Leveraging Interview-Informed LLMs to Model Survey Responses: Comparative Insights from AI-Generated and Human Data

Jihong Zhang
University of Arkansas
Fayetteville, U.S.
jzhang@uark.edu

Xinya Liang
University of Arkansas
Fayetteville, U.S.
xl014@uark.edu

Anqi Deng
University of Arkansas
Fayetteville, U.S.
adeng@uark.edu

Nicole Bonge
University of Arkansas
Fayetteville, U.S.
ngbonge@uark.edu

Lin Tan
University of Arkansas
Fayetteville, U.S.
lintan@uark.edu

Ling Zhang
University of Wyoming
Laramie, U.S.
lingzhangelaine@gmail.com

Nicole Zarrett
University of South Carolina
Columbia, U.S.
zarrettn@mailbox.sc.edu

Mixed methods research integrates quantitative and qualitative data but faces challenges in aligning their distinct structures, particularly in examining measurement characteristics and individual response patterns. Advances in large language models (LLMs) offer promising solutions by generating synthetic survey responses informed by qualitative data. This study investigates whether LLMs, guided by personal interviews, can reliably predict human survey responses, using the Behavioral Regulations in Exercise Questionnaire (BREQ) and interviews from after-school program staff as a case study. Results indicate that LLMs capture overall response patterns but exhibit lower variability than humans. Incorporating interview data improves response diversity for some models (e.g., Claude, GPT), while well-crafted prompts and low-temperature settings enhance alignment between LLM and human responses. Demographic information had less impact than interview content on alignment accuracy. Item-level analysis revealed higher discrepancies for negatively worded questions, suggesting LLMs struggle with emotional nuance. Person-level differences indicated varying model performance across respondents, highlighting the role of interview relevance over length. Despite replicating individual item trends, LLMs faltered in reconstructing the test's psychometric structure.

These findings underscore the potential of interview-informed LLMs to bridge qualitative and quantitative methodologies while revealing limitations in response variability, emotional interpretation, and psychometric fidelity. Future research should refine prompt design, explore bias mitigation, and optimize model settings to enhance the validity of LLM-generated survey data in social science research. The R code and the supplementary materials are available on the OSF platform (DOI:10.17605/OSF.IO/AFQG3).

Keywords: quantitative data, qualitative data, LLM-driven interview, survey, behavioral regulations in exercise

---

Correspondence concerning this article should be addressed to Jihong Zhang, Department of Counseling, Leadership, and Research Methods, University of Arkansas, 109 Graduate Education Building, Fayetteville, AR 72703, Email: jzhang@uark.edu

1. INTRODUCTION

Mixed methods design is a widely used research approach in psychology and education (Bishop, 2015; Johnson et al., 2007; Powell et al., 2008) due to the complementary strengths of integrating both quantitative and qualitative approaches. The framework of mixed methods study design requires the rigorous collection and analysis of both quantitative and qualitative data to answer research questions and test hypotheses (Creswell & Clark, 2017). However, because of the structural differences between qualitative (e.g., open-ended responses, textual richness, context-dependent interpretations) and quantitative (e.g., numeric ratings, standardized scales, fixed-response options; Schoonenboom, 2023) data, comparative analysis between quantitative and qualitative information—especially for examining measurement characteristics such as response formatting or question wording, as well as person-level attributes such as individual response styles and subjective interpretations—remains challenging, which hinders the broader adoption of mixed methods design (J. Wang et al., 2024). With the advent of large language models (LLMs) in natural language understanding tasks, bridging quantitative and qualitative insights has become increasingly feasible.

Generative artificial intelligence (GenAI), particularly LLMs, enables researchers to analyze survey data, automate scoring in educational assessment, and conduct large-scale social simulations by generating synthetic personas (A. Li et al., 2025). Personas are generalized representations of targeted users based on real users' data. The underlying assumption is that, when provided with relevant context (e.g., a scoring rubric or demographic information), LLMs can simulate personas (e.g., role information of teachers, survey respondents, interview coders) to represent population user profiles, which can produce population-level synthetic opinions that approximate those of real-world target populations. Researchers can then use these synthetic opinions to predict real-world responses. For instance, prior studies have relied on census data to generate personas that reflect varying political ideologies, which were subsequently used to predict survey responses on political attitudes and improve measurement accuracy by analyzing discrepancies between predicted and observed data (Argyle et al., 2023; Chang et al., 2024; Yu et al., 2024).

Prior research in this area has primarily focused on improving the representation of LLM-generated responses for various populations to improve the generalizability of the LLMs' output. This process often relies on large sample sizes in the training process of LLMs (A. Li et al., 2025). However, in real-world settings, the vast majority of studies in education and psychology routinely operate with relatively small sample sizes (Slavin & Smith, 2009), a challenge that remains underexplored in LLM literature. In addition, although Likert-scale surveys are widely

used in the social sciences, limited research has examined the performance of LLMs in generating valid Likert-scale responses (Liu et al., 2024) and to our knowledge, no research has examined the consistency or similarity between human survey responses and LLM-generated responses when rich individual-level interview data informs the LLMs.

In light of the growing interest in incorporating qualitative insights with LLM-driven survey methods, the present study proposes an interview-informed LLM approach for evaluating qualitative and quantitative data to provide a more comprehensive understanding of participants' behaviors. Our approach aligns with the convergence design model within mixed methods research (Ponce & Pagán-Maldonado, 2015), in which quantitative and qualitative approaches are integrated simultaneously to provide a comprehensive understanding of participants' experiences. Within this framework, LLM-generated survey responses serve as an analytical tool for interpreting and validating participants' interview and measurement data. Humans may be inconsistent between interviews and questionnaires due to factors such as emotions (e.g., feeling bored) or measurement issues (e.g., items not accurately reflecting participants' true thoughts). This is one reason LLMs can be valuable in mixed-method scenarios. Our approach leverages the richer, more nuanced information present in the interviews—which is widely used but often difficult to extract in social science studies—to generate synthetic survey responses. This allows us to examine how well LLMs can infer structured responses from unstructured, real-world data. These structured responses can then be compared with participants' actual questionnaire responses to further understand their characteristics.

We illustrate this approach using the *Behavioral Regulations in Exercise Questionnaire* (BREQ; Cid et al., 2012) and interviews from after-school program staff as an example. Our study aims to investigate whether LLMs can leverage human interview data to capture human response patterns in terms of both alignment and diversity. We define alignment as fidelity of reconstruction: the degree to which interview-grounded, item-level responses preserve the latent trait structure implied by participants' narratives, so that interview-informed LLMs can serve as a proxy for human responses. Operationally, alignment reflects how closely LLM-generated survey responses correspond to human survey responses when both are informed by the same interview-based insights. In addition, "diversity" refers to the extent of variation or heterogeneity in the LLM-generated survey responses when producing multiple outputs from the interview data across participants. Examining the alignment and diversity of generated responses enables researchers to capture the range and nuance of individual experiences via LLMs, supporting a more comprehensive understanding of both consistencies and discrepancies across qualitative and quantitative measures. Our novel interview-informed LLM approach allows researchers to: (1) provide a framework for triangulating qualitative and quantitative findings, (2) identify potential inconsistencies between qualitative narratives and quantitative self-reports, and (3) identify response patterns and prioritize items for review.

This study is organized as follows: the *Background* section reviews prior work on LLMs to generate survey responses; the *Current Study* section outlines the research purpose and guiding questions; the *Study Design* section describes the data, experimental settings, and LLM implementation; the *Results* section compares LLM-generated outputs with human survey data; and the *Discussion* section explores the empirical implications and suggests directions for future research.

2. BACKGROUND

LLMs have been investigated in educational and psychological assessment for varied purposes (May et al., 2025). Recent studies have demonstrated the versatility of LLMs in extracting insights from unstructured texts, either by treating unstructured texts as sequences of natural language processing tasks (Parker et al., 2024) or by leveraging multi-agent frameworks for natural language understanding and simulation (Rasheed et al., 2024). However, P. Wang et al. (2024) caution that although LLMs effectively capture broad patterns, they often struggle with item-level nuances and subtle individual differences. Additionally, the methods and effectiveness of using LLMs to generate human-like quantitative responses remain inadequately examined.

2.1. PERSONA-DRIVEN METHOD

The most widely used approach to simulating survey responses with LLMs centers on the persona-driven method (Jiang et al., 2023; A. Li et al., 2025; Liu et al., 2024). *Synthetic persona generation* (Chen et al., 2024) involves creating artificial user profiles for applications such as political attitude modeling and economic forecasting. These synthetic personas are typically generated by sampling demographic attributes (e.g., age, race, income) from census or survey distributions. Other data-driven techniques of persona generation include clustering respondent data, conducting factor analysis, or applying matrix decomposition methods to identify archetypal personas from extensive population datasets (Jansen et al., 2022). An alternative approach, termed *descriptive personas*, does not rely on real-world demographic attributes. Instead, descriptive personas utilize synthetic text inputs to condition LLMs to directly generate diverse personas. These inputs are commonly derived from extensive corpora of public web texts, allowing the models to simulate a broader spectrum of social perspectives and behavioral traits (Ge et al., 2024). This study employs a novel persona-based LLM method by combining research context, real-world personal interview data, and demographic information.

In our scenarios, persona-based predictions can be understood as a form of probabilistic profiling. The model infers likely survey responses based on patterns in the prompt text and selects the most probable option. This process aims to produce survey responses consistent with the traits and perspectives reflected in the qualitative narrative.

2.2. APPLICATION OF LLM-GENERATED RESPONSES

In addition to studies on persona generation, researchers have devoted increasing attention to using LLMs to generate human-like quantitative survey responses for various purposes (Liu et al., 2024; Xu & Zhang, 2023). These LLM-generated responses offer a cost-effective solution for data augmentation or prediction, reducing the need for time-intensive human data collection. The use of LLMs as substitutes for human participants in research contexts has generated considerable debate within the academic community. Proponents argue that LLMs offer cost-effective, scalable alternatives for pilot testing, data augmentation, and hypothesis generation, particularly in domains where human data collection is resource-intensive or ethically challenging (Huang, Wang, et al., 2024; Y. Li et al., 2024; Liu et al., 2025; Serapio-García et al., 2023). In this context, some studies viewed LLMs' generated responses as *silicon samples* (also referred to as "synthetic datasets") processing emergent personality characteristics shaped by prompts and training data (Sarstedt et al., 2024; Sun et al., 2024). Others minimize the personality traits of LLMs, instead conceptualizing LLMs as efficient computational tools for

measurement-development tasks, such as item calibration (Ding et al., 2024), automated scoring (Mendonça et al., 2025; Uto & Uchida, 2020), one-on-one tutoring (Fateen & Mine, 2025), cheating detection (Wang & Li, 2025), and item generation (Laverghetta & Licato, 2023).

However, some researchers raise concerns about the ecological validity of LLM-generated responses (e.g., Wang, P. et al., 2024), the potential for perpetuating biases embedded in training data (e.g., Shojaee et al., 2025), and the fundamental differences between human cognitive processes and LLM text generation mechanisms (e.g., Mancoridis et al., 2025). Our work contributes to this ongoing discussion by providing empirical evidence on the alignment between LLM-generated and human survey responses when both are informed by the same qualitative interview data. Rather than positioning LLMs as direct substitutes for human participants, our approach explores their potential as complementary tools for understanding response patterns, identifying measurement content validity, and enhancing the interpretability of mixed-methods research findings. This perspective aligns with emerging literature that advocates for thoughtful integration of LLMs into research workflows while maintaining rigorous validation against human data and acknowledging the limitations inherent in current AI systems.

2.3. LLM Variants and Configurations

Key factors that may influence the performance of LLM-generated survey responses include the choice of LLM-based chatbot, temperature settings, and prompt configurations. LLM-based chatbots are AI chatbots developed and trained by different companies. The current literature provides limited comparative analyses of these chatbots in the context of generating educational or psychological survey responses. Many existing studies employ only a single chatbot for simulation purposes. However, prior research suggests that investigations into LLM capabilities should extend beyond a single model (e.g., GPT) to include other AI chatbots for robust comparative analysis (e.g., Agarwal et al., 2023; Lozić & Štular, 2023; Wu et al., 2023). The release of GPT-4 by OpenAI in 2023 attracted global attention due to its strong test-taking performance (Nori et al., 2023; OpenAI et al., 2023). Similarly, Claude 2, released by Anthropic in June 2023, gained interest for its extended context window (up to 100,000 tokens), significantly expanding the working memory of LLMs (Anthropic, 2025). Gemini, developed by Google, has also demonstrated high performance on medical benchmarks (Saab et al., 2024). Furthermore, several open-source LLMs have proven effective across domains. In the context of survey response generation (commonly referred to as silicon sampling), widely used chatbots include GPT models (e.g., GPT-3 Turbo, GPT-4; Sarstedt et al., 2024) and Llama models (e.g., Llama-3; Peng et al., 2024).

LLM parameters and prompt design represent two additional important factors in LLM simulation studies. For instance, Sarstedt et al. (2024) discuss how model parameters (e.g., temperature) and prompt tuning affect the quality of silicon samples. Temperature is a parameter of LLMs used to control the degree of randomness when choosing tokens. A temperature of 0 yields highly deterministic outputs by consistently selecting the highest-probability tokens, resulting in minimal variation across runs. In contrast, higher temperatures introduce greater variability and creativity in generated responses. However, it is important to note that an excessively high temperature could degrade response quality faster than it adds originality. In addition, prompt-tuning is an important process of LLM-based generation that may impact outputs. Minor changes in the prompt can be pivotal in whether the AI tool fails or excels to understand the instruction (Ekin, 2023; Federiakin et al., 2024). Neither temperature nor prompt settings are well examined in the survey response generation literature.

2.4. CURRENT STUDY

The current study's purpose is to examine the performance of the proposed interview-informed LLMs by evaluating the alignment between generated quantitative survey responses and real human responses. Human survey data are treated as the reference measurement for this study's comparisons, while interview-informed LLM responses serve as a structured, interview-grounded proxy. This study has three general goals: (1) to understand the variability of LLM-generated survey responses across LLM chatbots, temperature settings, and prompt settings; (2) to identify key factors that affect alignment between LLM-generated and human responses; and (3) to explore measurement-level and person-level characteristics by comparing LLM-generated responses to human responses. In this study, measurement-level and person-level characteristics (e.g., demographics, interview length, item wordings) are defined as item-level and person-level information relevant to measurement and respondents that may have impacts on the alignment between LLM-generated responses and human responses. It should be noted that the LLM methods do not aim to replace human judgment; rather, they help statistically quantify the alignment and standardize the procedure.

To achieve these goals, we designed a simulation study to assess the feasibility of using LLMs as an evaluation tool. This study utilizes the real-world interview data and survey responses collected from adult afterschool program (ASP) personnel who were participating in the Connect through PLAY program, a randomized controlled efficacy trial, designed to support changes in the physical activity (PA) values, motivations, and behavior of ASP staff/teachers as a means for establishing sustainable social changes in the school setting for increasing the daily PA of underserved youth. The mixed methods research design that involved collecting complimentary survey and interview data made it an ideal data set to test this study's primary aims. The specific research questions of this study are as follows:

1. How do survey responses generated by LLMs vary in terms of means and variability of item responses across different settings, such as LLM chatbots, temperature settings, and prompt configurations?
2. How do factors such as LLM chatbots, prompt settings, and temperature settings influence the alignment between LLM-generated and human responses?
3. What do discrepancies between LLM-generated and human responses indicate about measurement-level and person-level characteristics?

3. METHOD

3.1. DATA

This study employed the Behavioral Regulation in Exercise Questionnaire (BREQ; Mullan et al., 1997; Mullan & Markland, 1997; Wilson et al., 2002) and semi-structured interviews as the primary instruments to assess health-based behavioral regulation among after-school program (ASP) staff and program directors (N = 55) across 10 ASPs serving underserved youth in the Southeastern United States from 2023 to 2024. Specifically, qualitative data were collected through semi-structured interviews conducted between researchers and ASP staff, each lasting approximately 15–25 minutes and subsequently transcribed for analysis. The staff semi-structured interview instrument is designed to assess ASP staff's physical activity (PA) experiences, perceptions, and readiness to implement new programming, particularly in underserved communities (see interview questions in the Supplementary). It captures key constructs including community engagement, job satisfaction, motivation towards PA, perceived support, staff

understanding of youth experiences and approaches to youth development, staff's beliefs and attitudes about youth's physical and mental health, and staff's perceptions of stress among both youth and staff. Additionally, it explores readiness for future program implementation by identifying potential motivators, barriers, and support needs, including professional development, incentives, and access to resources. In addition to the interviews, ASP staff were asked to complete a series of self-report questionnaires, including the BREQ. The BREQ instrument is used to assess the motivation behind exercise behavior. For the present analysis, data included responses from 19 ASP staff members who completed both the interview and the questionnaire (N = 19). All interview transcripts were de-identified by removing or masking personal information prior to analysis. The final sample consisted of 19 participants with a mean age of 35.5 years (range: 18-61 years). The majority of participants identified as female (84.2%, n = 16), with two male participants (10.5%, n = 2) and one participant missing gender information. Regarding racial/ethnic background, the sample was predominantly Black (68.4%, n = 13) and White (15.8%, n = 3), one participant was Asian (5.3%, n = 1) and two were Other (10.5%, n = 2).

3.2. MEASURES

3.2.1. The Behavioral Regulations in Exercise Questionnaire (BREQ)

BREQ (Mullan et al., 1997; Mullan & Markland, 1997; Wilson et al., 2002) is a 15-item survey measuring the self-determination in exercises. Each item of BREQ was rated on a 6-point Likert scale: 1 = Strongly disagree, 2 = Disagree, 3 = Somewhat disagree, 4 = Somewhat agree, 5 = Agree, 6 = Strongly agree. The BREQ contains four subscales: a four-item subscale of external regulation (e.g., *I exercise because other people say I should.*), a three-item subscale of introjected regulation subscale (e.g., *I feel guilty when I don't exercise.*), a four-item subscale of identified regulation (e.g., *I value the benefits of exercise.*), and a four-item subscale of intrinsic regulation (e.g., *I exercise because it's fun.*). The BREQ has been widely used in exercise motivation research, and studies have shown the instrument to be a reliable and valid measure of exercise motivation in various populations (Cronbach's α is .81 to .89; Cid et al., 2012).

3.3. STUDY DESIGN

For the simulation study, we considered three key design factors: (1) LLM chatbots, (2) temperature setting (Low = 0; High = 0.5), and (3) prompt components. First, we evaluated three widely adopted commercial generative AI models (LLM chatbots): OpenAI's GPT-4.1 (*gpt*), Google's Gemini 2.0 Flash (*gemini*), and Anthropic's Claude 3.7 Sonnet (*claude*). These models were selected given their prominence and accessibility in current applications of large language modeling. Second, we manipulated the temperature parameter to control the degree of randomness in token selection during the response generation. We selected 0.5 as the high-temperature condition based on preliminary analyses indicating that temperatures exceeding 0.5 (e.g., 0.7) produced highly variable and conversational outputs that lacked the structure necessary for generating coherent item-level survey responses (Jiang et al., 2023). We discuss prompt design and LLM response generation procedure in the next sections.

In total, there are 3 (LLM chatbots) × 4 (prompts) × 2 (temperature settings) = 24 conditions. The calling functions of *Application Programming Interface* (API) of LLMs were created in Python and the data analysis was performed in R. The R code and the supplementary materials are available on the OSF platform (DOI:10.17605/OSF.IO/AFQG3).

3.3.1. LLM Response Generation Procedure

The procedure of interview-informed LLM survey responses consists of three steps (see Figure 1): (1) data collection; (2) LLM simulation; and (3) comparison and evaluation. In Step 1, all participants were invited to complete a semi-structured interview and structured questionnaires. In Step 2, the collected information and research background information were used to create different prompts. These prompts differed in the components they included, which instructed LLM on how to generate survey responses. Finally, in Step 3, we compared the generated responses and actual survey responses using various evaluation metrics. Description of each prompt and evaluation metric are detailed below.

In each experimental condition, one response was generated for each participant and each survey item, yielding 24 (conditions) × 19 (participants) × 15 (items) = 6840 total observations. For the temperature condition specifically, our analysis focused on single realizations from the sampling distribution rather than characterizing the full distributional variability. We generated LLM responses at the individual level due to token output limitations. Specifically, given item specifications and research background (for prompt 3 and prompt 4, a participant's demographic information was also provided), the LLM generated responses for all 15 survey items simultaneously within a single condition, producing a semicolon-separated string (e.g., '1;1;1;1;3;3;3;5;5;5;4;5;5;5;5'). We then iterated across different experimental conditions for each participant. For each iteration, a new LLM chat session was initiated without retaining previous conversation history.

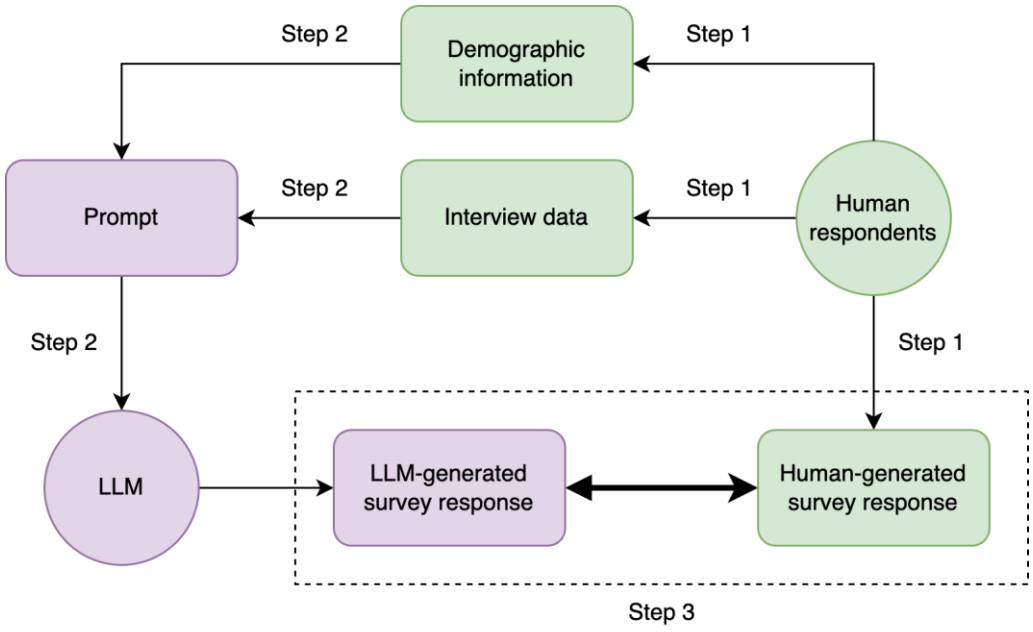

Figure 1: Procedure of the simulation study (green: human data collection; purple: response generation process of LLM)

3.3.2. Prompt Design and Implementation

Clarifying the content of prompts is crucial for advancing our understanding of how LLMs process and integrate different types of information (e.g., demographic data, interview transcripts, research context) to generate responses, which is essential for developing more effective and

interpretable LLM-based research methodologies. As shown in Figure 2, we developed four prompts based on varying combinations of four prompt components (see the example in Supplementary Materials): (1) Research Background, including program description and research aims (e.g., "*The program is a physical activity initiative designed to.... This research aims to...*"); (2) Item Information, containing the survey item content and response scale (e.g., "*Item 1: ...; Response Scale: 1 = Strongly disagree; ...; 6 = Strongly agree*"); (3) Personal Interview, containing the transcript of the interview of participants with masked identification; and (4) Demographic Information, containing age, race/ethnicity, and gender of participants (e.g., "*You are a 31-year-old white female*."). Prompts with personal interview and/or demographic components are considered as persona-based prompts.

These prompt components were combined in four different ways to evaluate whether the inclusion of specific information components influences the performance of LLM-based response simulation. Specifically, Prompt 1 (*P-BR*; Research Background and Item Information) served as the baseline condition, containing only the research background and survey information. Because Prompt 1 does not include any participant-level information, we ran the prompt 19 times—once for each human participant. This allowed us to generate a set of LLM responses that represent the 19 participants, which can be directly compared to the 19 human participants' survey responses, ensuring consistency in the analysis across all conditions. Prompt 2 (*P-BR-PI*; P-BR + Personal Interview) and Prompt 3 (*P-BR-DI*; P-BR + Demographic Information) extended P-BR by incorporating personal interview data or demographic information, respectively. Prompt 4 (*P-BR-PI-DI;* P-BR + Personal Interview + Demographic Information) was the most comprehensive, combining all four components (research background, survey item content, interview responses and demographic information).

To further evaluate the prompts used in current study, we quantified the length of the prompt by the number of tokens used for each prompt following the *o200k_base* encoding used by GPT-4o. The results showed that Prompt 1 has the lowest number of tokens (all prompts have same number of tokens, at 1,276), Prompt 2 and Prompt 3 have middle level of number of tokens (Mean=5,917; Min = 3,190; Max=9,641), and Prompt 4 has the highest number of tokens across all samples (Mean=5,952; Min=3,227; Max=9,676). The length of the prompt, however, does not indicate the richness of content. According to both length and content of the prompts, Prompt 4 has the highest amount of information followed by Prompt 2 and Prompt 3, and Prompt 1 contains the least amount of information. The amount of information for Prompt 2 and Prompt 4 depends on their person-level content (i.e., interview data). For each prompt, the LLMs were instructed to simulate the role of the interviewee and generate the response to each item of the BREQ mentioned above.

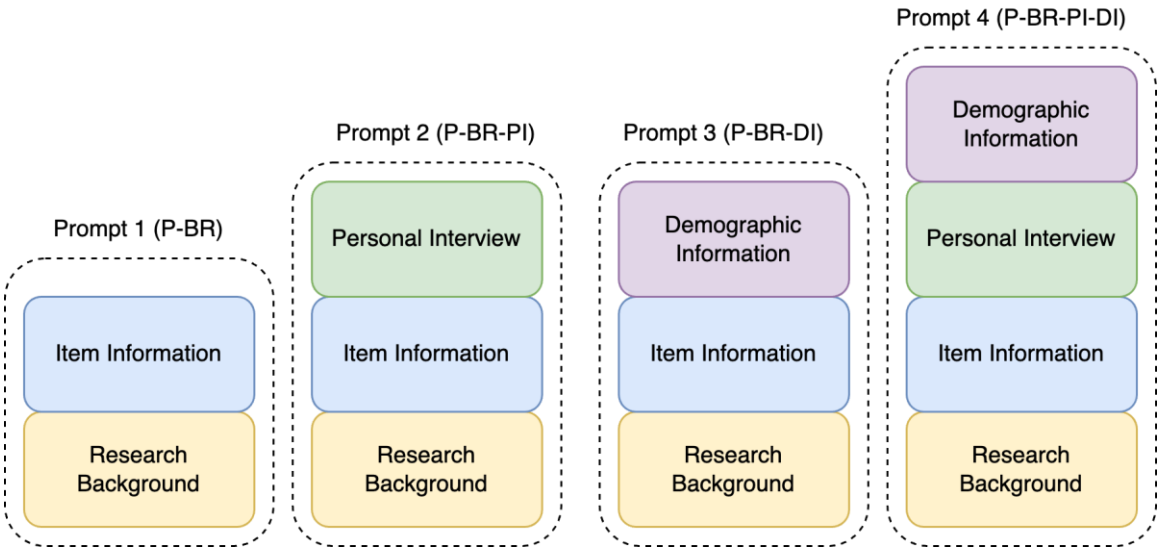

Figure 2: Components of four prompts

3.4. Evaluation Criteria

In this study, we investigated the alignment among various respondents regarding their tendency (overall responses across respondents) and diversity (overall variability across respondents). First, analyzing tendency helps identify whether LLMs can capture the central tendencies and response patterns that characterize human survey behavior, which is crucial for validating LLMs as reliable proxies for human responses in research contexts. Second, examining diversity allows researchers to assess whether LLMs can replicate the natural variability and individual differences present in human populations, which is essential for maintaining the ecological validity of simulated responses.

3.4.1. Alignment among LLMs

The alignment among three LLMs (*gpt*, *claude*, and *gemini*) was evaluated using the following methods. First, we calculated and visualized the average item means and variances of LLM-generated responses over participants in each condition. Second, we calculated the average *Pearson* correlations (ρ) between LLM-generated responses for each condition. Third, we performed a three-way ANOVA with Pearson's correlations as the dependent variable (DV) and three factors (LLM pairs, prompt settings, and temperature settings) as independent variables (IVs) to identify important factors that influenced the alignment among LLMs.

3.4.2. Alignment between LLMs and Human

We first presented descriptive statistics of LLM-generated responses and human-generated responses including item-level means and variances in different conditions to compare LLMs to human respondents. Second, we compared LLM-generated responses to human-generated survey responses in different layers. We utilized item-level ($RMSE_i$; Equation 1), person-level ($RMSE_p$; Equation 2), and test-level root-mean-square deviation ($RMSE_T$; Equation 3) to quantify the deviation between LLM-generated and human-generated responses in different aspects. Item- and person-level RMSEs can help identify which items or persons exhibit the largest deviation between LLM-generated responses and human survey responses. These may indicate

some person/item characteristics that LLMs struggle to capture or reveal potential measurement errors. The three types of RMSEs are defined as follows.

Item-level RMSE is calculated as the square root of the average squared difference between LLM-generated and human-generated responses for each item across all participants.

$$RMSE_i = \sqrt{\frac{\sum_1^P \left(X_{i,p,AI} - X_{i,p,Human}\right)^2}{P}} \qquad (1)$$

where $X_{i,p,AI}$ is the LLM-generated response for item i from participant p, and $X_{i,p,Human}$ is the human-generated response for item i from participant p. P denotes the total number of samples.

Person-level RMSE is used to quantify the difference between LLM-generated and human-generated responses for each participant across all items.

$$RMSE_p = \sqrt{\frac{\sum_1^I \left(X_{i,p,AI} - X_{i,p,Human}\right)^2}{I}} \qquad (2)$$

Test-level RMSE is used to evaluate the deviations between LLM-generated total test score and human-generated total test score across all LLM samples and human participants.

$$RMSE_T = \sqrt{\frac{\sum_1^P \left(RAI_{p,AI} - RAI_{p,Human}\right)^2}{P}} \qquad (3)$$

$$RAI_{p,\cdot} = -2 * S_{p,ext} - S_{p,int} + 2 * S_{p,ide} + S_{p,int} \qquad (4)$$

where $RAI_{p,\cdot}$ of the BREQ denotes the *relative autonomy index* of person p, which can be calculated as the weighted sum of four subscale scores using the formula in Equation 4. $S_{p,ext}$, $S_{p,int}$, $S_{p,ide}$, and $S_{p,int}$ denote the mean scores of external regulation, introjected regulation, identified regulation, and intrinsic regulation, respectively.

4. Results

4.1. Descriptive Statistics

As shown in Figure 3, all three LLMs produced approximately similar tendencies of item means compared to human participants across items, temperature settings, and prompts. Specifically, items 1 to 7 had relatively lower item means than items 8 to 15. The similarity of average item means across all samples between Prompt 1 and other prompts may suggest that LLMs may capture the items' linguistical features and potential human' responses to items regardless of the amount of personal information in the prompts. Compared to other LLMs, *gemini* appears to systematically underpredict when given additional context, indicated by lower mean scores for Prompt 4. Across all conditions, LLMs tended to generate more extreme responses than humans; that is, for items with lower human ratings (e.g., items 1–7), LLM mean scores were even lower, and for items with higher human ratings (e.g., items 8–15), LLM mean scores were even higher. In addition, temperature settings did not show much difference in item means across all three LLMs. Prompts only had slight differences in the means for certain LLMs, such as *gemini*, but not for others.

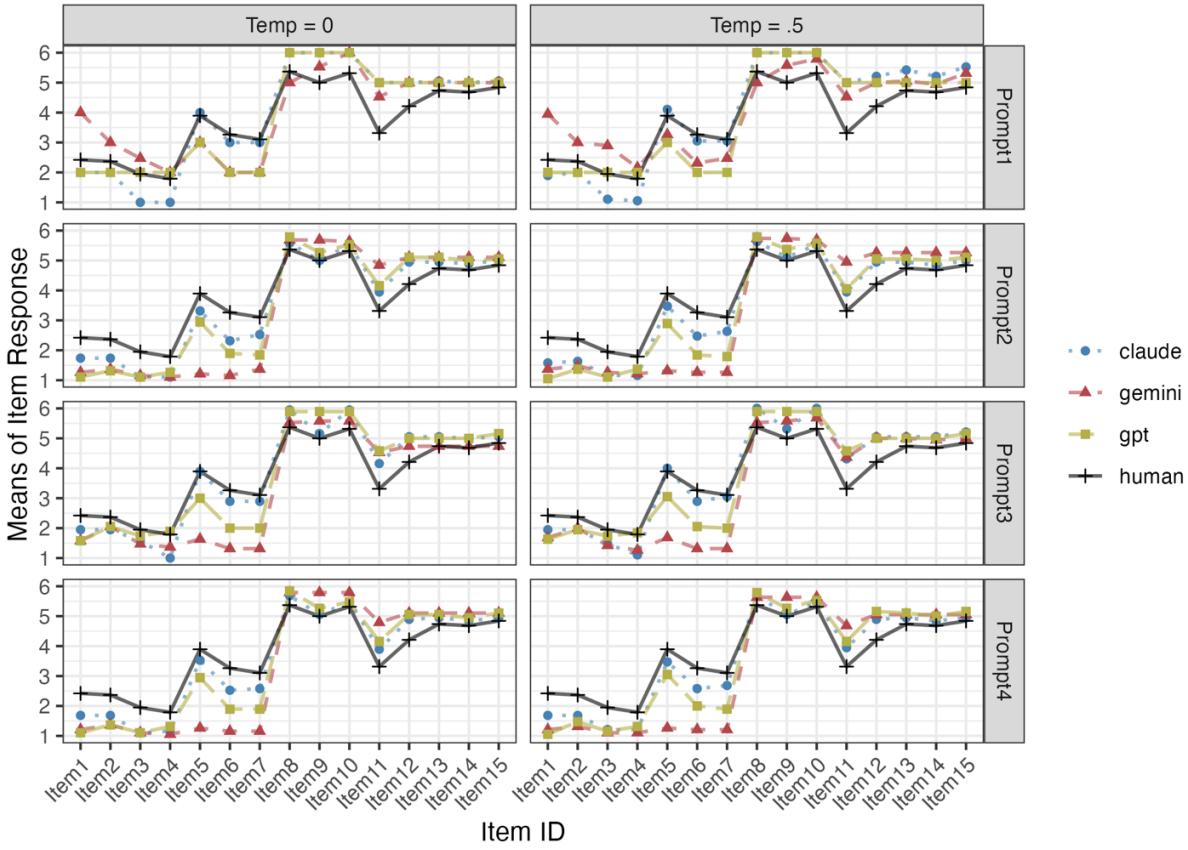

Figure 3: Item-level mean responses for LLMs and human across 15 items

Figure 4 presents the variances of item responses across all respondents or LLM-generated samples in different conditions. Unlike item means, variances of LLM-generated responses were much lower than human responses under all conditions, indicating that LLM-generated responses were less diverse than human responses. For Prompt 1, because the prompt content was identical across all participants, LLM variability is low to zero. That is, the more information one prompt contains, the higher variances LLM-generated responses yielded. Combined with evidence from Figure 3, this indicated that LLMs produced more consistently extreme values. In addition, the variances of LLM-generated responses interacted with LLM chatbots— *claude* had more diverse responses when prompts contained the interview data (Prompt 2 and Prompt 4). In contrast, *gemini* had more diverse responses when prompts contained demographic information only (Prompt 3). Also, higher temperature conditions (Temp = .5) showed higher variances of item responses than lower temperature conditions (Temp = 0) for all LLM chatbots, which was consistent with the definition of temperature.

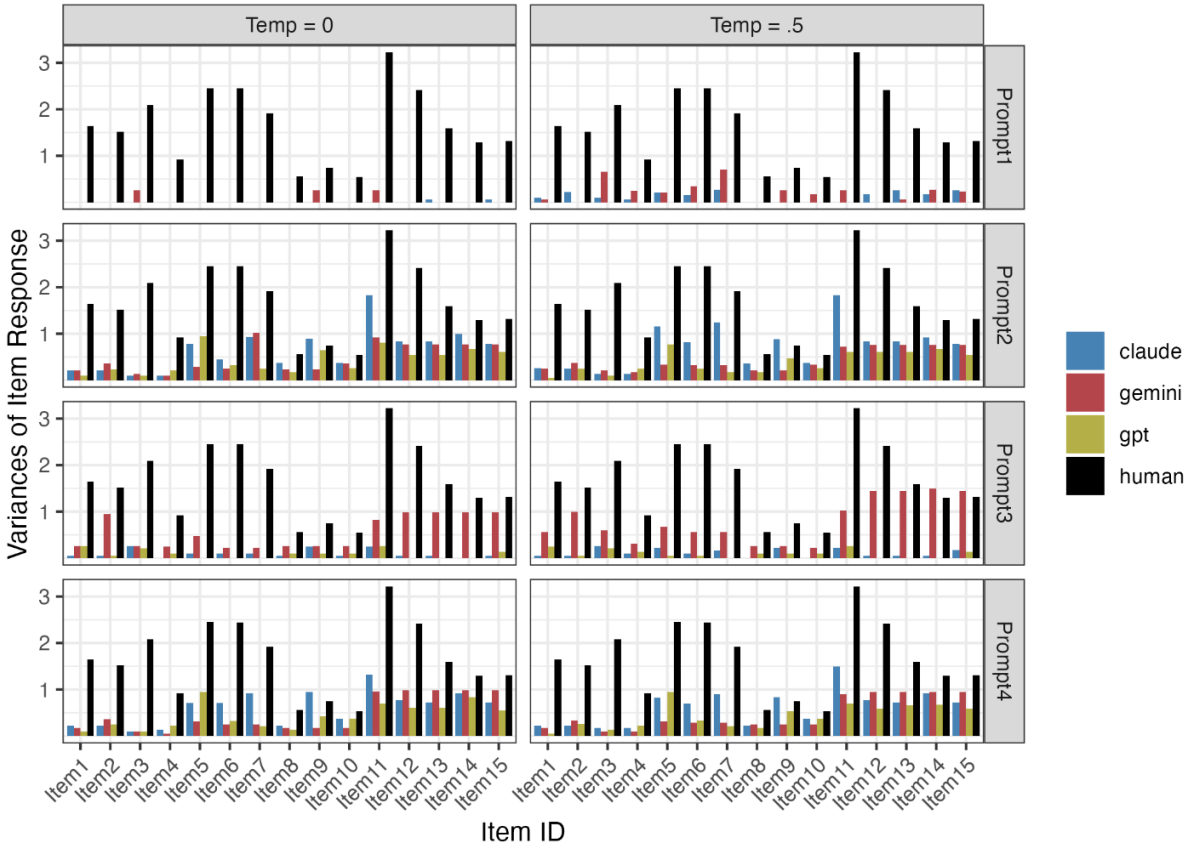

Figure 4: Item-level variances for LLMs and human across 15 items

4.2. ALIGNMENT AMONG LLMS

Table 1 presents the Pearson correlations among the three LLMs in different conditions. Results show that *gpt* and *claude* have the highest correlations across various temperatures and prompts ($\rho \in [0.92, 0.95]$), followed by *gpt* and *gemini* ($\rho \in [.87, .93]$), and *claude* and *gemini* ($\rho \in [.81, .88]$). The results also show that the correlations between the three LLM chatbots in the conditions of low temperature (Temp = 0; $\bar{\rho}$ = .91) are slightly higher than those in the conditions of higher temperature (Temp = .5; $\bar{\rho}$ = .89). In addition, the results show that the correlations in the conditions of prompts containing personal interview data (Prompt 2 and Prompt 3; $\bar{\rho}$ = .92) are higher than prompts without personal interview data (Prompt 1 and Prompt 3; $\bar{\rho}$ = .88). The results of three-way ANOVAs show that prompt settings ($F_{3,17} = 16.657, p < .001$) and temperature settings ($F_{1,17} = 8.133, p = .011$) have significant main effects on the correlations among the three LLM chatbots.

Table 1: Pearson correlations among three LLM chatbots across all conditions

Temp	Prompt	$\rho_{gpt,claude}$	$\rho_{gpt,gemini}$	$\rho_{gemini,claude}$
Low	Prompt 1	0.943	0.913	0.839
Low	Prompt 2	0.953	0.925	0.882
Low	Prompt 3	0.928	0.911	0.827
Low	Prompt 4	0.951	0.930	0.875

Temp	Prompt	$\rho_{gpt,claude}$	$\rho_{gpt,gemini}$	$\rho_{gemini,claude}$
High	Prompt 1	0.925	0.874	0.806
High	Prompt 2	0.940	0.932	0.862
High	Prompt 3	0.923	0.879	0.807
High	Prompt 4	0.944	0.935	0.875

4.3. Alignment between LLMs and Humans

As shown in Figure 5, the average Pearson correlations between LLMs chatbot responses with human responses show that there are medium to high relationships ($\rho \in [.5, .73]$). Among the three LLM chatbots, *claude* shows the highest correlations with humans for all four prompts, followed by *gpt*. In contrast, *gemini* has the lowest correlations with humans across conditions. The results also show that prompts containing interview data (Prompt 2 and Prompt 4) have relatively higher associations with humans than other prompts (Prompt 1 and Prompt 3). Prompt 2 and Prompt 4 have comparable levels of correlations across all LLM chatbots. Prompt 3 (prompt containing demographic information) has slightly higher correlations than the baseline (Prompt 1) for *claude*, *gemini* and *gpt*.

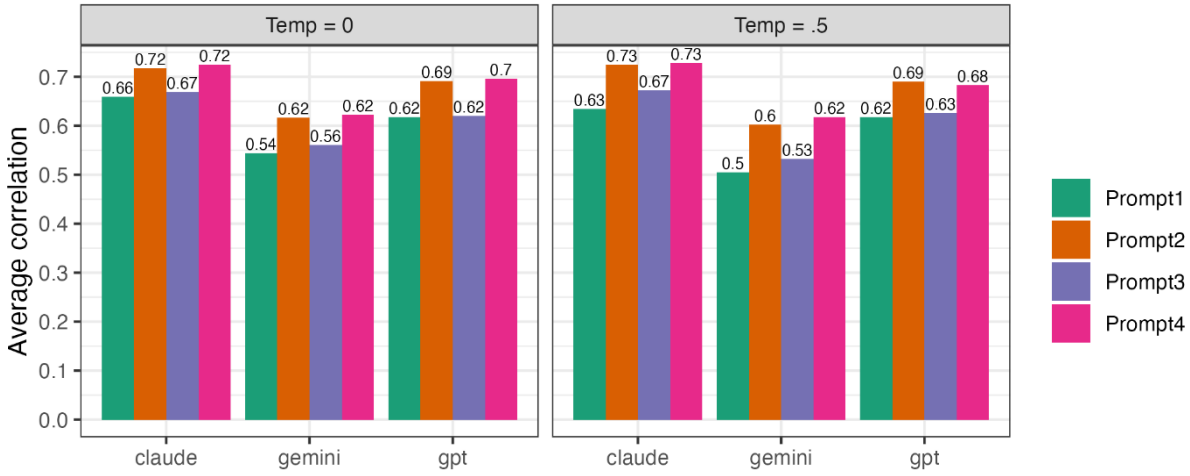

Figure 5: Pearson correlations between LLMs and human responses for four prompts

4.3.1. Item-level RMSEs

Figure 6 shows the item-level RMSEs for 15 items in different conditions. Overall, *claude* shows lower RMSEs than *gemini* and *gpt*, suggesting that *claude* has the highest alignment with human respondents. In comparison, *gemini* has the highest item-level RMSEs among the three LLM chatbots. As for specific items, we found that items 6, 7, and 11 displayed relatively higher RMSEs than other items. After screening those items carefully, we found that item 6 (*I feel ashamed when I miss an exercise session*), item 7 (*I feel like a failure when I haven't exercised in a while*), and item 11 (*I get restless if I don't exercise regularly*) contain relatively negative emotional words, such as "*ashamed*", "*failure*", and "*restless*" while other items have relatively positive words, such as item 8 (*I value the benefits of exercise*) or item 10 (*I think it is important to make the effort to exercise regularly*).

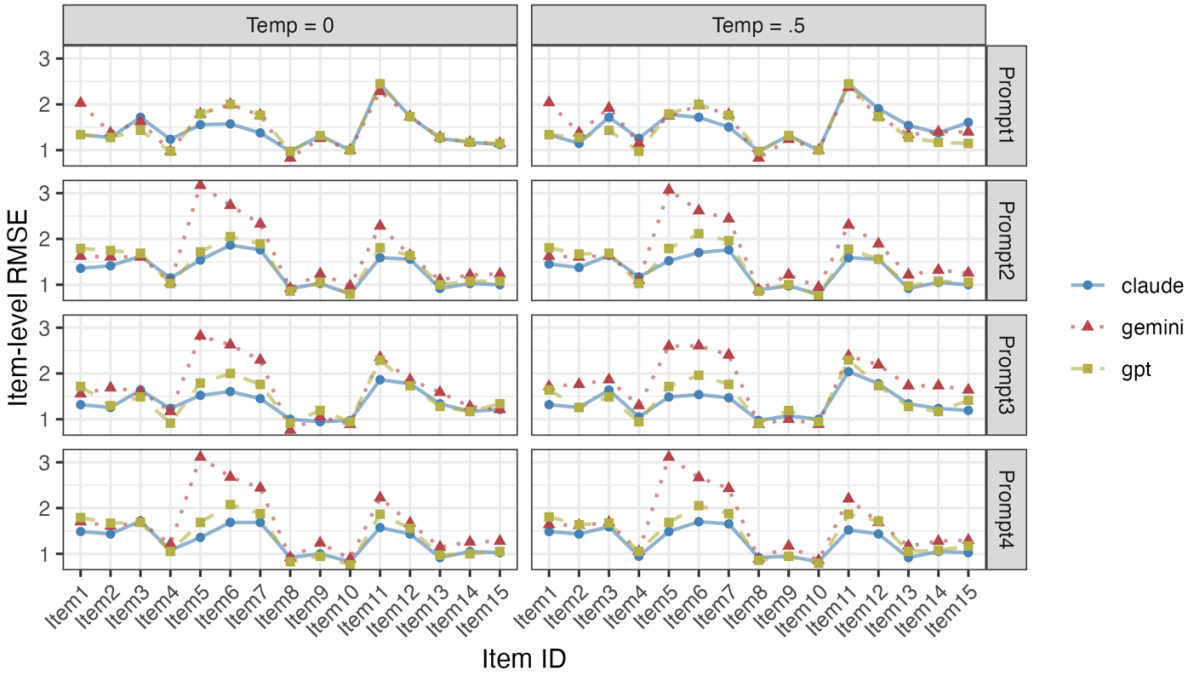

Figure 6: Item-level RMSEs between LLMs and human across samples

4.3.2. Person-level RMSEs

Figure 7 shows the person-level RMSEs between 19 human respondents and LLM-generated samples. When using Prompt 2 and Prompt 4, two respondents (ID 2111 and 2303) have the highest alignment between their survey responses with the LLMs' generated responses, indicated by the lowest person-level RMSEs. *claude* has the lowest person-level RMSEs when using Prompt 2 and Prompt 4, suggesting that it has the highest alignment with human respondents. In addition, the correlation between the number of tokens in individual interviews and the average person-level RMSE for Prompt 2 and Prompt 4 was moderate while not statistically significant ($\rho = .404$, $p = .086$), suggesting that longer interviews (greater number of tokens in prompts) did not necessarily lead to better alignment between LLM-generated and human responses. Nonetheless, such an outcome may stem from our small sample size and limited power.

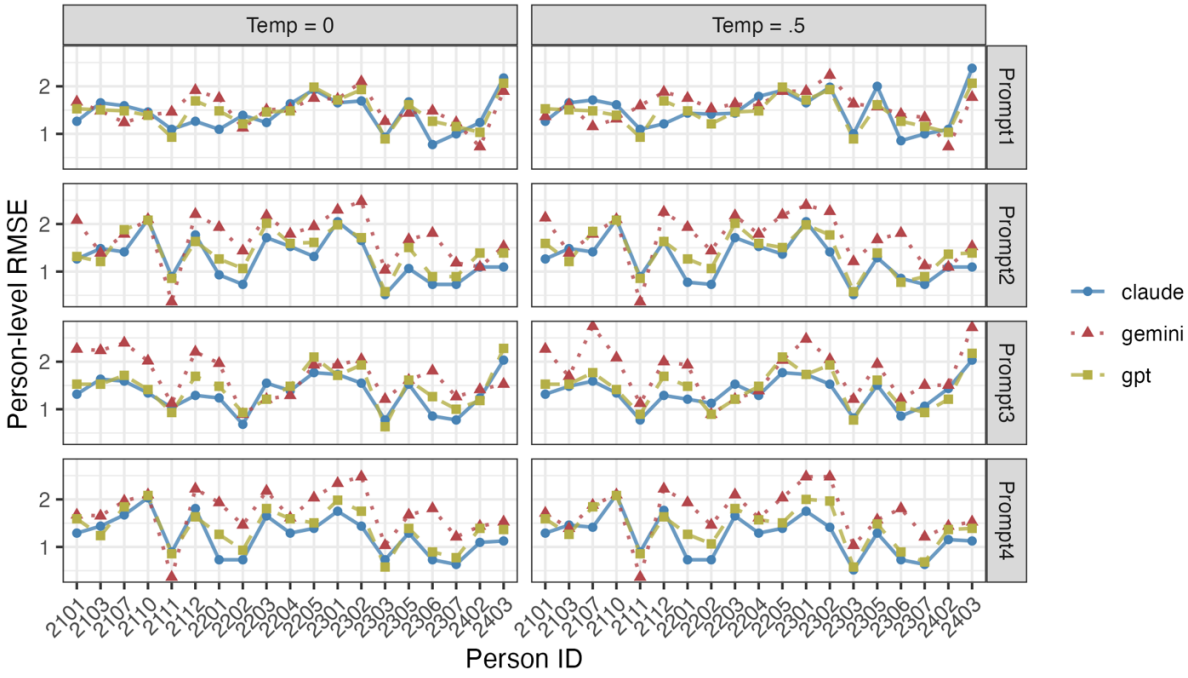

Figure 7: Person-level RMSEs between LLMs and human across 15 items

4.3.3. Test-level RMSEs

Finally, Figure 8 shows the test-level RMSEs of the *relative autonomy index* (RAI; see Equation 4) for all conditions. The results suggested that compared to Prompt 1, only *claude* with Prompt 2, Prompt 3, and Prompt 4 showed higher alignment with human respondents at the test level, which indicated, for some chatbots (*gemini* and *gpt*), more personal information in the prompts (e.g., interview data and demographic information) may not necessarily improve the alignment between LLMs and humans towards the RAI. The calculation of RAI requires theoretical understanding of the definitions and relationships among study variables (specifically for this study, external regulation, introjected regulation, identified regulation, and intrinsic regulation), which is not included in the training of LLM chatbots.

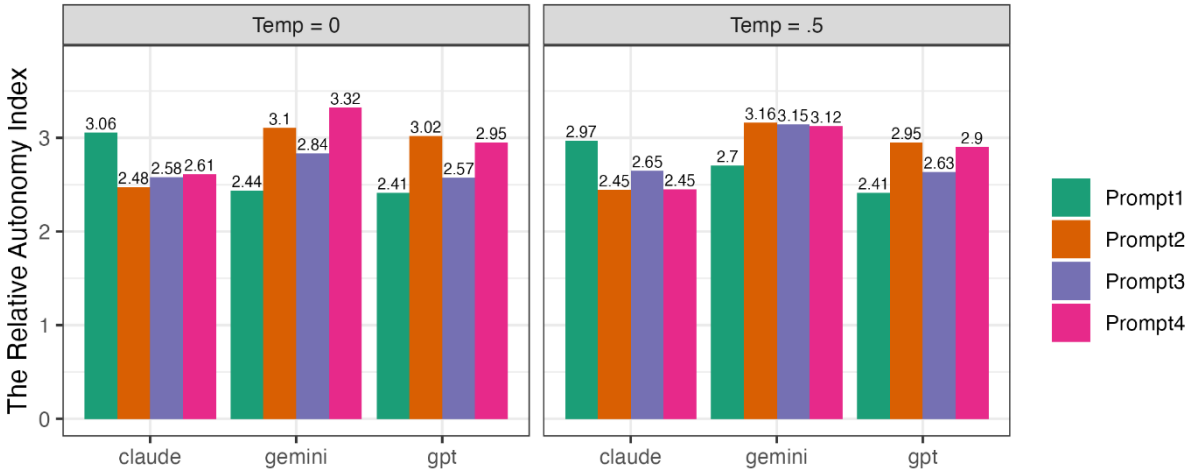

Figure 8: Test-level RMSEs between LLMs and Human

5. Discussion

This study introduces an LLM-driven approach that leverages personal interviews of after-school program staff to simulate quantitative survey responses. The comparisons between LLM-generated and staff actual survey data provide indications of tendency and variability of LLMs' survey responses, which can reveal measurement and personal characteristics that may help researchers improve their research design. Researchers could benefit from the LLM generation procedure and prompt design in this study when utilizing LLMs as a research tool in educational data mining (e.g., data validation or augmentation). However, we emphasize that LLM-generated data are not intended to replace human data, but rather to augment mixed-methods research by offering new ways to connect qualitative and quantitative findings.

Our first finding is that post-trained LLM chatbots can capture the overall pattern of BREQ responses from after-school program staff across different temperature and prompt settings; however, LLM-generated responses tend to exhibit lower variability than those of human respondents (see Figure 3 and Figure 4). In general, incorporating personal interview data can improve the diversity of generated responses for some LLM chatbots (i.e., *claude* and *gpt*). This finding is consistent with other LLM-driven simulation studies that well-crafted personas can yield results that approximate real-world human behaviors on average but hardly approximate the variability of human behaviors (e.g., A. Li et al., 2025).

Our second finding is that prompt content (e.g., persona-based prompts) can affect the alignment among LLMs chatbots and between LLM-generated responses and human survey responses (see Figure 5). Findings indicated that persona-based prompts containing personal interview data improved alignment in both response patterns and diversity between LLM-generated and human responses (Prompts 2 & 4 vs. Prompts 1 & 3), while additional demographic information in prompts showed minimal improvement in alignment (Prompt 4 vs. Prompt 2). Individuals' demographic information was less impactful on alignment perhaps because the relationship between individuals' demographic information and their survey responses is relatively weaker than the interview. Further research is needed to investigate the importance of different types of qualitative data in training LLMs for research in social sciences.

Our final finding is that the discrepancy between LLMs and human responses suggests further investigation into item and person characteristics. Specifically, the item-level RMSEs between LLM and human tended to be higher for survey items with negative emotional words (e.g., *ashamed*, *failure*, and *restless*; see Figure 6), suggesting that LLMs may struggle to align with human responses on negative emotional wording in the survey in our example. This discrepancy may arise because our tasks involve mimicking human survey response patterns, which differs fundamentally from the objective tasks examined in previous studies. Another possibility is that the underlying decision-making mechanisms of LLMs operate through different cognitive processes than human respondents, even when producing superficially similar responses. This distinction is crucial for understanding the limitations of using LLMs as proxies for human participants in survey research contexts. Additionally, Figures 3 and 6 showed that items with high RMSE (e.g., items 6, 7, and 11) exhibit extreme mean responses from LLMs but medium-level responses from humans. This pattern suggests that either survey items or interview questions may not accurately capture true human attitudes, leading LLMs to provide extreme responses on certain items. These items/questions might need to be revisited by content experts, and human response characteristics may be further investigated through person-level RMSE analyses.

The person-level RMSEs suggest that LLMs differ in their ability to learn from diverse respondents. We found a moderate but not significant association between the length of individual interviews (measured by the number of tokens) and the alignment. This may suggest that the relevance of interview content, rather than its length, plays a more critical role in enhancing the alignment of LLM-generated responses. However, given our small sample size (n=19), the impact of interview length on alignment remains inconclusive and warrants further investigation with larger samples. The test-level RMSEs suggest that LLMs struggle to grasp the overall psychometric structure of the test when item scores are aggregated and weighted by subscales. In other words, while they can mimic individual item patterns, LLMs have difficulty reproducing the underlying psychometric relationships of the whole test. This finding is consistent with prior studies that LLMs can capture individual differences in personality traits and emotional expressions (Y. Li et al., 2024; Serapio-García et al., 2023) but may not adhere to psychometric principles rigorously (Huang, Jiao, et al., 2024; Huang, Wang, et al., 2024; A. Li et al., 2025; Liu et al., 2025; P. Wang et al., 2024).

5.1. Limitations and Future Directions

5.1.1. Designing Effective Prompts Using Qualitative Data

Although prior studies have demonstrated that when guided by appropriately designed prompts, LLMs can serve as a viable method for validating the consistency between qualitative (e.g., interview) and quantitative (e.g., survey) data, the scientific principles underlying prompt construction remain underdeveloped. A central challenge lies in identifying which attributes of respondents should be included in prompts to optimize LLM-generated responses for specific research purposes. In the current study, demographic and interview-based information were incorporated. However, future research could benefit from considering additional respondent characteristics such as psychographic (e.g., personality traits, values), behavioral (e.g., past actions), or contextual variables (e.g., social or interview setting), as suggested by A. Li et al. (2025). In addition, the four prompts differ in both their length (number of tokens) and content (included components), making it difficult to distinguish the roles of these two factors. Additionally, LLM responses may be sensitive to prompt wording and structure while our prompts

are limited to four types with similar structure. Systematic investigations are needed to determine how to tailor these elements for domain-specific applications.

5.1.2. Limited Diversity in LLM-Generated Responses

Despite the small sample size, our findings are consistent with prior research that LLM-generated responses often exhibit lower diversity than actual human respondents. This is a general pattern observed across various domains, including political science (Bisbee et al., 2024) and educational technology (Liu et al., 2025). This consistency strengthens the generalizability and highlights a fundamental characteristic of LLM-generated responses that researchers should be aware of when using LLMs as proxies for human participants in survey research contexts. Nevertheless, the underlying causes of this limitation remain poorly understood, and the conclusions of this study should be interpreted with caution given the small sample size. Further investigation is required to identify whether this constraint stems from small sample size, model architecture, training data, or the input prompt features.

5.1.3. Bias in Prediction Accuracy Across Individuals

The accuracy of LLM-generated responses in approximating real individual responses appears to vary across participants and items, raising questions about LLMs' bias from measurement and personal characteristics. Previous studies have shown that LLMs are likely to present various types of biases depending on the demographic information of individuals (Binz & Schulz, 2023; Dillion et al., 2023; Yan et al., 2024). One possibility is that emotionally valenced tokens (e.g., sad) may act as surface-level cues that influence outputs. Additionally, emotional inconsistencies across data collection modes (e.g., differing emotional states during interviews vs. surveys) may contribute to mismatches between LLMs' predictions and actual responses. If emotional tone or opinion varies across these modalities, LLMs' predictions may reflect that divergence. This suggests that interview protocols could be improved by encouraging emotional neutrality or consistent perspectives, potentially enhancing the alignment between qualitative data and LLM-generated survey responses.

5.1.4. Leveraging Interview Data to Reduce Bias

It should be noted that both human survey responses and interview-informed LLM responses are imperfect proxies for the latent constructs of interest. That is, the observed discrepancies may be due to multiple possible sources (e.g., LLM limitations, human response behaviors such as satisficing or impression management, and potential instrument issues). More research is needed to explore how researchers can leverage interview data to mitigate bias in LLM-generated outputs. Qualitative data analysis techniques, such as word cloud visualizations or topic modeling, may help identify language features—such as topic richness or information density—that influence the fidelity and variability of generated responses. Understanding these dynamics may offer new pathways for refining prompt design and improving model performance.

5.1.5. Model Temperature and Response Stability

Other factors that significantly affect LLM-generated survey responses include the temperature settings of the LLM. Our findings suggest that increasing the temperature may enhance the diversity of generated responses. However, temperatures exceeding a certain threshold (e.g., >0.8) tend to produce outputs that lack coherence and structural integrity. Currently, there is a lack of empirical guidance on optimal temperature settings for survey simulation tasks, particularly in

education and psychology. Future research should focus on identifying appropriate model settings that balance diversity with consistency, particularly in high-stake applications such as educational assessment. In addition, our primary objective was to compare deterministic (temperature = 0) and stochastic (temperature = 0.5) response generation procedures given the interview content. To make both temperature settings comparable, our simulation focused on single realizations from the sampling distribution while we acknowledge that the stochasticity inherent to higher temperature values can yield different outputs given the same input. Systematically investigating the dispersion and representativeness of LLM-generated responses under stochastic settings will be one of our future research directions.

Declaration of Generative AI Software Tools in the Writing Process

During the preparation of this work, the authors used ChatGPT in the Abstract section in order to improve clarity and refine language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgement

Funding Support: Research reported in this publication was supported by the National Institute of Nursing Research of the National Institutes of Health under Award Number 1R01NR017619-01 (Zarrett, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Acknowledgements: We would like to thank our school and after school program community partners.
Disclosure Statement: The authors report there are no competing interests to declare.

References

Agarwal, M., Goswami, A., Sharma, P., Agarwal, M., Goswami, A., & Sharma, P. 2023. Evaluating ChatGPT-3.5 and claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. *Cureus 15*, 9, e46222. https://doi.org/10.7759/cureus.46222

Anthropic 2025. *Claude 2* [Large language model]. https://www.anthropic.com/news/claude-2

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis 31*, 3, 337–351. https://doi.org/10.1017/pan.2023.2

Binz, M. & Schulz, E. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences 120*, 6, e2218523120. https://doi.org/10.1073/pnas.2218523120

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. 2024. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis 32*, 4, 401-416. https://doi.org/10.1017/pan.2024.5

BISHOP, F. L. 2015. Using mixed methods research designs in health psychology: An illustrated discussion from a pragmatist perspective. *British Journal of Health Psychology 20*, 1, 5–20. https://doi.org/10.1111/bjhp.12122

CHANG, S., CHASZCZEWICZ, A., WANG, E., JOSIFOVSKA, M., PIERSON, E., & LESKOVEC, J. 2024. LLMs generate structurally realistic social networks but overestimate political homophily. *arXiv*. https://doi.org/10.48550/arXiv.2408.16629

CHEN, J., WANG, X., XU, R., YUAN, S., ZHANG, Y., SHI, W., XIE, J., LI, S., YANG, R., ZHU, T., CHEN, A., LI, N., CHEN, L., HU, C., WU, S., REN, S., FU, Z., & XIAO, Y. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv*:2404.18231. https://doi.org/10.48550/arXiv.2404.18231

CID, L., MOUTÃO, J., LEITÃO, J., & ALVES, J. 2012. Behavioral Regulation Assessment in Exercise: Exploring an Autonomous and Controlled Motivation Index. *The Spanish Journal of Psychology 15*, 3, 1520–1528. https://doi.org/10.5209/rev_SJOP.2012.v15.n3.39436

CRESWELL, J. W. & CLARK, V. L. P. 2017. *Designing and conducting mixed methods research*. SAGE Publications.

DILLION, D., TANDON, N., GU, Y., & GRAY, K. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences 27*, 7, 597–600. https://doi.org/10.1016/j.tics.2023.04.008

DING, M., DENG, C., CHOO, J., WU, Z., AGRAWAL, A., SCHWARZSCHILD, A., ZHOU, T., GOLDSTEIN, T., LANGFORD, J., ANANDKUMAR, A., & HUANG, F. 2024. Easy2Hard-bench: Standardized difficulty labels for profiling LLM performance and generalization. *Advances in Neural Information Processing Systems 37*, 44323–44365.

EKIN, S. 2023. Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices. *TechRxiv*:22683919.v1.

FATEEN, M. & MINE, T. 2025. Developing a tutoring dialog dataset to optimize LLMs for educational use. *arXiv:2410.19231*. https://doi.org/10.48550/arXiv.2410.19231

FEDERIAKIN, D., MOLEROV, D., ZLATKIN-TROITSCHANSKAIA, O., & MAUR, A. 2024. Prompt engineering as a new 21st century skill. *Frontiers in Education 9*. https://doi.org/10.3389/feduc.2024.1366434

GE, T., CHAN, X., WANG, X., YU, D., MI, H., & YU, D. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv:2406.20094*. https://doi.org/10.48550/arXiv.2406.20094

HUANG, J., JIAO, W., LAM, M. H., LI, E. J., WANG, W., & LYU, M. R. 2024. Revisiting the reliability of psychological scales on large language models. *arXiv:2305.19926*. https://doi.org/10.48550/arXiv.2305.19926

HUANG, J., WANG, W., LI, E. J., LAM, M. H., REN, S., YUAN, Y., JIAO, W., TU, Z., & LYU, M. R. 2024. Who is ChatGPT? Benchmarking LLMs' psychological portrayal using PsychoBench. *arXiv:2310.01386*. https://doi.org/10.48550/arXiv.2310.01386

JANSEN, B. J., SALMINEN, J., JUNG, S., & GUAN, K. 2022. *Data-driven personas*. Springer Nature.

JIANG, H., ZHANG, X., CAO, X., BREAZEAL, C., ROY, D., & KABBARA, J. 2023. PersonaLLM: Investigating the ability of large language models to express personality traits. *arXiv:2305.02547*. https://doi.org/10.48550/arXiv.2305.02547

JOHNSON, R. B., ONWUEGBUZIE, A. J., & TURNER, L. A. 2007. Toward a definition of mixed methods research. *Journal of Mixed Methods Research 1*, 2, 112–133. https://doi.org/10.1177/1558689806298224

LAVERGHETTA JR., A. & LICATO, J. 2023. Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications*, E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch, Eds. Association for Computational Linguistics, 414–428. https://doi.org/10.18653/v1/2023.bea-1.34

LI, A., CHEN, H., NAMKOONG, H., & PENG, T. 2025. LLM generated persona is a promise with a catch. *arXiv:2503.16527*. https://doi.org/10.48550/arXiv.2503.16527

LI, Y., HUANG, Y., WANG, H., ZHANG, X., ZOU, J., & SUN, L. 2024. Quantifying AI psychology: A psychometrics benchmark for large language models. *arXiv:2406.17675*. https://doi.org/10.48550/arXiv.2406.17675

LIU, Y., BHANDARI, S., & PARDOS, Z. A. 2025. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology 56*, 3, 1028–1052. https://doi.org/10.1111/bjet.13570

LIU, Y., SHARMA, P., OSWAL, M. J., XIA, H., & HUANG, Y. 2024. PersonaFlow: Boosting research ideation with LLM-simulated expert personas. *arXiv:2409.12538*. https://doi.org/10.48550/arXiv.2409.12538

LOZIĆ, E. & ŠTULAR, B. 2023. Fluent but not factual: A comparative analysis of ChatGPT and other AI chatbots' proficiency and originality in scientific writing for humanities. *Future Internet 15*, 1010, 336. https://doi.org/10.3390/fi15100336

MANCORIDIS, M., WEEKS, B., VAFA, K., & MULLAINATHAN, S. 2025. Potemkin Understanding in Large Language Models. *arXiv:2506.21521v2.* https://doi.org/10.48550/arXiv.2506.21521

MAY, T. A., STONE, G. E., FAN, Y., SONDERGELD, C. J., LAPLANTE, J. N., PROVINZANO, K., KOSKEY, K. L. K., & JOHNSON, C. C. 2025. Using generative artificial intelligence tools to develop multiple-choice assessment items: An effectiveness study. *Education Sciences 15,* 2, educscu15020144.

MENDONÇA, P. C., QUINTAL, F., & MENDONÇA, F. 2025. Evaluating LLMs for automated scoring in formative assessments. *Applied Sciences 15*, 55, 2787. https://doi.org/10.3390/app15052787

MULLAN, E., MARKLAND, D., & INGLEDEW, D. 1997. A graded conceptualisation of self-determination in the regulation of exercise behaviour: development of a measure using confirmatory factor analytic procedures. *Pers Individ Differ 23*, 745–752. https://doi.org/10.1016/S0191-8869(97)00107-4

MULLAN, E. & MARKLAND, D. 1997, Variations in self-determination across the stages of change for exercise in adult. *Motivation and Emotion 21*, 349–362. https://doi.org/10.1023/A:1024436423492

NORI, H., KING, N., MCKINNEY, S. M., CARIGNAN, D., & HORVITZ, E. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv:2303.13375*. https://doi.org/10.48550/arXiv.2303.13375

OPENAI, ACHIAM, J., ADLER, S., AGARWAL, S., AHMAD, L., AKKAYA, I., ALEMAN, F. L., ALMEIDA, D., ALTENSCHMIDT, J., ALTMAN, S., ANADKAT, S., AVILA, R., BABUSCHKIN, I.,

BALAJI, S., BALCOM, V., BALTESCU, P., BAO, H., BAVARIAN, M., BELGUM, J., …& ZOPH, B. 2023. GPT-4 technical report. *arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774

PARKER, M. J., ANDERSON, C., STONE, C., & OH, Y. 2024. A Large Language Model Approach to Educational Survey Feedback Analysis. *International Journal of Artificial Intelligence in Education 35*, 444–481. https://doi.org/10.1007/s40593-024-00414-0

PENG, Q., LIU, H., XU, H., YANG, Q., SHAO, M., & WANG, W. 2024. Review-LLM: Harnessing large language models for personalized review generation. *arXiv:2407.07487*. https://doi.org/10.48550/arXiv.2407.07487

PONCE, O. A. & PAGÁN-MALDONADO, N. 2015. Mixed methods research in education: Capturing the complexity of the profession. *International Journal of Educational Excellence 1*, 1, 111–135. https://doi.org/10.18562/ijee.2015.0005

POWELL, H., MIHALAS, S., ONWUEGBUZIE, A. J., SULDO, S., & DALEY, C. E. 2008. Mixed methods research in school psychology: A mixed methods investigation of trends in the literature. *Psychology in the Schools 45*, 4, 291–309. https://doi.org/10.1002/pits.20296

RASHEED, Z., WASEEM, M., AHMAD, A., KEMELL, K.-K., XIAOFENG, W., DUC, A. N., & ABRAHAMSSON, P. 2024. Can large language models serve as data analysts? A multi-agent assisted approach for qualitative data analysis. *arXiv:2402.01386*. https://doi.org/10.48550/arXiv.2402.01386

SAAB, K., TU, T., WENG, W.-H., TANNO, R., STUTZ, D., WULCZYN, E., ZHANG, F., STROTHER, T., PARK, C., VEDADI, E., CHAVES, J. Z., HU, S.-Y., SCHAEKERMANN, M., KAMATH, A., CHENG, Y., BARRETT, D. G. T., CHEUNG, C., MUSTAFA, B., PALEPU, A., … NATARAJAN, V. 2024. Capabilities of gemini models in medicine. *arXiv:2404.18416*. https://doi.org/10.48550/arXiv.2404.18416

SLAVIN, R., & SMITH, D. 2009. The Relationship Between Sample Sizes and Effect Sizes in Systematic Reviews in Education. *Educational Evaluation and Policy Analysis 31*, 4, 500–506. https://doi.org/10.3102/0162373709352369

SARSTEDT, M., ADLER, S. J., RAU, L., & SCHMITT, B. 2024. Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing 41*, 6, 1254–1270. https://doi.org/10.1002/mar.21982

SCHOONENBOOM, J. 2023. The fundamental difference between qualitative and quantitative data in mixed methods research. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research 24*, 11. https://doi.org/10.17169/fqs-24.1.3986

SERAPIO-GARCÍA, G., SAFDARI, M., CREPY, C., SUN, L., FITZ, S., ROMERO, P., ABDULHAI, M., FAUST, A., & MATARIĆ, M. 2023. Personality traits in large language models. *arXiv:2307.00184*. https://doi.org/10.48550/arXiv.2307.00184

SHOJAEE, P., MIRZADEH, I., ALIZADEH, K., HORTON, M., BENGIO, S., & FARAJTABAR, M. 2025. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *arXiv:2506.06941*. https://doi.org/10.48550/arXiv.2506.06941

SUN, S., LEE, E., NAN, D., ZHAO, X., LEE, W., JANSEN, B. J., & KIM, J. H. 2024. Random silicon sampling: Simulating human sub-population opinion using a large language model based on

group-level demographic information. *arXiv:2402.18144*. https://doi.org/10.48550/arXiv.2402.18144

UTO, M. & UCHIDA, Y. 2020. Automated short-answer grading using deep neural networks and item response theory. In *Proceedings of Artificial intelligence in education*, I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán, Eds. Springer International Publishing, 334–339. https://doi.org/10.1007/978-3-030-52240-7_61

WANG, J., HIDA, R. M., PARK, J., KIM, E. K., & BEGENY, J. C. 2024. A systematic review of mixed methods studies published in six school psychology journals: Prevalence, characteristics, and trends from 2011 to 2020. *Psychology in the Schools 61*, 4, 1302–1317. https://doi.org/10.1002/pits.23114

WANG, P., ZOU, H., YAN, Z., GUO, F., SUN, T., XIAO, Z., & ZHANG, B. 2024. Not yet: Large language models cannot replace human respondents for psychometric research. *OSF:rwy9b*. https://doi.org/10.31219/osf.io/rwy9b

WANG, Q. & LI, H. 2025. On continually tracing origins of LLM-generated text and its application in detecting cheating in student coursework. *Big Data and Cognitive Computing 9*, 33, 50. https://doi.org/10.3390/bdcc9030050

WILSON, P., RODGERS, W. & FRASER, S. 2002. Examining the Psychometric Properties of the Behavioral Regulation in Exercise Questionnaire. *Measurement & Evaluation in Exercise & Sport Science 6*, 1-21. https://doi.org/10.1207/S15327841MPEE0601_1

WU, S., KOO, M., BLUM, L., BLACK, A., KAO, L., SCALZO, F., & KURTZ, I. 2023. A comparative study of open-source large language models, GPT-4 and claude 2: Multiple-choice test taking in nephrology. *arXiv*. https://doi.org/10.48550/arXiv.2308.04709

XU, S. & ZHANG, X. 2023. Leveraging generative artificial intelligence to simulate student learning behavior. *arXiv:2310.19206*. https://doi.org/10.48550/arXiv.2310.19206

YAN, L., SHA, L., ZHAO, L., LI, Y., MARTINEZ-MALDONADO, R., CHEN, G., LI, X., JIN, Y., & GAŠEVIĆ, D. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology 55*, 1, 90–112. https://doi.org/10.1111/bjet.13370

YU, C., YE, J., LI, Y., LI, Z., FERRARA, E., HU, X., & ZHAO, Y. 2024. A large-scale simulation on large language models for decision-making in political science. *arXiv:2412.15291*. https://doi.org/10.48550/arXiv.2412.15291