTS conditions; one of them included conversational humor and small talk alongside the item variants. Striking a balance between personalization and standardization will be crucial for maintaining high-quality data collection and boosting response rates in the context of 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.

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] 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.

[3] 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.

[4] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[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] 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.

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

[14] 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.

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

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

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

[18] 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.

[19] Mindy Duffourc and Sara Gerke. 2023. Generative AI in Health Care and Liability Risks for Physicians and Safety Concerns for Patients. *JAMA* (2023).

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

[21] 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.

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

[23] 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.

[24] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–23.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Maria Kjærup, Mikael B Skov, Peter Axel Nielsen, Jesper Kjeldskov, Jens Gerken, and Harald Reiterer. 2021. *Longitudinal studies in HCI research: a review of CHI publications from 1982–2019*. Springer.

[30] 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.

[31] 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.

[32] Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do We Still Need Clinical Language Models? *arXiv preprint arXiv:2302.08091* (2023).

[33] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. *arXiv e-prints* (2023), arXiv–2305.

[34] Baihan Lin, Djallel Bouneffouf, Guillermo Cecchi, and Kush R Varshney. 2023. Towards Healthy AI: Large Language Models Need Therapists Too. *arXiv preprint arXiv:2304.00416* (2023).

[35] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).

[36] 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.

[37] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786* (2021).

[38] 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.

[39] 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.

[40] 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.

[41] 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).

[42] 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.

[43] 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.
[44] Danila Morozovskii and Sheela Ramanna. 2023. Rare words in text summarization. *Natural Language Processing Journal* 3 (2023), 100014.
[45] 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.
[46] S Olafsson, T O'Leary, and T Bickmore. 2020. Motivating Health Behavior Change with Humorous Virtual Agents.
[47] 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.
[48] 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.
[49] 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.
[50] 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).
[51] 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.
[52] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.
[53] 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.
[54] Michael F Schober and Frederick G Conrad. 1997. Does conversational interviewing reduce survey measurement error? *Public opinion quarterly* (1997), 576–602.
[55] 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.
[56] 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.
[57] Angela Sinickas. 2007. Finding a cure for survey fatigue. *Strategic Communication Management* 11, 2 (2007), 11.
[58] Kirsten P Smith and Nicholas A Christakis. 2008. Social networks and health. *Annu. Rev. Sociol* 34 (2008), 405–429.
[59] 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.
[60] 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.
[61] Laura Vardoulakis. 2013. Social desirability bias and engagement in systems designed for long-term health tracking. (2013).
[62] Vasja Vehovar, Zenel Batagelj, Katja Lozar Manfreda, and Metka Zaletel. 2002. Nonresponse in web surveys. *Survey nonresponse* (2002), 229–242.
[63] 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.
[64] 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.
[65] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.
[66] 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.
[67] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*. PMLR, 12697–12706.

A PROMPTS FOR LLMS

We used GPT-3 and ChatGPT to generate the questionnaire item variants and various conversation content for the agent. We attempted various prompting strategies before finalizing the prompts. In our study, we used the hyperparameters `temperature` = 0.7 and `top_p` = 1 for GPT-3 model as we saw the best results despite some duplicate outputs when we made up to 20 requests per prompt. The actual prompts we used for the study are provided below.

A.1 Prompts for Generating Item Variants

GPT-3

> Statement: "[Original Item]"
> Ask the above statement as a long question. Also, change the time from past 7 days to from yesterday. The answer of the question should be between: Never, Rarely, Sometimes, Often, Always.
> Question:

ChatGPT

> Statement: "[Original Item]"
> Give me ten different conversational questions which tries to measure the same concept in the statement. Also, ask it from yesterday or since the last time we talked.
> The question should be formatted for the following answers only: Never, Rarely, Sometimes, Often, Always.

A.2 Prompts for Generating Stories

For generating personal anecdotes, we chose 15 general everyday topics to prompt the LLMs. The 15 topics were: "grocery shopping", "commuting to work", "having lunch", "making dinner", "going to a party with a friend", "watching a basketball game", "trying a new restaurant the past weekend", "reading a new book", "going to a live theatre show", "hosting a potluck", "having a boring day at work", "spilling coffee all over the place while getting ready for work", "feeling a bit under the weather", "going to the dentist", and "going for a hike."

GPT-3

> Give an engaging, interesting first person story of a character on the topic of [Topic] that can be shared with someone in 3 sentences.

ChatGPT

> Marie is a character who is responsible for asking different questions of other people's mood and depression symptoms. Please give me a backstory for Marie in first person.

A.3 Prompts for Generating Jokes

GPT-3

> Marie is a computer animated agent who is responsible for asking different questions of user's mood and depression symptoms. She asks serious questions everyday and would like to use humor to make the conversation more engaging and uplifting. What joke should Marie tell?

> What funny joke can I tell to make a very serious conversation more uplifting?

> Give a short joke that a mental health counselor can tell to a patient.

GPT-3 & ChatGPT

> Tell me a short joke.

> Tell me a really funny computer-related joke.

> Give me a joke about computer animated character.

A.4 Prompts for Generating Empathetic Responses

The [Original Item Content] refers to the main content of the original question or item. For example: if the original item is "In the past 7 days, I felt worthless.", we would prompt the model with "Provide a short empathetic response for someone who never or rarely feels worthless."

GPT-3 & ChatGPT

> Provide a short empathetic response for someone who never or rarely feels [Original Item Content].

> Provide a short empathetic response for someone who often or always feels [Original Item Content].

> Provide a short empathetic response for someone who sometimes feels [Original Item Content].

A.5 Prompts for Generating Motivating Messages

GPT-3 & ChatGPT

> Give me a short inspiring message for the day.

> Share a short uplifiting quote for me.

A.6 Prompts for Generating Farewells

GPT-3 & ChatGPT

What can I tell someone to signal an end to a conversation before saying good bye?

How can I end a conversation in a nice way before saying good bye to someone who I was talking to?

Marie is a character who met someone and have had a good conversation. Give me a short thing that Marie should say before saying good bye in first person.

I am feeling very low and down today. Give me a short uplifting message.

Marie is a character whose job is to only check in on different clients' moods and feelings and does not provide advice. Give me a short script that Marie can use before saying good bye to one of her clients in first person.