Leveraging Interview-Informed LLMs to Model Survey Responses

Comparative Insights from AI-Generated and Human Data

Jihong Zhang[1], Xinya Liang[1], Anqi Deng[1], Nicole Bonge[1],
Lin Tan[1], Ling Zhang[2], & Nicole Zarret[3]

[1]University of Arkansas · [2]University of Wyoming · [3]University of South Carolina

American Educational Research Association, 2026

Section 1

Introduction

Background: The Mixed Methods Challenge

Mixed methods research integrates quantitative & qualitative data

- Qualitative data (interview, focus group): open-ended, textual, context-rich
- Quantitative data (survey, test scores): numeric, standardized scales
- Structural mismatch hinders comparison of measurement characteristics & individual response patterns

LLMs: bridge quantitative and qualitative insights by playing the role of a "simulated respondent" informed by qualitative data

Key Gap

No prior work examined consistency between human Likert-scale responses and LLM-generated responses informed by individual-level interview data.

The Core Idea

1. Understand how well LLMs can simulate human survey responses based on interview data.
2. Introduce LLM as a novel tool for diagnosing consistency or coherence between qualitative and quantitative data.

LLM Personas & Synthetic Survey Responses

Persona-driven methods

- **Role-playing**: LLMs adopt specific identities (e.g., "a 35-year-old female ASP staff with 10 years of experience")
- **Personas**: generalized user representations built from real data

Why: LLMs as research tools

- Cost-effective for pilot testing & data augmentation
- Produce synthetic datasets with emergent personality shaped by prompts & training

Key Configuration Factors

Chatbot GPT, Gemini, Claude APIs.

Temperature Controls randomness. $T = 0$: deterministic; $T = 0.5$: stochastic. Too high degrades coherence.

Prompt design Minor wording changes can be pivotal. Prompt effects are understudied.

Gaps & Contribution

Prior Work

- LLMs capture broad patterns but struggle with item-level nuances *(Wang et al., 2024)*
- May not adhere to **psychometric principles** *(Huang et al., 2024; Li et al., 2025)*
- Proponents: scalable, cost-effective proxies *(Liu et al., 2025)*
- Critics: validity concerns, training bias, cognitive mismatch *(Mancoridis et al., 2025)*

Three Gaps Our Study Addresses

1. Small-N settings underexplored *(Slavin & Smith, 2009)*
2. Limited work on valid **Likert-scale** generation *(Liu et al., 2024)*
3. No work on LLM–human consistency with **individual-level interview data**

Our Approach

Research context + **real interview transcripts** + demographics → predict individual Likert-scale responses → compare to actual human data

Research Questions

Three Research Questions

RQ1 How do LLM-generated survey responses vary in **means and variability** across different chatbots, temperature settings, and prompt configurations?

RQ2 How do chatbot choice, prompt design, and temperature **influence alignment** between LLM-generated and human responses?

RQ3 What do discrepancies between LLM and human responses reveal about **measurement-level and person-level characteristics**?

Alignment: how closely LLM simulated responses match human responses informed by the same inter

Section 2

Method

Data & Instrument

Setting: Interview and Survey Responses from Connect through PLAY program (RCT)

- After-school program (ASP) staff, 10 ASPs
- Southeastern U.S., 2023–2024
- $N = 19$ who completed *both* interview + survey regarding ASP staff's physical activity (PA) motivation

Sample characteristics:

- Mean age = 35.5 years (range: 18–61)
- 84.2% female
- 68.4% Black, 15.8% White

Interviews: Semi-structured, 15–25 min each; assessed PA experiences, motivation, perceived support, attitudes toward youth health

BREQ Instrument

Behavioral Regulation in Exercise Questionnaire (BREQ)

- 15 items, 4 subscales
- 6-point Likert scale (1 = Strongly disagree, 6 = Strongly agree)

Subscale	Items
External regulation	4
Introjected regulation	3
Identified regulation	4
Intrinsic regulation	4

Cronbach's α = .81–.89

Study Design

3 Design Factors:

1. **LLM Chatbot** (3 levels)
   - GPT-4.1
   - Gemini 2.0 Flash
   - Claude 3.7 Sonnet
2. **Temperature** (2 levels)
   - Low $= 0$ (deterministic)
   - High $= 0.5$ (stochastic)
3. **Prompt** (4 levels; see right)

$$\underbrace{3}_{\text{LLMs}} \times \underbrace{4}_{\text{prompts}} \times \underbrace{2}_{\text{temps}} = \mathbf{24} \text{ conditions}$$

Total observations:
$24 \times 19(persons) \times 15(items) = \mathbf{6{,}840}$

Four Prompt Configurations

Prompt	Components RB	II	PI	DI
P1 (baseline)	✓	✓		
P2 (+Interview)	✓	✓	✓	
P3 (+Demographics)	✓	✓		✓
P4 (Full)	✓	✓	✓	✓

RB = Research Background; II = Item Information;
PI = Personal Interview; DI = Demographic Info

Prompt Length

P1: 1,276 tokens — P2/P3: avg 5,917 tokens
P4: avg 5,952 tokens (most information-rich)

Evaluation Metrics

Alignment among LLMs

- Average item means and variances
- Pearson correlations of survey responses across LLMs (ρ)
- ANOVA (ρ as DV; LLMs, prompts, and temperature as IVs)

Alignment between LLMs and Human

Item-level RMSE$_i$

$$\sqrt{\frac{\sum_{p=1}^{P}(X_{i,p,AI} - X_{i,p,H})^2}{P}}$$

Which items are hardest for LLMs to predict?

Person-level RMSE$_p$

$$\sqrt{\frac{\sum_{i=1}^{I}(X_{i,p,AI} - X_{i,p,H})^2}{I}}$$

Which respondents are hardest to simulate?

Test-level RMSE$_T$

$$\sqrt{\frac{\sum_{p=1}^{P}(RAI_{p,AI} - RAI_{p,H})^2}{P}}$$

How well does LLM replicate the RAI score?

Section 3

Results

Alignment among LLMs

Among-LLM Pearson Correlations:

LLM Pair	Low T	High T	Avg
GPT–Claude	.943	.925	.94
GPT–Gemini	.913	.874	.91
Claude–Gemini	.839	.806	.85

3-way ANOVA on inter-LLM ρ:

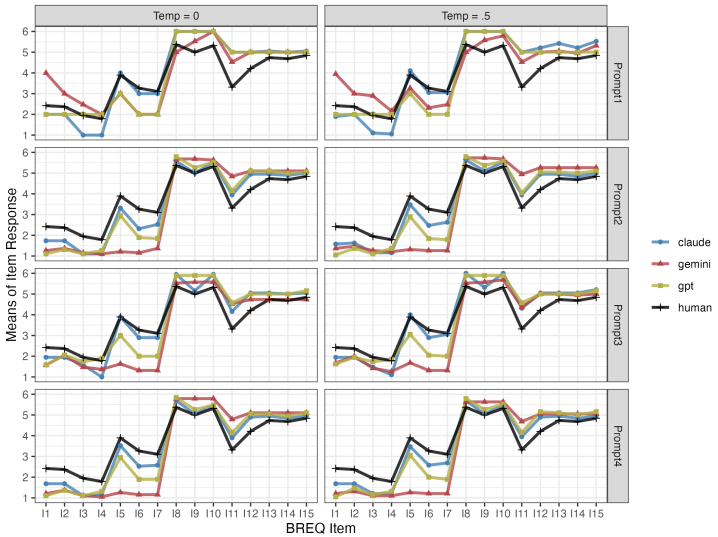
- **Prompt:** $F(3, 17) = 16.66,\ p < .001$
- **Temperature:** $F(1, 17) = 8.13,\ p = .011$

LLM–Human Correlations: $\rho \in [.50,\ .73]$

Findings

1. LLMs show high consistency with each other, especially at low temperature. They are quite reliable.
2. LLMs are sensitive to prompt design, with interview-informed prompts improving alignment.
3. LLMs are also sensitive to temperature, with higher temperature reducing inter-LLM consistency.

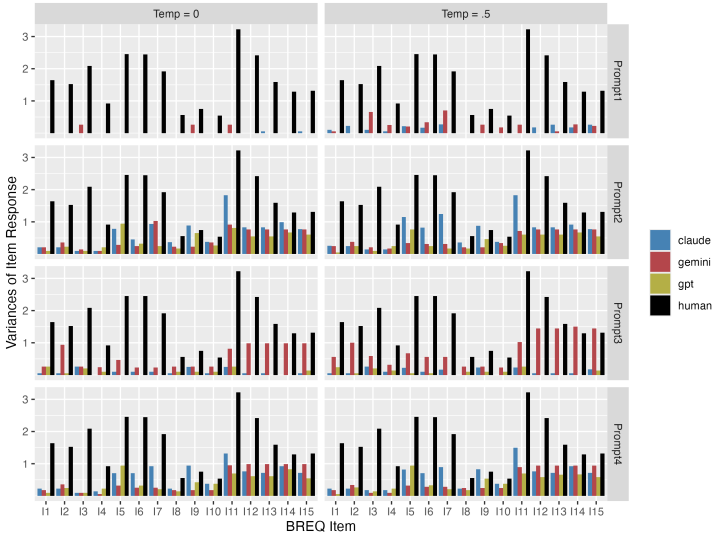
Item Means



LLMs are able to track human item mean patterns:

- Items 1–7: lower means; Items 8–15: higher means
- LLMs amplify extremes — undershooting low items, overshooting high items
- Temperature had little impact; Gemini underpredicts with richer prompts (P4)

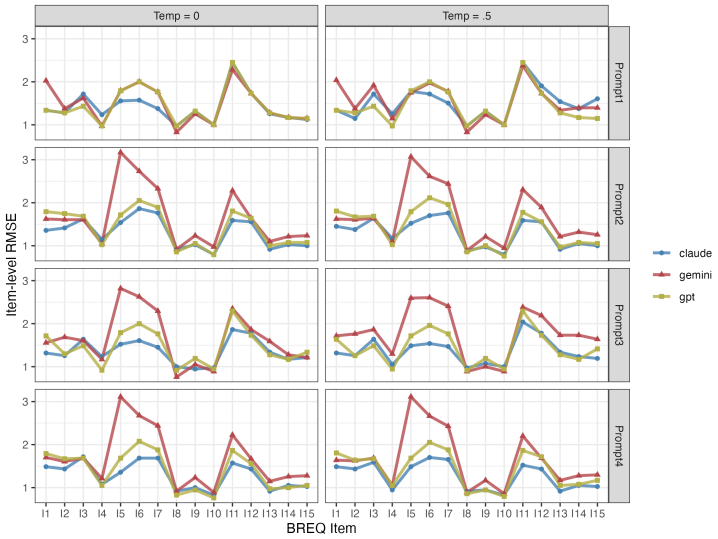
Item Variances



LLMs are NOT able to mimic human variability:

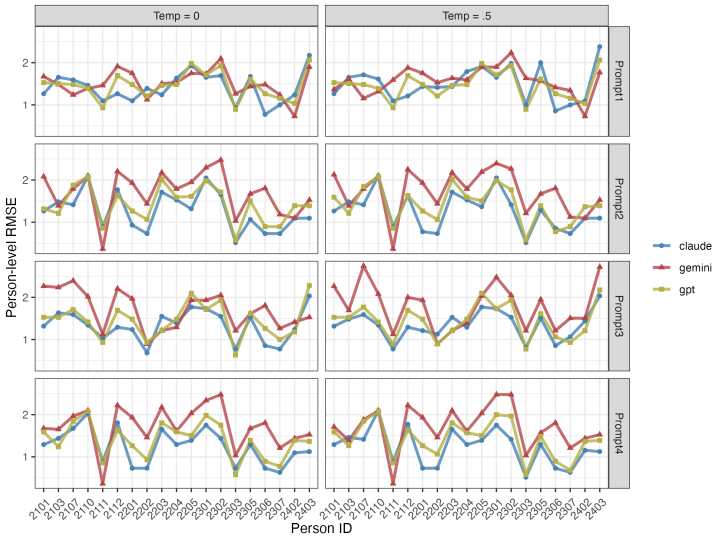
- LLM variance $\ll$ Human variance.
- Prompt Deisgns: P1: near-zero variance; P2/P4 (+interview): increases for *claude* & *gpt*; P3 (+demographics): *gemini* has closest variance to human data
- Higher temperature $\rightarrow$ higher variance

Alignment between LLMs and Human: Item-level RMSEs



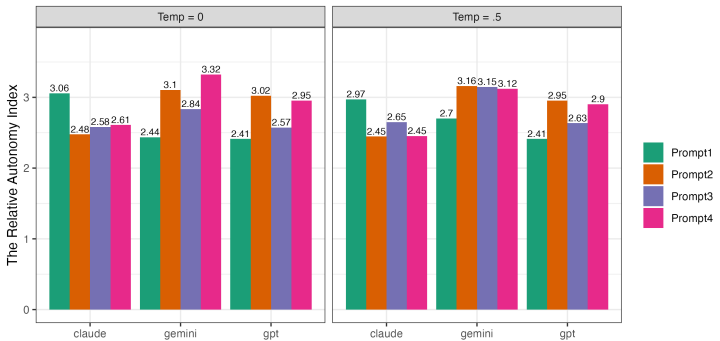
- *Claude* has lowest RMSE; *Gemini* highest
- Items 6, 7, & 11 show highest RMSE across all models
- LLMs struggle with *negative emotional wording*

Alignment between LLMs and Human: Person-level RMSEs



- Large variability across participants
- Best alignment for IDs 2111 & 2303 (with P2/P4)
- Interview *length* (tokens): $\rho = .40,\ p = .086$
- Relevance $>$ length of interview content

Alignment between LLMs and Human: Test-level RMSEs



- Only *claude* with P2, P3, P4 improved over baseline
- GPT & Gemini: more prompt info did NOT improve RAI alignment
- LLMs lack understanding of **subscale weighting relationships**

Section 4

Discussion

Summary & Implications

1. **LLMs have internal consistency**
   - High inter-LLM correlations, especially at low temperature
   - Prompt design significantly impacts LLM consistency
2. **LLMs capture patterns but not variability**
   - Interview prompts expand diversity (esp. Claude, GPT)
3. **Interview prompts improve alignment**
   - P2/P4 > P1/P3; demographics add minimal value
   - Low temperature = more stable outputs
4. **Discrepancies diagnose measurement issues**
   - Negative wording & psychometric structure (RAI) hardest to replicate

Practical Contributions

- Triangulate qual + quant data via LLMs
- Flag items for revision (high RMSE = measurement signal)
- Augment small-N studies without replacing human data

Important Caveat

LLM data **augments**, not replaces, human responses. Discrepancies reflect both LLM limitations and human inconsistencies.

Thank You

Zhang, J., Liang, X., Anqi, D., Bonge, N., Tan, L., Zhang, L., & Zarret, N. (2026). *Leveraging Interview-Informed LLMs to Model Survey Responses: Comparative Insights from AI-Generated and Human Data*. Journal of Educational Data Mining, 18(1), 1–24. https://doi.org/10.5281/zenodo.18247291

Contact Information

Jihong Zhang

jzhang@uark.edu

University of Arkansas

Questions and Comments?