The Reliability Illusion in Synthetic Patients: Psychometric Misalignment of Open-weight LLMs on PHQ-9 and GAD-7

Qian Shen
University of Florida
qian.shen1@ufl.edu

Yu Han
University of Mississippi
yhan3@go.olemiss.edu

Abstract

Globally, the incidence of depression and anxiety continues to rise, and the importance of mental health assessment scales as diagnostic tools has grown accordingly. Researchers are increasingly employing generative AI to produce large volumes of items and entire scales, which in turn elevates the costs of validating their reliability and validity. In this study, we used four open-weight LLMs to complete the GAD-7 and PHQ-9, varying prompts, sampling temperature, and dynamic contextual scenarios to emulate realistic human response patterns. Using multi-group confirmatory factor analysis, differential item functioning analyses, and other psychometric methods, we evaluate the factor structure of LLM-generated responses and assess measurement invariance relative to human responses. Our findings reveal a critical paradox: although open-weight LLMs exhibit exceptionally high internal consistency, they demonstrate severe structural mismatch and fail to achieve scalar measurement invariance against human baselines. Furthermore, pervasive differential item functioning and extreme prompt fragility indicate that these models rely on superficial, stereotype-driven semantic matching rather than simulating stable latent psychological dynamics.

1 Introduction

Recent advances in large language models (LLMs) have motivated growing interest in their use for social science studies and practice (Thapa et al., 2025), such as mental health (Lawrence et al., 2024). As one of the most prevalent mental health issues, anxiety and depression are exhibiting a growing trend on a global scale (Chen et al., 2026). Consequently, researchers are actively exploring research and clinical practices concerning the effective application of LLMs in the diagnosis and treatment of these conditions (Pavlopoulos et al., 2024).

Psychological scales such as Generalized Anxiety Disorder (GAD)-7 (Spitzer et al., 2006) and Patient Health Questionnaire (PHQ)-9 (Kroenke et al., 2001) are a common class of instruments used for preliminary diagnosis or self-assessment of anxiety and depression (Costello and Comrey, 1967). Currently, LLMs are widely used by scholars in education and psychometric measurement to develop new items and questionnaires (Tan et al., 2025; Adhikari et al., 2025). Undeniably, LLMs can design a large number of new items and scales in a short period, greatly improving the efficiency and time cost of item and questionnaire development. However, the ensuing problem is that as the number of newly designed items and questionnaires increases, the temporal and economic costs of collecting human responses to validate their reliability and validity according to traditional methods have also increased significantly.

Due to their robust comprehension and learning capabilities, LLMs are being utilized by researchers to simulate human behavior and are considered a promising approach (Park et al., 2023). However, when such models are used as synthetic respondents for psychological scales, the central question is not merely whether they can generate plausible answers, but whether those answers remain psychometrically comparable to authentic human responses. In this study, we examine this question using two widely used mental health screening scales, the GAD-7 and PHQ-9, and evaluate whether responses generated by open-weight LLMs exhibit measurement invariance relative to human data. To make the comparison more realistic, we vary respondent personas, situational context, prompting conditions, and decoding temperatures. Our primary research question is therefore whether LLM-generated responses can function as psychometrically aligned synthetic substitutes for human respondents on clinical screening scales.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 88–99

2 Related Work

Given their strong natural language capabilities, LLMs are increasingly used in the screening and treatment of anxiety and depression. For example, Tao et al. (2023) proposed an LLM-based virtual interaction framework to explore ChatGPT's potential for detecting these disorders, while Liu et al. (2025a) developed EmoScan, which screens for depression and anxiety through brief text dialogues. Xu et al. (2025) fine-tuned LLMs on clinical recordings of outpatients with depression and anxiety to automate scale assessments, achieving robust diagnosis and symptom classification through an ensemble pipeline and 10-fold cross-validation. In addition, Zhao et al. (2025) found that LLM-driven conversational agents can reduce mild symptoms in young adults, and García-Méndez et al. (2026) combined machine learning and LLMs to improve detection in unstructured conversations.

Recent studies have also examined LLM performance on academic tests and psychometric scales. Using Item Response Theory, Liu et al. (2025b) compared LLMs with undergraduates on college algebra problems and found that individual LLMs show a narrower proficiency range than humans, while ensemble models better approximate student ability distributions. Similarly, Ferreira et al. (2025); Cipriani et al. (2025) evaluated LLM responses to personality and climate change perception scales with descriptive statistics and exploratory factor analysis, showing that LLMs only partially reproduce human patterns. This suggests that LLM-generated data may be useful for early group-level psychometric prototyping, but cannot replace individual-level validation. A fundamental driver of these psychometric discrepancies is highlighted by Licht et al. (2025), who demonstrate that LLMs process ordinal scales idiosyncratically through heaping, mechanically concentrating probability mass on arbitrary numeric tokens rather than reflecting true continuous distributions. This mechanical token bias provides a critical mechanistic context for why synthetic responses may mimic human averages but fail structural validation.

Battista et al. (2026) further showed that LLMs such as GPT-3.5 can simulate depressive symptoms and score highly on relevant scales, although they may also display malingering-like patterns. Moreover, while LLMs remain limited in capturing authentic cross-cultural psychological nuances in personality inventories (Li and Qi, 2025), they show social desirability bias comparable to human responses (Salecha et al., 2024). Thus, using LLMs to pilot personality scales and questionnaires may be a valuable step before human administration (de Winter et al., 2024).

From the perspective of LLM simulation, the aforementioned studies generally rely on very small sample sizes (fewer than 100) or use fixed demographic profiles, which fails to mitigate the impact of inherent LLM stochasticity on research results. Regarding analytical rigor, most of these studies employ basic statistical methods and lack formal psychometric validation to establish measurement invariance between LLM-generated and human responses. Furthermore, the limited sample sizes in these studies can potentially undermine the precision and reliability of their conclusions.

3 Method

In this study, we selected four open-weight LLMs to respond repeatedly to the GAD-7 and PHQ-9 by systematically varying combinations of personas, prompt types, dynamic situational modifiers, and temperatures. Subsequently, we applied psychometric techniques, specifically Confirmatory Factor Analysis (CFA) and ordinal logistic differential item functioning (DIF), to evaluate the measurement invariance between LLM-generated outputs and human response data. Figure 1 shows our comprehensive methodological pipeline.

3.1 Data

For this study, we selected the GAD-7 and PHQ-9, classic self-report scales for anxiety and depression, as our primary research instruments. To facilitate a robust comparison with human responses, we used the GAD-7 data from 2019 National Health Interview Survey (NHIS), and the PHQ-9 data from 2017 to 2019 of the National Health and Nutrition Examination Survey (NHANES) after removing all samples with missing values.

3.2 LLM-Based Response Generation

We constructed a layered respondent simulation pipeline to generate synthetic respondent data for PHQ-9 and GAD-7. The pipeline was implemented with four open weight LLMs: Llama-3.1-8B-Instruct, Mistral-7B-Instruct, Llama-3.3-70B-Instruct, and GPT-OSS-120B. The inclusion of models at both the $\sim$8B and $\sim$100B parameter scales reflects a deliberate range of model capacities; prior work has demonstrated that models

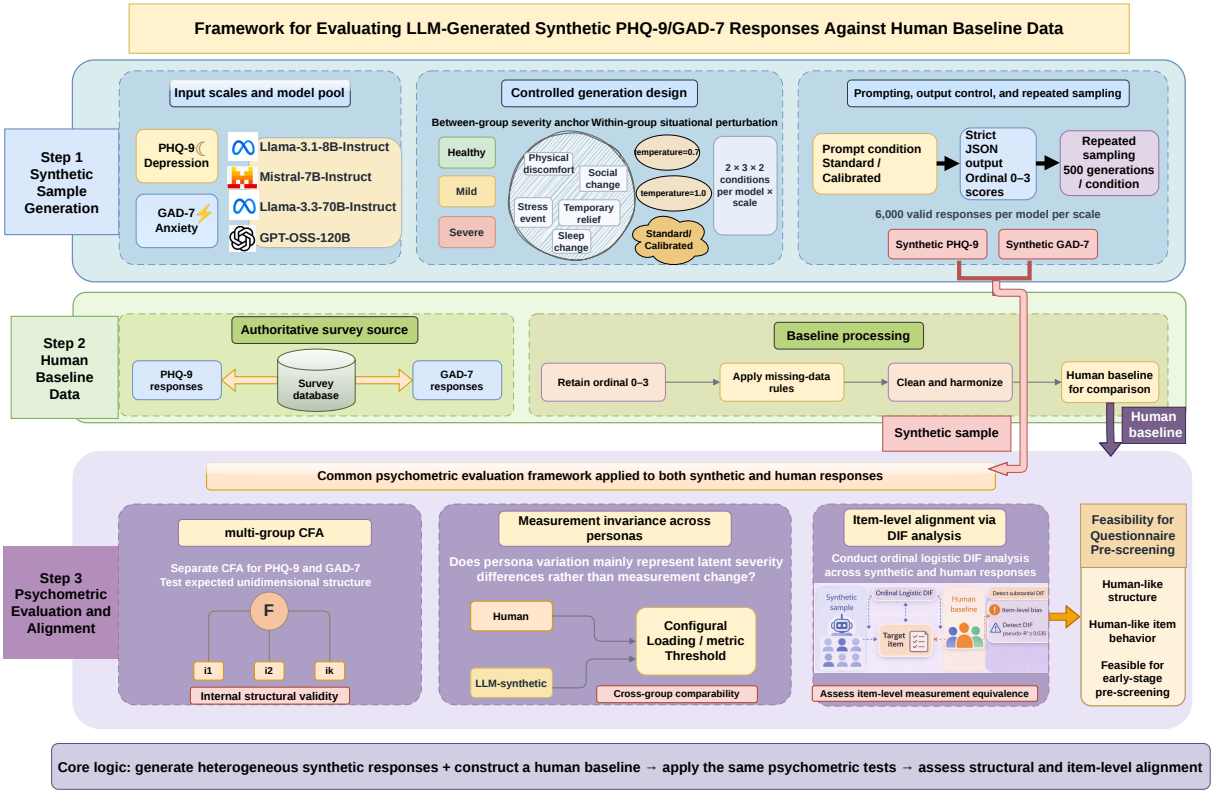

Figure 1: Overall framework of the study for generating LLM-based synthetic PHQ-9 and GAD-7 responses, and evaluating psychometric alignment between synthetic and human data.

of these scales exhibit reasonably representative language understanding and generation behavior suitable for the tasks of psychology and education (Mantena et al., 2025; Xiao and Shen, 2026; Loweimi et al., 2026; Shen et al., 2026). First, real respondents typically exhibit relatively stable individual differences when completing questionnaires (Julian, 2011). To reflect this, we assigned each model a baseline persona specifying the respondent's symptom background. For both scales, personas were defined at three severity levels, healthy, mild, and severe, so that generated responses showed a clear severity structure. This layer captured relatively stable differences in respondent symptom profiles and provided a basis for later modeling of local fluctuation and item-level variation. However, stable severity differences alone are insufficient to explain response variation within the same severity group. In real self-report data, responses reflect not only overall symptom level but also short-term fluctuations and situational influences (Armey et al., 2015; Mengelkoch et al., 2024). To model this, we introduced dynamic situational modifiers. Before each generation, the system randomly appended a brief description of the respondent's current state or recent experience, such as "you slept poorly last night." These modifiers did not alter overall severity, but influenced item-level responses, preserving the overall severity structure while retaining more natural within-group fluctuation. Even with baseline personas and dynamic situational modifiers, LLM-generated responses could still deviate from real self-report response patterns. Whereas real self-report responses reflect both relatively stable symptom tendencies and short-term situational fluctuation, LLM-based questionnaire generation may produce overly regular or internally over-consistent response patterns (Jeon et al., 2024; Chan et al., 2025). To reduce this mismatch, we introduced two prompting conditions, Standard and Calibrated, to constrain response style. Under the Standard condition, the model completed the questionnaire directly from the assigned persona and modifier. Under the Calibrated condition, the prompt discouraged uniformly extreme responding and allowed moderate item-level variation, temporary relief, and mild inconsistency. For example, severe respondents were not expected to choose the highest option for every item, while healthy respondents could still report

mild distress. This setting helped reduce overly regular response patterns. To account for sampling effects, we repeated generation under different temperature conditions to introduce additional output variability within the same respondent condition, rather than fixing responses to a single decoding setting. For each scale and model, generation conditions were defined by temperature, persona, and prompt condition. Temperature had two levels, persona had three levels (healthy, mild, severe), and prompt condition had two levels (Standard, Calibrated), yielding $2 \times 3 \times 2 = 12$ generation conditions per model for each scale. Each condition was repeated for 500 iterations, resulting in 6,000 valid responses per scale for each model, with matched sample sizes across models. All outputs followed a strict JSON schema with one integer score per item, and a mixed-score example response was provided to constrain output structure. The resulting outputs formed the synthetic respondent dataset used for subsequent psychometric evaluation.

3.3 Psychometric Methods

To ensure comparable group sizes in subsequent analyses(Chen, 2007; Herrera and Gómez, 2008), we randomly sampled 6,000 observations without replacement from each human dataset to match the number of LLM-generated responses per scale. We then fitted unidimensional CFA models separately to the human and LLM-generated data (Löwe et al., 2008; Bianchi et al., 2022), treating item responses as continuous and using robust maximum likelihood estimation (MLR). We adopted this specification because preliminary ordinal CFA models produced improper solutions (e.g., non-positive definite covariance matrices and negative residual variances), whereas MLR yielded stable and interpretable estimates for the present comparisons. Model fit was evaluated using the criteria recommended by Hu and Bentler (1999), with acceptable fit indicated by CFI and TLI ≥ 0.95, RMSEA ≤ 0.06, and SRMR ≤ 0.08. Chi-square tests for all individual CFA models were statistically significant ($p < .001$), which is expected given the large sample sizes involved and does not in itself indicate model misfit.

Measurement invariance between human and LLM-generated responses was examined using multi-group CFA under the conventional sequence of configural, metric, and scalar models, and evaluated primarily by changes in CFI between nested models. Following common practice in social science research, strict invariance was not tested, as it is widely regarded as an overly stringent criterion that is rarely achieved in practice and is generally not required for meaningful cross-group comparisons (Vandenberg and Lance, 2000; Putnick and Bornstein, 2016). To assess item-level bias, we additionally conducted ordinal logistic regression DIF analyses, using a change in McFadden's pseudo-R^2 greater than 0.035 as the criterion for substantial DIF (Zumbo, 1999; Choi et al., 2011). We also conducted within-model multi-group CFA to compare responses generated under the Standard and Context-Calibrated prompts, assessing whether internal prompt variation altered the underlying measurement structure.

4 Results and Discussion

Tables 1 and 2 show the results of our psychometric evaluation and analysis of data generated by LLMs, as well as the evaluation of the measurement invariance between the LLM-generated data and human response data. The results show a critical paradox in LLM-generated psychometric data: the illusion of high reliability masking structural collapse. Across the PHQ-9 and GAD-7, LLMs exhibited exceptionally high internal consistency (α = 0.86 to 0.96), often surpassing human baselines. However, CFA results expose this reliability as spurious. Chi-square tests were statistically significant ($p < .001$) for all models including the human baseline, which is expected given the large sample size ($N = 6{,}000$) and does not in itself indicate misfit (Hu and Bentler, 1999). Human data showed excellent unidimensional fit across all practical indices (CFI $= 0.972$, RMSEA $= 0.047$), while all LLM yielded a substantially degraded fit, with RMSEA ranging from 0.124 to 0.364 and CFI as low as 0.724, confirming that the significant chi-squares for LLM models reflect genuine structural misfit rather than a sample-size artifact. Notably, Llama-3.3-70B showed the poorest fit, and even the largest model evaluated (GPT-OSS-120B) failed to reach acceptable structural thresholds. This suggests LLMs rely on stereotype-driven semantic matching, mechanically assigning high scores to negative descriptors, rather than simulating a coherent latent psychological trait.

This structural mismatch is further corroborated by testing measurement invariance against human baselines. No LLM achieved partial scalar invariance on either scale, indicating fundamentally mis-

aligned intercepts. Furthermore, DIF analysis revealed that LLMs systematically distort specific symptom weights, exhibiting significant DIF on 7 to 8 out of the 9 PHQ-9 items. Consequently, LLM raw scores cannot be mathematically equated to human populations.

Crucially, we uncovered extreme prompt fragility within the models themselves. Tables 3 and 4 report within-model multi-group CFA results comparing the two prompt conditions. Across both scales, none of the models supported metric or scalar invariance, and configural fit was already weak in several cases, suggesting that changes in prompting affected the latent response architecture. In humans, contextual shifts may alter latent means but rarely obliterate the underlying measurement geometry. For LLMs, altering the persona prompt completely rewires the response architecture.

Collectively, these findings indicate that while LLM outputs appear clinically relevant, their underlying psychometric geometry remains fundamentally non-human. The complete failure to maintain structural consistency, both against humans and across internal prompt variations, poses severe validity risks for deploying LLMs as synthetic respondents in psychological scale pre-testing.

5 Conclusion

This study examines whether open-weight LLMs can function as psychometrically valid synthetic respondents on the GAD-7 and PHQ-9. Our results show that the answer is currently no. Despite high internal consistency, the generated responses do not preserve the same structural properties as human data, fail to support scalar invariance, and display substantial item-level bias. The central finding is a reliability illusion: LLM-generated responses achieved Cronbach's α values of 0.86 to 0.96, often exceeding the human baseline, yet this apparent reliability masked severe structural misfit. This paradox likely reflects stereotype-driven semantic matching, whereby models assign scores based on surface-level associations with symptom-related language rather than sampling from a coherent latent psychological distribution. As a result, high inter-item consistency emerges not from a shared underlying construct, but from systematic response tendencies that happen to co-vary across items, a fundamentally different mechanism from what reliability coefficients are designed to capture. The prompt fragility results further underscore this interpretation. Across all four models, even minor changes in prompt framing were sufficient to disrupt metric and scalar invariance within the same model, indicating that the latent response architecture itself shifts with surface-level wording. In human respondents, contextual factors may shift symptom expression but rarely alter the underlying measurement geometry of a well-validated scale. The fact that LLMs do not exhibit this stability suggests that they lack a stable internal representation of the clinical constructs being measured. These findings carry practical implications for the use of LLMs in psychometric scale development. While LLM-generated data may offer utility in early-stage item generation or qualitative piloting, the present results caution against their use as synthetic respondents for quantitative pre-validation, particularly when the goal is to estimate reliability or factor structure prior to human administration. Treating high Cronbach's α as sufficient evidence of data quality in this context risks systematically misleading scale development decisions. More broadly, our results speak to ongoing debates in the digital twin and LLM-as-coder literature: any claim of LLM-human substitutability must be vetted through domain-expert-led, construct-appropriate validation rather than surface-level performance metrics alone.

We therefore argue that reliability alone is an inadequate benchmark for synthetic psychometric data, and that measurement invariance relative to human baselines should serve as a more stringent criterion in future evaluations. Subsequent work might explore whether instruction-tuned models with richer demographic and clinical grounding, or models fine-tuned on large-scale self-report corpora, can better approximate the psychometric geometry of human responses. Until such evidence is available, researchers should interpret LLM-generated scale responses with considerable caution.

Limitations

Our study has several limitations. First, to enhance applicability across clinical settings with varying economic resources, we used only freely downloadable, deployable open-weight LLMs and did not employ stronger, paid models (e.g., Gemini-3, GPT-5), which may perform differently on the same tasks. Second, in our experimental design, the combination of conditions yielded approximately

balanced proportions of different mental health states; however, the true distribution in real populations completing these scales may deviate from our simulated proportions. For example, among respondents completing these scales in mental health clinics, the prevalence of actual depression or anxiety is likely higher than in large community surveys, and in neither context is a balanced prevalence assured. Third, we did not systematically ablate the role of dynamic contexts or prompting components, so the specific contribution of each design choice remains unclear. Finally, because we drew only a single sample of human responses, multi-group CFA comparisons may vary across different samples due to sampling variability.

Ethical Statement

This study used publicly available, de-identified human response data from NHIS and NHANES and did not involve new participant recruitment or direct interaction with human subjects. These data are fully open to the public; anyone may access them unconditionally, without the need to apply for special permissions or sign data usage agreements. When researchers acquire and analyze data, they must never re-link the data to an individual's identity through any encoding or key. Rather than an ethical concern intrinsic to this study, our findings serve as a cautionary signal: using LLM-generated responses as substitutes for real patient data, without rigorous psychometric validation, would constitute a failure on the part of those who disregard such evidence.

References

Divya Mani Adhikari, Alexander Hartland, Ingmar Weber, and Vikram Kamath Cannanure. 2025. Exploring llms for automated generation and adaptation of questionnaires. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–23.

Michael F. Armey, Heather T. Schatten, Natasha Haradhvala, and Ivan W. Miller. 2015. Ecological momentary assessment (ema) of depression-related phenomena. *Current Opinion in Psychology*, 4:21–25.

Fabiana Battista, Tiziana Lanciano, Raffaella Maria Ribatti, and Antonietta Curci. 2026. Malingering depression: a comparative study of human and gpt-3.5 performance. *Current Psychology*, 45(4):379.

Renzo Bianchi, Jay Verkuilen, Sharon Toker, Irvin Sam Schonfeld, Markus Gerber, Elmar Braehler, and Kurt Kroenke. 2022. Is the phq-9 a unidimensional measure of depression? a 58,272-participant study. *Psychological assessment*, 34(6):595.

Callum Chan, Sunveer Khunkhun, Diana Inkpen, and Juan Antonio Lossio-Ventura. 2025. Prompt engineering for capturing dynamic mental health self states from social media posts. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 256–267, Albuquerque, New Mexico. Association for Computational Linguistics.

Fang Fang Chen. 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling: a multidisciplinary journal*, 14(3):464–504.

Manliang Chen, Yang Wang, and Chao Tan. 2026. Epidemiological trends of depression and anxiety at global, regional, and national level: A population-based observational study from 1990 to 2021 based on global burden of disease 2021. *Medicine*, 105(2):e47094.

Seung W Choi, Laura E Gibbons, and Paul K Crane. 2011. Lordif: An r package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of statistical software*, 39:1–30.

Enrico Cipriani, Pavel Okopnyi, Danilo Menicucci, and Simone Grassini. 2025. In silico development of psychometric scales: Feasibility of representative population data simulation with llms. *arXiv preprint arXiv:2512.02910*.

CG Costello and Andrew L Comrey. 1967. Scales for measuring depression and anxiety. *The Journal of psychology*, 66(2):303–313.

Joost CF de Winter, Tom Driessen, and Dimitra Dodou. 2024. The use of chatgpt for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences*, 228:112729.

Gregorio Ferreira, Jacopo Amidei, Rubén Nieto, and Andreas Kaltenbrunner. 2025. Matching gpt-simulated populations with real ones in psychological studies—the case of the epqr-a personality test. *ACM Transactions on Computing for Healthcare*, 6(2):1–33.

Silvia García-Méndez, Francisco de Arriba-Pérez, Julen Beiro-Suso, and Francisco J González-Castaño. 2026. Real-time anxiety and depression detection by combining large language models and machine learning with explainability capabilities on a user-centric, engaging conversational assistant. *Expert Systems*, 43(4):e70229.

Aura-Nidia Herrera and Juana Gómez. 2008. Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the mantel–haenszel and logistic regression techniques. *Quality & Quantity*, 42(6):739–755.

Li-tze Hu and Peter M Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55.

Hyolim Jeon, Dongje Yoo, Daeun Lee, Sejung Son, Seungbae Kim, and Jinyoung Han. 2024. A dual-prompting for interpretable mental health language models. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 247–255, St. Julians, Malta. Association for Computational Linguistics.

Laura J. Julian. 2011. Measures of anxiety: State-trait anxiety inventory (stai), beck anxiety inventory (bai), and hospital anxiety and depression scale-anxiety (hads-a). *Arthritis Care & Research*, 63(S11):S467–S472.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

Chihao Li and Yue Qi. 2025. Toward accurate psychological simulations: Investigating llms' responses to personality and cultural variables. *Computers in human behavior*, 170:108687.

Hauke Licht, Rupak Sarkar, Patrick Y Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle. 2025. Measuring scalar constructs in social science with llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32132–32159.

June M Liu, Mengxia Gao, Sahand Sabour, Zhuang Chen, Minlie Huang, and Tatia MC Lee. 2025a. Enhanced large language models for effective screening of depression and anxiety. *Communications Medicine*, 5(1):457.

Yunting Liu, Shreya Bhandari, and Zachary A Pardos. 2025b. Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052.

Bernd Löwe, Oliver Decker, Stefanie Müller, Elmar Brähler, Dieter Schellberg, Wolfgang Herzog, and Philipp Yorck Herzberg. 2008. Validation and standardization of the generalized anxiety disorder screener (gad-7) in the general population. *Medical care*, 46(3):266–274.

Erfan Loweimi, Sofia de la Fuente Garcia, and Saturnino Luz. 2026. Predicting psychological well-being from spontaneous speech using llms. *arXiv preprint arXiv:2605.11303*.

Sriya Mantena, Anders Johnson, Marily Oppezzo, Narayan Schütz, Alexander Tolas, Ritu Doijad, C Mikael Mattson, Allan Lawrie, Mariana Ramirez-Posada, Paul Schmiedmayer, and 1 others. 2025. Fine-tuning llms in behavioral psychology for scalable health coaching. *NPJ Cardiovascular Health*, 2(1):48.

Summer Mengelkoch, Daniel P. Moriarity, Anne M. Novak, Michael P. Snyder, George M. Slavich, and Shahar Lev-Ari. 2024. Using ecological momentary assessments to study how daily fluctuations in psychological states impact stress, well-being, and health. *Journal of Clinical Medicine*, 13(1):24.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Adrianos Pavlopoulos, Theodoros Rachiotis, and Ilias Maglogiannis. 2024. An overview of tools and technologies for anxiety and depression management using ai. *Applied Sciences*, 14(19):9068.

Diane L Putnick and Marc H Bornstein. 2016. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, 41:71–90.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models display human-like social desirability biases in big five personality surveys. *PNAS nexus*, 3(12):pgae533.

Qian Shen, Fanghua Cao, Min Yao, Shlok Gilda, Bonnie J Dorr, and Walter L Leite. 2026. Children's english reading story generation via supervised fine-tuning of compact llms with controllable difficulty and safety. *arXiv preprint arXiv:2605.13709*.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.

Bin Tan, Nour Armoush, Elisabetta Mazzullo, Okan Bulut, and Mark Gierl. 2025. A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2):317–340.

Yongfeng Tao, Minqiang Yang, Hao Shen, Zhichao Yang, Ziru Weng, and Bin Hu. 2023. Classifying anxiety and depression through llms virtual interactions: a case study with chatgpt. In *2023 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 2259–2264. IEEE.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem.

2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):4.

Robert J Vandenberg and Charles E Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, 3(1):4–70.

Bushi Xiao and Qian Shen. 2026. Personality-driven student agent-based modeling in mathematics education: How well do student agents align with human learners? *arXiv preprint arXiv:2603.21358*.

Shihao Xu, Yiming Yan, Yanli Ding, Feng Li, Shu Zhang, Haoyun Tang, Chao Luo, Yan Li, Hao Liu, Yu Mei, and 1 others. 2025. Identifying psychiatric manifestations in outpatients with depression and anxiety: a large language model-based approach. *npj Mental Health Research*, 4(1):63.

Yuqing Zhao, Wei Qian, Yaru Chen, Donghong Wu, Yujia Luo, Cong Gao, Kankan Wu, and Zhengkui Liu. 2025. Effect of an ai agent trained on a large language model (llm) as an intervention for depression and anxiety symptoms in young adults: A 28-day randomized controlled trial. *Applied Psychology: Health and Well-Being*, 17(5):e70067.

Bruno D Zumbo. 1999. A handbook on the theory and methods of differential item functioning (dif). *Ottawa: National Defense Headquarters*, 160:53.

A Appendix

A.1 Psychometric Results for PHQ-9 and GAD-7

Here we report the full psychometric evaluation results for the sampled human data and the four open-weight LLMs on PHQ-9 and GAD-7.

Table 1: Psychometric evaluation results for the sampled human data and four open-weight LLMs on GAD-7.

Model	SRMR	CFI	TLI	RMSEA	Cronbach's α	Partial Metric	Partial Scalar	DIF Items
Human	0.017	0.972	0.958	0.047	0.97	–	–	–
Llama-3.3-70B	0.088	0.724	0.586	0.364	0.93	Pass	Fail	3
GPT-OSS-120B	0.038	0.926	0.889	0.178	0.94	Fail	Fail	3
Llama-3.1-8B	0.056	0.878	0.817	0.216	0.92	Fail	Fail	5
Mistral-7B	0.051	0.901	0.907	0.173	0.90	Pass	Fail	4

Table 2: Psychometric evaluation results for the sampled human data and four open-weight LLMs on PHQ-9.

Model	SRMR	CFI	TLI	RMSEA	Cronbach's α	Partial Metric	Partial Scalar	DIF Items
Human	0.042	0.931	0.908	0.054	0.84	–	–	–
Llama-3.3-70B	0.060	0.819	0.758	0.272	0.96	Pass	Fail	7
GPT-OSS-120B	0.039	0.919	0.891	0.167	0.96	Fail	Fail	8
Llama-3.1-8B	0.056	0.897	0.863	0.124	0.86	Fail	Fail	8
Mistral-7B	0.057	0.918	0.891	0.151	0.92	Fail	Fail	7

Table 3: Measurement invariance across prompt conditions for LLM-generated GAD-7 responses.

Model	Invariance Level	CFI	TLI	RMSEA	SRMR	ΔCFI	Status
llama-3.3-70b-instruct	Configural	0.658	0.487	0.526	0.129	-	-
llama-3.3-70b-instruct	Metric	0.557	0.453	0.543	0.322	-0.101	Fail
llama-3.3-70b-instruct	Scalar	0.491	0.466	0.536	0.346	-0.066	Fail
gpt-oss-120b	Configural	0.910	0.866	0.218	0.037	-	-
gpt-oss-120b	Metric	0.899	0.876	0.210	0.084	-0.011	Fail
gpt-oss-120b	Scalar	0.850	0.842	0.236	0.099	-0.049	Fail
llama-3.1-8b-instruct	Configural	0.840	0.761	0.266	0.054	-	-
llama-3.1-8b-instruct	Metric	0.809	0.764	0.264	0.119	-0.031	Fail
llama-3.1-8b-instruct	Scalar	0.785	0.774	0.258	0.126	-0.024	Fail
mistral-7b-instruct	Configural	0.869	0.804	0.252	0.050	-	-
mistral-7b-instruct	Metric	0.839	0.802	0.253	0.099	-0.030	Fail
mistral-7b-instruct	Scalar	0.797	0.787	0.262	0.128	-0.042	Fail

Table 4: Measurement invariance across prompt conditions for LLM-generated PHQ-9 responses.

Model	Invariance Level	CFI	TLI	RMSEA	SRMR	ΔCFI	Status
llama-3.3-70b-instruct	Configural	0.718	0.624	0.419	0.106	-	-
llama-3.3-70b-instruct	Metric	0.590	0.524	0.471	0.196	-0.128	Fail
llama-3.3-70b-instruct	Scalar	0.572	0.560	0.453	0.201	-0.018	Fail
gpt-oss-120b	Configural	0.900	0.867	0.212	0.035	-	-
gpt-oss-120b	Metric	0.868	0.847	0.228	0.140	-0.032	Fail
gpt-oss-120b	Scalar	0.826	0.821	0.246	0.151	-0.042	Fail
llama-3.1-8b-instruct	Configural	0.875	0.833	0.145	0.060	-	-
llama-3.1-8b-instruct	Metric	0.838	0.811	0.155	0.107	-0.037	Fail
llama-3.1-8b-instruct	Scalar	0.766	0.759	0.175	0.123	-0.072	Fail
mistral-7b-instruct	Configural	0.921	0.895	0.163	0.051	-	-
mistral-7b-instruct	Metric	0.895	0.878	0.176	0.120	-0.026	Fail
mistral-7b-instruct	Scalar	0.851	0.846	0.197	0.147	-0.044	Fail

A.2 Persona Definitions

To improve methodological transparency without including full implementation details, we summarize here the baseline persona settings used in the prompt construction process for PHQ-9 and GAD-7 generation.

A.2.1 PHQ-9 Persona Definitions

Table 5: Baseline persona definitions used for PHQ-9 response generation.

Persona	Intended severity	Description
Healthy	low symptom level	A completely healthy adult with a regular routine, stable mood, and generally positive outlook on life.
Mild Depression	moderate symptom level	An adult with mild depressive symptoms who occasionally feels tired or down but can still manage daily work and life with effort.
Severe Depression	high symptom level	An adult experiencing a severe major depressive episode, characterized by persistent hopelessness, low energy, loss of interest, and substantial impairment in daily functioning.

A.2.2 GAD-7 Persona Definitions

Table 6: Baseline persona definitions used for GAD-7 response generation.

Persona	Intended severity	Description
Healthy	low symptom level	A completely healthy adult with a regular routine, a calm mind, and little tendency to worry unnecessarily.
Mild Anxiety	moderate symptom level	An adult with mild anxiety who occasionally feels nervous or worries about everyday matters but can still manage daily life.
Severe Anxiety	high symptom level	An adult experiencing severe generalized anxiety, characterized by persistent tension, uncontrollable worry, a sense of impending doom, and substantial impairment in daily functioning.

A.3 Situational Modifiers

In addition to baseline personas, each generation prompt incorporated a short situational modifier sampled from a predefined modifier pool. These modifiers were used to introduce controlled within-person fluctuation while preserving the overall symptom profile implied by the baseline persona.

A.3.1 PHQ-9 Situational Modifiers

Table 7: Situational modifiers used for PHQ-9 response generation.

No.	Modifier text
1	Even though this is your baseline state, you had a relatively peaceful morning today.
2	Your mood has been fluctuating a bit more than usual this week.
3	You've been trying to distract yourself with routines recently, with mixed success.
4	You are just going about an ordinary, uneventful day.
5	You recently had a pleasant but brief interaction with a neighbor.
6	You had a poor night's sleep which is making everything feel a bit more intense today.
7	You are feeling a slight, unexpected shift in your energy levels right now.
8	You feel slightly more easily irritated or distracted than usual today.

A.3.2 GAD-7 Situational Modifiers

A.4 Prompt Assembly Summary

For each generation condition, the final prompt was constructed by combining four components: (1) a baseline persona definition corresponding to the target symptom severity, (2) a short situational modifier drawn from the predefined modifier pool, (3) the questionnaire completion instruction for PHQ-9 or

Table 8: Situational modifiers used for GAD-7 response generation.

No.	Modifier text
1	Even though this is your baseline state, you had a relatively calm morning today.
2	Your anxiety levels have been fluctuating a bit more than usual this week.
3	You just received an ambiguous email from your boss, which is making you overthink.
4	You are just going about an ordinary, uneventful day.
5	You recently had a pleasant and relaxing conversation with a friend.
6	You had too much coffee today, making you feel physically jittery and more restless than usual.
7	You are experiencing a rare, fleeting moment of complete relief right now.
8	You feel slightly more easily startled or on edge than usual today.

GAD-7, and (4) a structured output constraint requiring item-level responses in JSON format. This design allowed us to preserve the overall symptom profile implied by the persona while introducing controlled within-condition variability through the situational modifier.

A.5 Code

The code we used to generate responses to the scales using LLMs is available at the following link https://osf.io/d5wrm/overview?view_only=889cc6bd0aed44659fd5aa2ee98c675a.