Keeping Users Engaged During Repeated Interviews by a Virtual Agent

Using Large Language Models to Reliably Diversify Questions

Hye Sun Yun, Mehdi Arjmand, Phillip Sherlock, Michael Paasche-Orlow, James W. Griffith, Timothy Bickmore

IVA 2024

Self-Report Questionnaires

Source: http://www.aldenhampsychology.com/self-reports.html

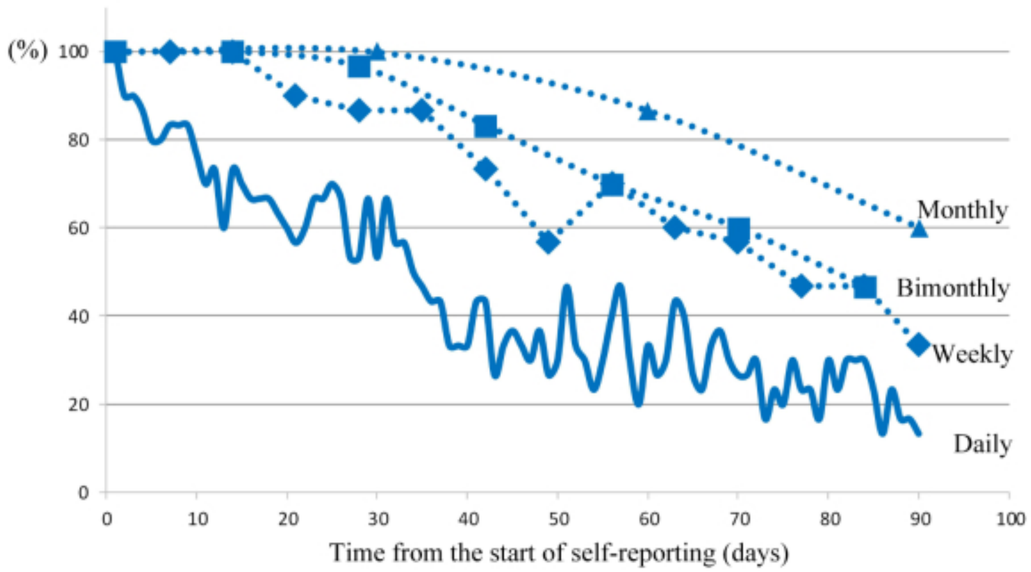
Virtual Agent-Administered Questionnaires

- Virtual agent-administered questionnaires **=** self-administered questionnaires (Jaiswal et al., 2019; Bickmore et al., 2020)
- Studies have shown the feasibility and reliability of using virtual agents (VAs) to administer questionnaires simulating interviews for **a single session**

Source: Jaiswal et al., 2019

Engagement

Repeated-Measures Evaluation



Source: Min et al., 2014

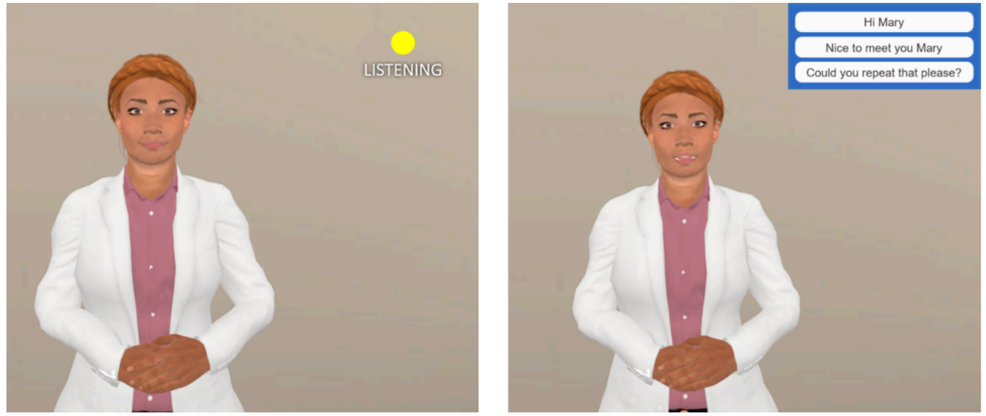
- Patient-Reported Outcomes (PROs)
- Fatigue leads to declining response rates over time (Porter et al., 2004; Min et al., 2014; Dean & Crittenden, 2016)
- PRO longitudinal survey completion rates **can be as low as 48%** (Min et al., 2014; Dean & Crittenden, 2016; Huynh et al., 2021)
  - nonresponse measurement bias (Groves & Peytcheva, 2008)

Large Language Models

For Creating Engaging Dialogue Content

- **Scenarios** (Antunes et al., 2023) and **dialogue utterances** (Hanschmann et al., 2023; Sevilla-Salcedo et al., 2023; Olafsson et al., 2023) for agents
- **Diverse texts or paraphrases** in a scalable way while preserving the original meaning (Yu et al., 2023; Cox et al., 2023; Pehlivanoğlu et al., 2023)

Source: Hanschmann et al., 2023



Source: Olafsson et al., 2023

Research Questions

1. Will VA administration of LLM-generated item **variants retain similar validity and reliability** to the VA administration of the original questionnaire?

2. Are questionnaires delivered in a different form **using LLM-generated variants daily more engaging for participants**, based on the number of questionnaires completed and feedback from participants?

3. Are **questionnaires delivered with LLM-generated conversational small talk, humor, and empathy more engaging** compared to those delivered as strictly question-and-response interviews by a VA?

PROMIS® Depression Questionnaire

Short Form

- self-report PRO questionnaire using the eight-item PROMIS® short form depression questionnaire (version 8a) (Cella et al., 2010)
- assess a respondent's level of emotional distress caused by depressed mood
  - a five-point scale from 1 = "Never" to 5 = "Always"

In the past 7 days...

	Never	Rarely	Sometimes	Often	Always
I felt worthless	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
I felt helpless.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

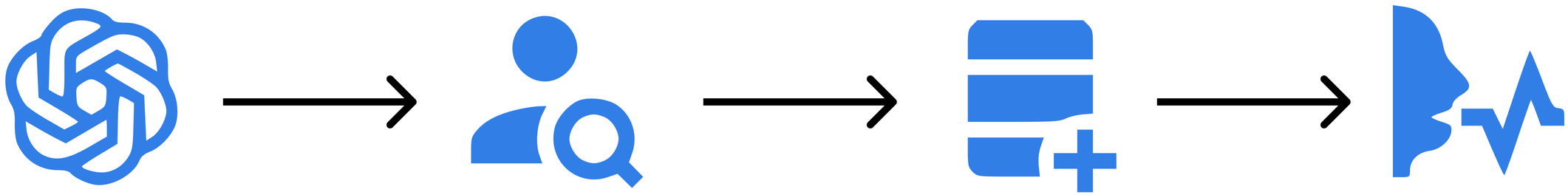
System Design

Generate Diverse Questions with LLMs

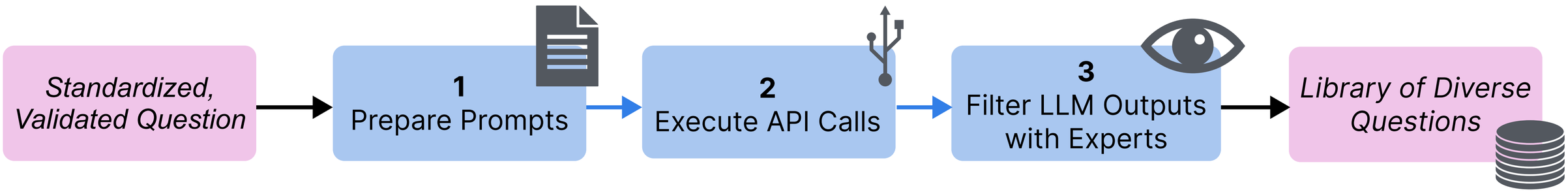
Manually Review Outputs

Add Item Variants to Database

Develop Agent System with Variants



Item Variants with LLMs



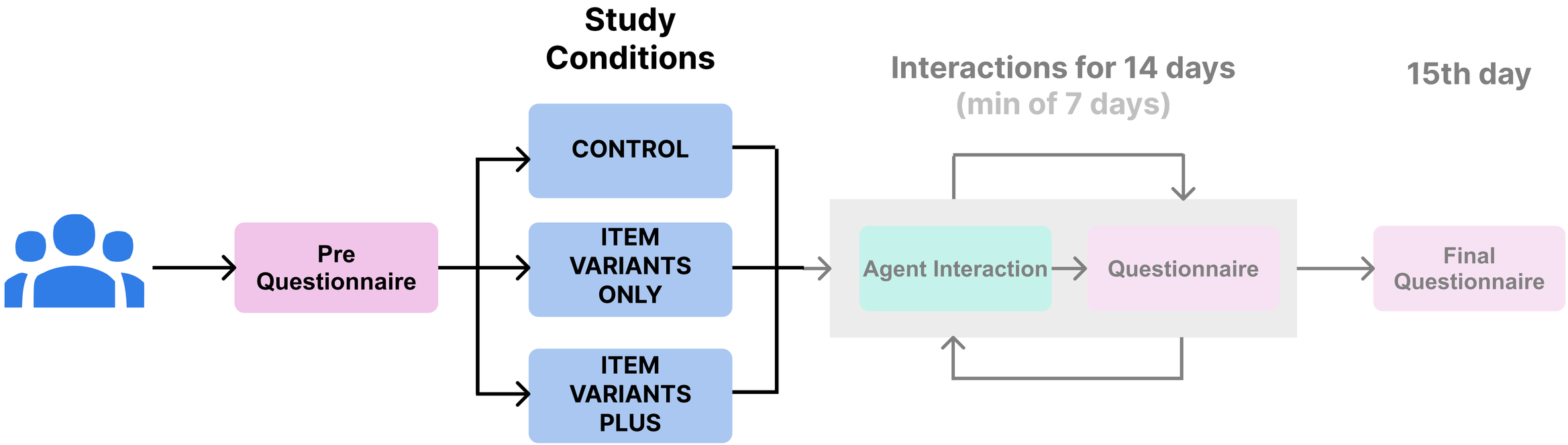
Examples

Original	Sample Variant	# of Variants
In the past 7 days, I felt worthless.	Since we last spoke, have you ever felt like you were a burden to others?	8
In the past 7 days, I felt helpless.	How often have you felt like you were unable to control a situation in the past day?	7

Agent

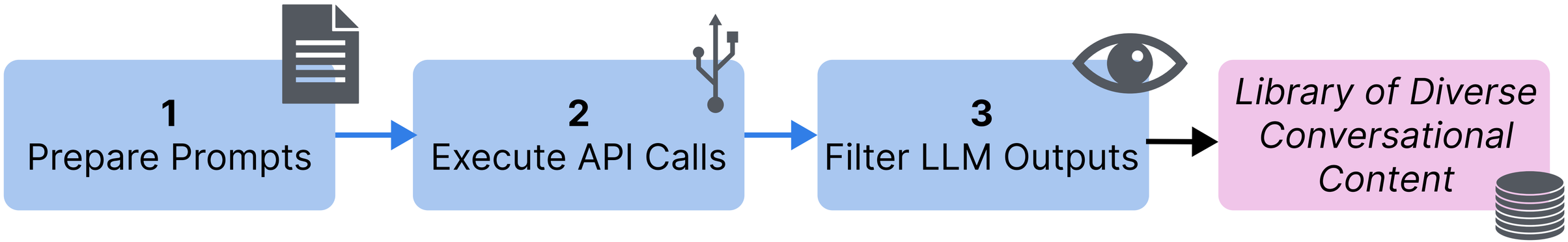

Never
Rarely
Sometimes
Often
Always
I would like to SKIP
Could you repeat that?

Longitudinal Validation Study



Conversational Contents with LLMs

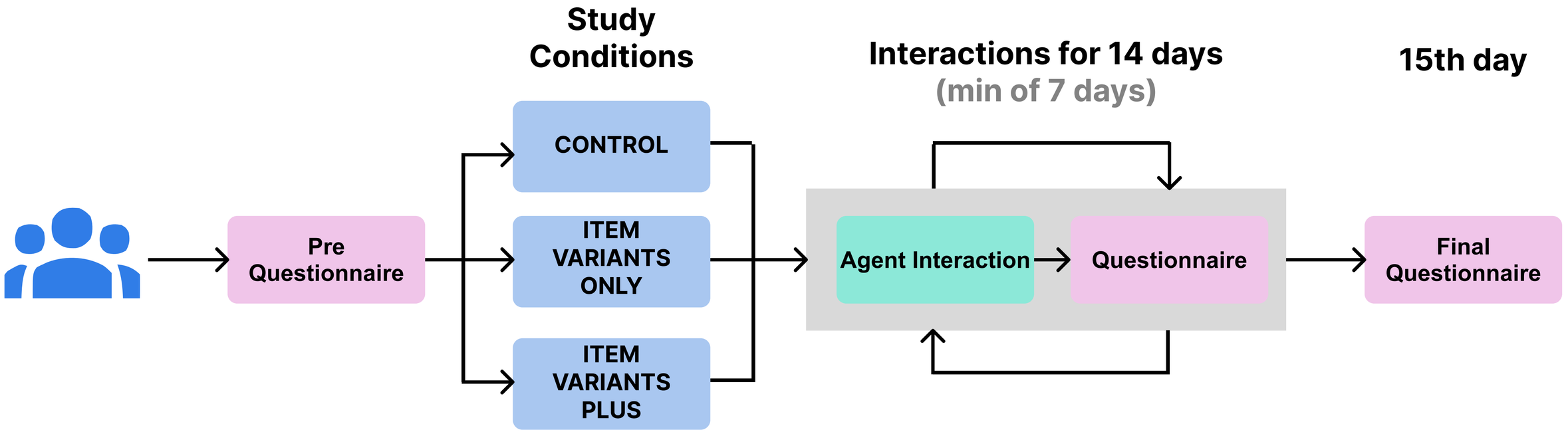
Stories, Jokes, Empathetic Responses, Messages, & Farewells



Examples

Category	Example	# of Unique Content
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!	37
Jokes	Why did the smartphone need glasses? Because it lost all its contacts!	24

Longitudinal Validation Study



Participants

- **105** total participants were recruited via Prolific
  - 35 per study condition
- **Age**: Mean = 39, *SD* = 12
- **Gender**: women = 49.5%, men = 46.7%, non-binary 2.9%, & others = 1.0%
- **Education**: all had at least a high school degree or equivalent
- **Depression Therapy or Medication**: “No” = 80.0%, “Yes” = 19.1% said “yes”, & preferred not to answer = 1.0%

Psychometric Properties

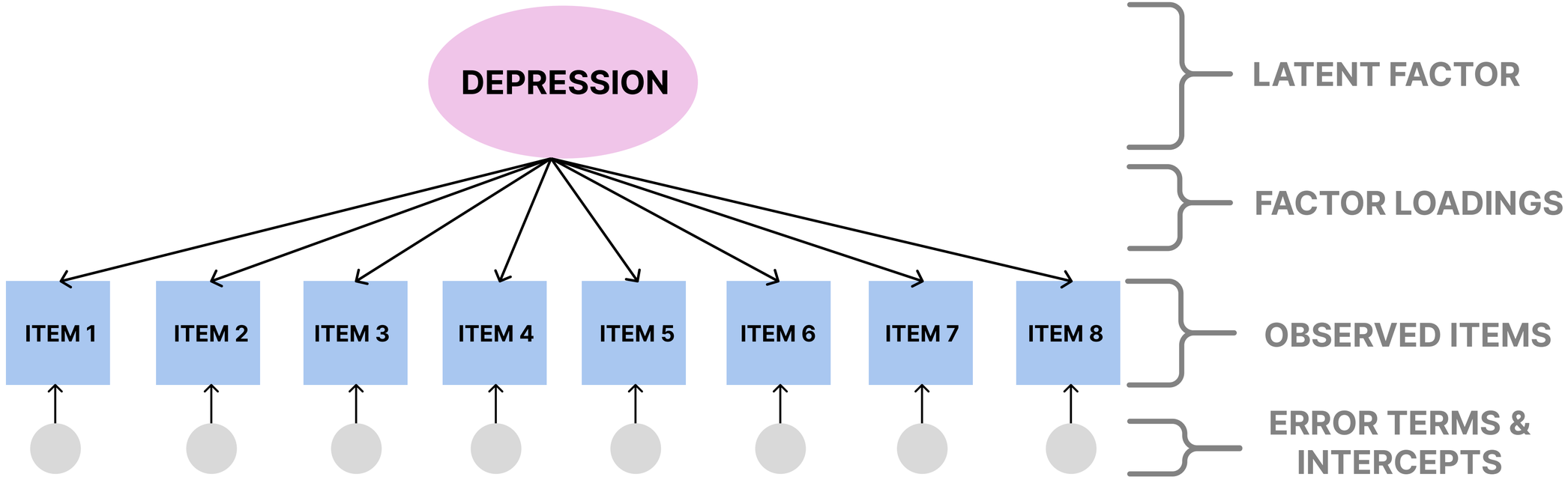
Internal Consistency

- Internal Consistency/Reliability of 8 Depression Questions
  - Cronbach's alpha
    - CONTROL (*original in daily question format*): α **= 0.76**
    - ITEM VARIANTS (*LLM-generated variants*): α **= 0.65**

Psychometric Properties

Consistency Across 3 Study Groups

- measurement alignment analysis (Han, 2024)
  - method for multiple-group confirmatory factor analysis (CFA)



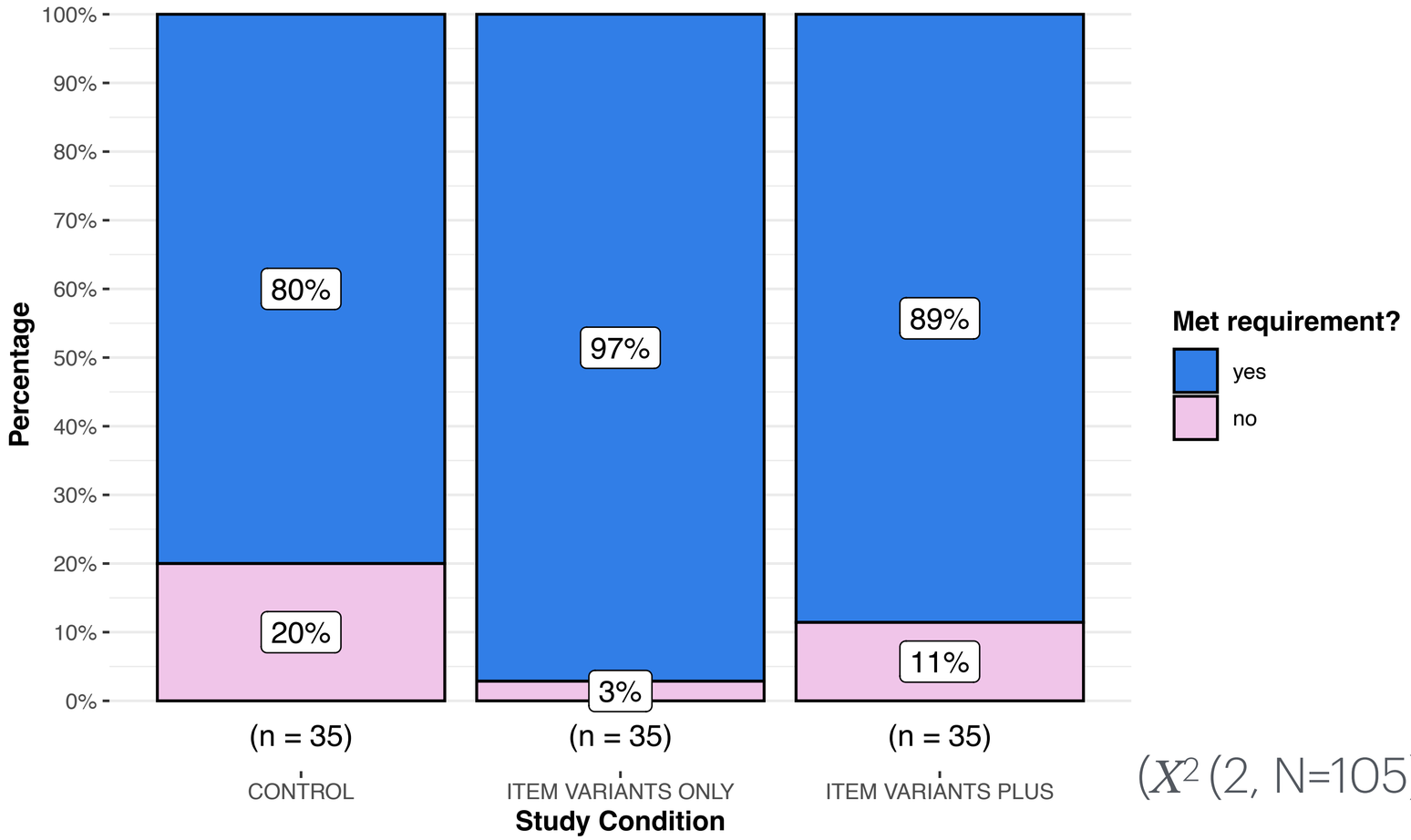
Psychometric Properties

Consistency Across 3 Study Groups

- $R^2 \geq$ **0.98** (*reliable & trustworthy alignment results*)
- External Criterion - PHQ-8 (Kroenke et al., 2009; Razykov et al., 2012)
  - correlations between the PROMIS® questionnaire and the PHQ-8 $\geq$ **0.80 across all study conditions**

Engagement

Minimum Interaction Requirement



(X^2 (2, N=105) = 5.1, ***p* = .08**)

User Perceptions

System

Item	CONTROL	ITEM VARIANTS ONLY	ITEM VARIANTS PLUS
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

User Perceptions

Agent

Item	CONTROL	ITEM VARIANTS ONLY	ITEM VARIANTS PLUS
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

User Perceptions

Questions

Item	CONTROL	ITEM VARIANTS ONLY	ITEM VARIANTS PLUS
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

Content Analysis

- Mentions of **“repetitiveness”** in open-ended responses
  - CONTROL vs two VARIANTS groups
    - X^2 (1, *N*=93) = 5, ***p*=.029**

Qualitative Analysis

Comforting vs Uncanny Agents

“I like how someone was checking in with me daily to make sure I was alright”
[P43 - ITEM VARIANTS PLUS]

“The attempt to make the robot AI feel human looking—it was uncanny valley to the max”
[P80 - CONTROL]

Qualitative Analysis

Various Reasons for Repetitiveness

“How repetitive the responses were…”
[P88 - CONTROL]

“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]

“The feedback was repetitive”
[P66 - ITEM VARIANTS ONLY]

Qualitative Analysis

Humor and Small Talk Does Not Always Work

“[Favorite part was] hearing the jokes she had”
[P85]

“[Wish I could skip] the bad dad jokes”
[P11]

“Probably the ‘let me tell you about myself’ stupidity. It was ridiculously patronizing that I was expected to take that seriously.”
[P87]

Conclusion

1. Will VA administration of LLM-generated item **variants retain similar validity and reliability** to the VA administration of the original questionnaire? **YES**
2. Are questionnaires delivered in a different form **using LLM-generated variants daily more engaging for participants**, based on the number of questionnaires completed and feedback from participants? **MAYBE**
3. Are **questionnaires delivered with LLM-generated conversational small talk, humor, and empathy more engaging** compared to those delivered as strictly question-and-response interviews by a VA? **NO**

A step forward in integrating LLMs into VAs to diversify and enhance questionnaire administration while maintaining validity and reliability

Thank you!
Any questions?

yun.hy@northeastern.edu

@hyesunyun

Data & Code