



Validation And Ethics of the use of Large Language Models (Llms) as Affective Simulation Models in Psychopathology Research

Sanya Tri Aqilla^{1*}, & Chintia Adinda Prameswari²

^{1*}Poltekkes Kemenkes Padang, Indonesia, ²Poltekkes Kemenkes Padang, Indonesia

*Co e-mail: sanyaaqilla@gmail.com¹

Article Information

Received: March 19, 2026

Revised: April 21, 2026

Online: May 07, 2026

Keywords

Large Language Model, Agreement Validity, Affective Simulation, Artificial Intelligence Ethics, Psychopathology Research

ABSTRACT

Large Language Models (LLMs) are increasingly used as affective simulation models in psychopathology research, yet their psychometric validity and safety-related ethical implications remain underexamined. This study evaluated agreement between affective severity scores generated by three LLMs and human reference scores for depression, anxiety, and stress, while examining the effects of Baseline, Grounded, and Safety prompting conditions on simulation fidelity and bias. Using a quantitative 3×3 factorial design, 720 human–LLM score pairs were analyzed through correlation, Intraclass Correlation Coefficient (ICC), Bland–Altman analysis, error metrics, ANOVA, and chi-square tests. Results showed excellent agreement between human and LLM scores (ICC = .87–.89; $r = .87-.89$; $p < .001$) with no significant overall systematic bias ($p > .30$). However, prompting condition significantly affected simulation error (partial $\eta^2 = .167$; $p < .001$), with Grounded yielding the lowest and Baseline the highest error. Safety produced significant negative bias for depression (bias = -0.42 ; $p = .025$), indicating symptom underestimation. These findings demonstrate high agreement validity while highlighting potential ethical risks of safety-related symptom attenuation.

INTRODUCTION

The rapid advancement of Large Language Models (LLMs) in recent years has opened a new chapter in clinical psychology and psychopathology research. LLMs are no longer solely used as conversational aids, but are increasingly being used as patient simulation models (patient simulators) that can portray individuals with specific emotional disorders for clinician training, assessment instrument development, and as research proxies in clinical populations that are difficult to recruit directly due to ethical considerations, cost, or case rarity. Frameworks such as PATIENT-

Ψ have demonstrated that LLMs can produce expert rated symptom presentations that more closely mimic real life patient interactions than generic models, significantly improving trainees' therapeutic competence in repeated simulation sessions (Wang et al., 2024). These developments position LLMs as a tool with significant potential to address the limited access to diverse clinical cases in psychology education and research.

Despite their significant potential, the use of LLMs as affective simulation models raises two fundamental, interrelated issues: validity and ethics. In terms of validity, the primary question is the extent to which the affective scores or presentations generated by LLMs truly represent the experience and severity of symptoms equivalent to those of humans. A recent systematic review of LLMs in mental health services found that while LLMs demonstrate promising performance in detecting mental health problems, comprehensive assessment of their reliability and validity beyond text detection tasks is limited (Guo et al., 2024; see also Bucher et al., 2025). The recently developed PSI Bench evaluation framework revealed that LLM based depression patient simulators tend to produce overly long and lexically rich responses, exhibit reduced emotional variability, and follow a uniform and less naturalistic pattern of emotional transitions than real patients (Hoang et al., 2026), an indication that the affective fidelity of LLM simulations cannot be assumed without careful empirical testing.

From an ethical perspective, a key issue is how the safety alignment process additional training directed at LLMs avoiding potentially harmful outputs, including content related to severe psychological distress could inadvertently distort the fidelity of the affective simulations themselves. PSI Bench explicitly notes that simulating depression presents a particular challenge because LLM safety training limits their ability to authentically mimic depressive speech patterns (Hoang et al., 2026; Liu et al., 2025). This means that mechanisms designed to protect users from potential harm have the potential to create systematic bias in the form of underestimating the severity of severe psychopathology symptoms in research and clinical simulation contexts an ethical paradox where efforts to maintain safety can compromise scientific validity and the quality of clinician training.

Other ethical concerns have been raised in recent studies evaluating LLM based chatbots in the broader mental health context. Research from Brown University found that AI chatbots systematically violate core mental health ethical standards, including over validating inaccurate user beliefs and using deceptive empathy that can foster a false emotional attachment to a system that lacks sentience (Iftikhar et al., 2025). Another study developing the EmoAgent framework reported that emotionally engaging dialogues with LLM-based characters led to psychological deterioration in more than 34% of simulations involving vulnerable users, underscoring the need for additional safeguards in human–AI interactions in mental health (Qiu et al., 2025). These findings reinforce the urgency of empirically examining how prompting strategies and safety alignment impact the balance between simulation fidelity and the safety of LLM use in psychopathology research.

Most studies on LLM-based patient simulators to date have used qualitative evaluation approaches or ratings by other LLMs acting as judges (LLM as judge), which the PSI Bench study



found are not always reliable and can show trends that contradict human expert assessments (Hoang et al., 2026). Quantitative psychometric approaches that directly compare LLM affective severity scores with human reference scores using standard agreement metrics such as the Intraclass Correlation Coefficient (ICC) and Bland-Altman analysis, commonly used in the validation of psychological and medical instruments (Bland & Altman, 1986; Koo & Li, 2016) have rarely been systematically applied, especially when manipulating prompting conditions as an experimental variable to test the effect of safety alignment on validity. This gap is what this study attempts to address, by combining rigorous psychometric agreement validity testing and experimental manipulation of prompting conditions to examine both the validity and the ethical implications of LLM-based affective simulation. Without such evidence, simulated patients risk being adopted for clinician training and research on the strength of fluent output alone, and any systematic distortion introduced by safety alignment would remain undetected. The findings are therefore expected to carry three implications. Theoretically, they extend psychometric validation practice to LLM-based affective simulation by combining agreement analysis with experimental manipulation of prompting conditions. Methodologically, they offer a reproducible, human-referenced alternative to LLM-as-judge evaluation that can be reused across simulators and prompting strategies (Hoang et al., 2026; Sabour et al., 2026). Practically and ethically, they can inform researchers, educators, and developers about how prompting and safety-alignment choices shape the fidelity of simulated patients used for clinician training and for testing conversational agents (Shawon et al., 2026; Stade et al., 2024).

METHODS

This study used a quantitative design with a psychometric validation approach (agreement/validation study) combined with a 3×3 factorial between-subjects-case design, namely three Large Language Model models (LLM-A, LLM-B, LLM-C) crossed with three prompting conditions (Baseline, Grounded, and Safety). This design allows testing the validity of the agreement between LLM simulation scores and human reference scores as well as testing the influence of models and prompting conditions on simulation fidelity experimentally.

The unit of analysis in this study was the affective simulation case, a pair of psychopathology severity scores consisting of a human-anchored score and a LLM score for an identical case. Each model × condition combination yielded 80 cases, resulting in a total of 3 models × 3 conditions × 80 cases = 720 human–LLM score pairs. The human-anchored score was derived from a structured case profile that served as the ground truth for the three psychopathology domains measured, while the LLM score was derived from the output of each model when presented with the same case profile under one of three prompting conditions. The Baseline condition represented general instructions without explicit clinical guidelines or additional safety instructions; the Grounded condition represented instructions explicitly anchored to structured clinical criteria and rubrics to guide symptom severity estimation; and the Safety condition represented additional safety-oriented and

risk mitigation instructions (safety-alignment framing) commonly applied to LLM systems used in mental health contexts.

Three domains of psychopathology were measured in each case: depression (dep), anxiety (anx), and stress (stress), each measured using either a human-referenced version of the scores (dep_h, anx_h, stress_h) or a simulated LLM score (dep_llm, anx_llm, stress_llm). The depression score used a range of 0–27, equivalent to the range of depression severity instruments commonly used in clinical practice (e.g., the Patient Health Questionnaire-9). The anxiety score used a range of 0–21, equivalent to the range of anxiety severity instruments commonly used (e.g., the Generalized Anxiety Disorder-7). The stress score used a range of 0–27 as a parallel index of stress severity. Higher scores on all three domains indicate greater symptom severity.

Data analysis was performed using the Python 3 programming language with the pandas, SciPy, statsmodels, pingouin, and scikit-learn libraries. The first stage involved a Shapiro-Wilk normality test on human scores, LLM scores, and difference scores for each domain to determine the appropriate statistical approach for the subsequent analysis. The Shapiro-Wilk test was preferred over the Kolmogorov-Smirnov test because it has been shown to have greater statistical power to detect departures from normality across a wide range of distributions and sample sizes, whereas the Kolmogorov-Smirnov test has low power and is not recommended when the mean and standard deviation are estimated from the sample, even with the Lilliefors correction (Ghasemi & Zahediasl, 2012; Razali & Wah, 2011). Because normality tests become oversensitive to trivial deviations in large samples ($N = 720$), both parametric (Pearson correlation, paired t-test) and nonparametric (Spearman correlation, Wilcoxon signed-rank test) procedures were reported, so that the conclusions did not depend on the normality assumption (Ghasemi & Zahediasl, 2012).

Agreement validity between human scores and LLM scores was tested using three complementary approaches. First, Pearson and Spearman correlations were calculated to assess the strength of the linear and monotonic relationship between the two score sources. Second, the Intraclass Correlation Coefficient (ICC) was calculated using a two-way random model with absolute agreement (ICC agreement, single measure) to assess the degree of agreement, taking into account both rank consistency and absolute value agreement between human and LLM scores a more rigorous approach than Pearson correlation alone because it also detects systematic shifts between score sources. Third, a Bland-Altman analysis was used to estimate the mean difference between LLM and human scores, along with 95% limits of agreement, both overall and by prompting condition, to identify patterns of systematic bias that may not be captured by the correlation coefficient.

To test the significance of the overall systematic bias, a paired-samples t-test and a nonparametric Wilcoxon signed-rank test were performed, with Cohen's d_z effect sizes reported. Next, the absolute error (LLM score – human score) was calculated as an index of simulation fidelity at the case level, summarized as the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for each model $\times$ condition combination. The effect of model and prompting condition on the mean absolute error was tested using a two-way ANOVA (model $\times$ condition), with effect sizes



reported as partial η^2 . The effect of prompting conditions on the direction of bias (not just the magnitude of the error) was further tested using a one-way ANOVA on the difference scores (LLM human) for each psychopathology domain.

Finally, to assess the equivalence of the extreme distribution patterns between human and LLM scores, the percentage of cases falling at the upper (ceiling) and lower (floor) limits of the scale for each domain was calculated, and the differences were tested using a chi-square test of proportions. All statistical tests used a significance level of $\alpha = 0.05$.

RESULTS

1. Descriptive Statistics and Normality Test

Table 1 presents descriptive statistics and the results of the Shapiro-Wilk normality test for the human scores and LLM scores across the three psychopathology domains. All variables showed significant deviations from a normal distribution ($p < 0.001$), so further analyses included a nonparametric approach as a comparison to the parametric tests.

Table 1. Descriptive Statistics and Normality Test of Human Scores and LLM Scores (N = 720)

Variabel	M	SD	Shapiro-Wilk (W)	p	Distribusi
Depression (Human)	13,950	5,677	0,994	0,007	Abnormal
Depression (LLM)	13,935	6,365	0,990	< 0,001	Abnormal
Anxiety (Human)	12,537	5,361	0,972	< 0,001	Abnormal
Anxiety (LLM)	12,427	5,689	0,966	< 0,001	Abnormal
Stress (Human)	17,300	6,477	0,969	< 0,001	Abnormal
Stress (LLM)	17,214	6,735	0,965	< 0,001	Abnormal

Descriptively, the mean human scores and LLM scores in the three domains appear very close: depression (human M = 13.95 vs. LLM M = 13.94), anxiety (human M = 12.54 vs. LLM M = 12.43), and stress (human M = 17.30 vs. LLM M = 17.21), indicating no significant mean differences at the aggregate level.

2. Validity of Agreement between Human Scores and LLM Scores

Table 2 and Figure 1 present the results of the correlation analysis and the Intraclass Correlation Coefficient (ICC) as an index of agreement validity. All three domains showed strong Pearson and Spearman correlations between human scores and LLM scores ($r = 0.88$ – 0.89 ; $p < 0.001$ for all domains). The ICC values which more strictly account for absolute value agreement, rather than simply ordinal consistency also indicated very good agreement across all three domains (ICC = 0.877 – 0.892), with 95% confidence intervals all above the 0.75 threshold commonly interpreted as “very good” agreement in the psychometric validation literature.

Table 2. Correlation Coefficient and Intraclass Correlation Coefficient (ICC) between Human Scores and LLM Scores

Domain	Pearson r	Spearman ρ	ICC (agreement)	95% CI ICC	p
Depression	0,882	0,881	0,877	[0,86, 0,89]	< 0,001
Anxiety	0,866	0,863	0,865	[0,85, 0,88]	< 0,001
Stress	0,892	0,886	0,892	[0,88, 0,91]	< 0,001

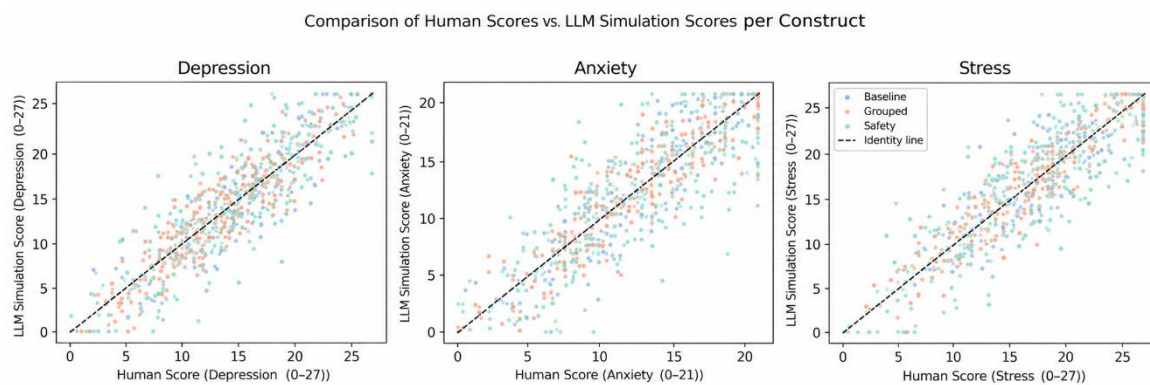


Figure 1. Scatterplot of the Concordance of Human Scores and LLM Simulation Scores per Psychopathology Domain, by Prompting Condition

3. Systematic Bias Analysis: Bland-Altman Test and Paired Difference Test

Bland-Altman analysis (Table 3, Figure 2) showed very small, near-zero mean bias across all three domains (bias = -0.015 to -0.110 on a 0–27/0–21 scale), with 95% limits of agreement ranging from -6 to +6 points. Paired t-tests confirmed no statistically significant mean differences between human scores and overall LLM scores across all three domains ($p > 0.30$ for all domains), with very small Cohen's d_z effect sizes ($|d_z| < 0.04$), indicating the absence of large systematic bias at the aggregate level.

Table 3. Results of Bland-Altman Analysis and Paired t-Test between Human Scores and LLM Scores

Domain	Bias (LLM–Human)	SD Difference	95% Lower Limit	95% Upper Limit	t	p (uji t)	d_z
Depression	-0,015	2,996	-5,886	5,857	-0,132	0,895	-0,005
Anxiety	-0,110	2,873	-5,741	5,522	-1,024	0,306	-0,038
Stres	-0,087	3,076	-6,116	5,943	-0,756	0,450	-0,028



4. Simulation Fidelity based on Model and Prompting Conditions

Although no significant systematic bias was found at the aggregate level, analysis of errors at the individual case level revealed a more complex pattern. Table 4 presents the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for each combination of model and prompting condition. Descriptively, the Grounded condition consistently produced the lowest MAE across all three domains across all models, while the Baseline condition consistently produced the highest MAE.

Table 4. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) based on Model and Prompting Conditions

Model	Condition	MAE Depression	MAE Anxiety	MAE Stres	Depression RMSE	Anxiety RMSE	Stress RMSE
LLM-A	Baseline	2,95	3,52	2,80	3,74	4,31	3,63
LLM-A	Grounded	2,02	1,48	1,80	2,49	1,91	2,36
LLM-A	Safety	1,94	1,86	2,61	2,61	2,52	3,17
LLM-B	Baseline	2,88	2,59	3,06	3,69	3,32	3,71
LLM-B	Grounded	2,06	1,64	1,69	2,51	2,28	2,24
LLM-B	Safety	2,09	1,88	2,23	2,64	2,40	2,85
LLM-C	Baseline	2,68	2,57	3,17	3,58	3,26	4,15
LLM-C	Grounded	1,83	1,78	1,54	2,33	2,33	1,92
LLM-C	Safety	2,32	2,10	2,26	2,91	2,77	2,88

The results of a two-way ANOVA on the combined mean absolute error of the three domains (Table 5) indicate that the main effect of prompting condition was statistically significant and practically large, $F(2, 711) = 71.44$, $p < 0.001$, partial $\eta^2 = 0.167$, indicating that prompting condition explained approximately 16.7% of the variance in simulation error. In contrast, the main effect of model, $F(2, 711) = 0.57$, $p = 0.564$, and the interaction effect of model $\times$ condition, $F(4, 711) = 0.84$, $p = 0.498$, were not significant, with very small effect sizes (partial $\eta^2 < 0.005$). These findings suggest that the fidelity of LLM affective simulations is determined much more by the prompting strategy used than by the specific LLM model employed.

5. Condition-Specific Bias Direction: Ethical Implications of Safety Alignment

To test whether the prompting condition not only affects the magnitude of the error but also the direction of the bias, a one-way ANOVA was conducted on the difference scores (LLM – human) for each domain by condition (Table 6). The results showed that the prompting condition had a significant effect on the direction of the bias in the depression domain, $F(2, 717) = 3.72$, $p = 0.025$, but was not significant in the anxiety ($p = 0.866$) or stress ($p = 0.503$) domains.

Table 6. Mean Bias (LLM – Human) by Prompting Condition and One-Way ANOVA Results

Domain	Bias Baseline	Bias Grounded	Bias Safety	F (ANOVA)	p
Depression	0,314	0,060	-0,418	3,722	0,025
Anxiety	-0,185	-0,045	-0,099	0,144	0,866
Stres	-0,275	-0,015	0,030	0,688	0,503

The pattern of bias direction in the depression domain shows a clinically meaningful gradient: the Baseline condition shows a positive bias (LLM scores slightly higher than human scores; bias = +0.31), the Grounded condition shows almost zero bias (bias = 0.060), while the Safety condition shows a clear negative bias (LLM scores lower than human scores; bias = -0.42). This pattern is consistent with the notion that safety-oriented instructions cause the model to underestimate the severity of simulated depressive symptoms, in line with the finding that LLM safety training can limit the model's ability to authentically imitate depressive speech patterns.

6. Effects of Ceiling and Floor on Human Scores and LLM Scores

Table 7 compares the percentages of cases at the ceiling and floor of the scale for human and LLM scores. The ceiling effect is most pronounced in the anxiety and stress domain (9–10% of human cases at the maximum score of the scale), reflecting the high-severity nature of the simulated case design. Chi-square tests indicate that the proportions of cases at the ceiling and floor were not significantly different between human and LLM scores in all three domains ($p > 0.30$ for all domains), indicating that the LLM generally replicates the distribution pattern of extreme human scores reasonably well, without a tendency for excessive compression of scores at the aggregate level.

Tabel 7. Comparison of Percentage of Cases on Ceiling and Floor Scale between Human Score and LLM Score

Domain	% Ceiling human	% Ceiling LLM	% Floor human	% Floor LLM	χ^2	p
Depression	1,25	2,08	0,28	1,53	1,059	0,303
Anxiety	9,03	8,19	0,97	1,25	0,221	0,639
Stres	10,14	8,89	0,00	1,39	0,516	0,472

DISCUSSION

This study aimed to test the validity and agreement of Large Language Models (LLMs) as affective simulation models in psychopathology research, while also investigating the ethical implications of aligning safety with the fidelity of such simulations. Overall, the findings supported the study's first hypothesis: affective severity scores from LLM simulations demonstrated excellent agreement with human benchmark scores across the three psychopathology domains tested, with



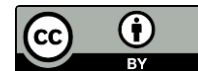
both correlation coefficients ($r = 0.87\text{--}0.89$) and Intraclass Correlation Coefficients ($ICC = 0.87\text{--}0.89$), all exceeding the "very good" threshold for agreement (Cicchetti, 1994). These findings provide quantitative evidence supporting the potential of LLMs as valid affective simulation models, consistent with reports that LLM-based patient simulation frameworks such as PATIENT- Ψ are rated by experts as more closely approximating real-life patient interactions than generic models (Wang et al., 2024), and reports that ChatGPT even demonstrated superior performance compared to humans in evaluating emotional awareness in some testing contexts (Elyoseph et al., 2023).

The absence of significant systematic bias at the aggregate level (paired t -test, $p > 0.30$ across all three domains; Cohen's $d_z < 0.04$) is an important finding supporting the convergent validity of the LLM as a general affective simulation model. However, these results should be interpreted with caution, as seemingly stable patterns of agreement at the aggregate level may mask significant variation at the individual condition and case levels as further analyses of this study found (Bland & Altman, 1986). The prompting condition was shown to have a significant and practically significant effect on the magnitude of simulation error (partial $\eta^2 = 0.167$), with the Grounded condition explicitly anchored to structured clinical criteria producing the highest fidelity, while the Baseline condition, without explicit guidance, produced the greatest error. Under the more conservative benchmarks of Koo and Li (2016), which grade reliability by the lower bound of the 95% confidence interval and reserve "excellent" for values above 0.90, the agreement observed here would be classified as good rather than excellent.

This pattern is consistent with the idea that the prompting strategy plays a more important role than the choice of the specific LLM model in determining simulation quality, as confirmed by the PSI-Bench study, which found that the simulation framework (rather than the underlying LLM model) plays a more decisive role in simulator behavior (Hoang et al., 2026). The most ethically relevant finding in this study is the condition-specific bias pattern in the depression domain, where the Safety condition exhibited a significant negative bias (underestimation) compared to the Baseline and Grounded conditions. This pattern directly supports the study's third hypothesis and aligns with the finding that safety training in LLM inherently limits the model's ability to authentically reproduce depressive speech patterns and symptom presentations, as safety mechanisms tend to encourage more cautious, positive, and less emotionally intense output (Hoang et al., 2026; Liu et al., 2025).

The implications are ethically double-edged: on the one hand, safety alignment is crucial to prevent LLM from generating content that could potentially trigger or exacerbate psychological distress in vulnerable users, as highlighted in numerous studies on the risks of mental health chatbots (Iftikhar et al., 2025; Moore et al., 2025; Rahsepar Meadi et al., 2025); but on the other hand, these same mechanisms have the potential to distort scientific validity when LLM is used as a simulation tool for research or clinical training purposes, as the severity of severe symptoms the very ones most crucial for future clinicians to accurately recognize is likely to be attenuated.

This ethical paradox is of particular relevance to researchers and developers of LLM-based patient simulation systems. The finding that LLM-based chatbots frequently violate core mental



health ethical standards including over-validation and deceptive empathy (Iftikhar et al., 2025) underscores the importance of safety alignment in the context of direct interactions with vulnerable users. However, this research suggests that applying the same safety alignment to simulated clinical research and training contexts which have fundamentally different goals than direct support interactions can lead to undesirable validity trade-offs. This suggests the need to differentiate LLM safety policies based on the context of use: strict safety mechanisms may be appropriate for chatbots interacting directly with users in distress, but require adjustments or ethically governed exceptions when LLM is used as a controlled simulation tool for research and clinical education purposes that do not directly involve vulnerable users. Such context-specific governance is consistent with recent calls for clinically grounded, use-case-specific evaluation and responsible development of LLMs in behavioral healthcare (Stade et al., 2024).

In contrast, the consistent superiority of the Grounded condition in producing the highest simulation fidelity offers a promising compromise: explicit anchoring to structured clinical criteria appears to enhance simulation validity without requiring a relaxation of the safety mechanisms themselves. These findings align with recent work that builds simulated patients from structured clinical profiles, diagnostic criteria, and symptom-severity information and validates them against expert judgment (Liu et al., 2025; Wang et al., 2025). They also support recent calls for standardized, reproducible simulation and evaluation pipelines so that simulators can be compared fairly across methods and models (Hoang et al., 2026; Sabour et al., 2026). However, because the present study evaluated severity scores rather than dialogue-level linguistic behavior, it remains to be tested whether grounding also restores the linguistic authenticity that current simulators lack (Hoang et al., 2026; Liu et al., 2025).

CONCLUSIONS

This study demonstrates that Large Language Models (LLMs) have excellent agreement validity as affective simulation models in psychopathology research, with high correlations and Intraclass Correlation Coefficients (r and $ICC = 0.87\text{--}0.89$) between LLM-simulated severity scores and human reference scores in the domains of depression, anxiety, and stress, without significant systematic bias at the aggregate level. However, simulation fidelity was significantly affected by the prompting strategy used, with the Grounded condition yielding the highest accuracy and the Baseline condition yielding the greatest error. The most ethically important finding was that the Safety condition yielded a significant negative bias in the depression domain, indicating that safety alignment may cause LLMs to systematically underestimate the severity of severe psychopathology symptoms.

These findings highlight the inherent ethical trade-off between safety and validity in the use of LLMs as affective simulation models: safety mechanisms designed to protect users potentially distort the simulation fidelity required for research and clinical training purposes. Theoretically, this study enriches the psychometric validation literature on artificial intelligence in clinical psychology with a quantitative testing framework that combines agreement analysis (ICC, Bland-Altman) and



experimental manipulation of prompting conditions. Practically, the findings recommend the use of grounded-prompting strategies anchored in structured clinical criteria as a practical mitigation to enhance simulation validity without compromising safety mechanisms, and emphasize the need for differentiated safety alignment policies based on the context of LLM use between direct interactions with vulnerable users and use as a controlled simulation tool in clinical research and education. Further research involving human clinical validators, a broader range of psychopathology domains, and a wider variety of models and prompting strategies is needed to strengthen the generalizability of these findings.

REFERENCES

- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476), 307–310. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8)
- Bucher, A., Egger, S., Vashkite, I., Wu, W., & Schwabe, G. (2025). "It's not only attention we need": Systematic review of large language models in mental health care. *JMIR Mental Health*, 12, e78410. <https://doi.org/10.2196/78410>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>
- Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14, 1199058. <https://doi.org/10.3389/fpsyg.2023.1199058>
- Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2), 486–489. <https://doi.org/10.5812/ijem.3505>
- Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health*, 11, e57400. <https://doi.org/10.2196/57400>
- Hoang, N. K., Mehri, S., Hsu, T.-A., Sun, Y.-J., Truong, Q. X. N., Doan, K. D., & Hakkani-Tür, D. (2026). PSI-Bench: Towards clinically grounded and interpretable evaluation of depression patient simulators. *arXiv preprint arXiv:2604.25840*. <https://doi.org/10.48550/arXiv.2604.25840>
- Iftikhar, Z., Xiao, A., Ransom, S., Huang, J., & Suresh, H. (2025). How LLM counselors violate ethical standards in mental health practice: A practitioner-informed framework. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2), 1311–1323. <https://doi.org/10.1609/aies.v8i2.36632>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>



- Li, B., Jin, B., Lan, K., Wang, M., & Wu, M. (2026). Synthetic or authentic? Building mental patient simulators from longitudinal evidence. arXiv preprint arXiv:2603.22704. <https://doi.org/10.48550/arXiv.2603.22704>
- Liu, S., Brie, B., Li, W., Biester, L., Lee, A., Pennebaker, J., & Mihalcea, R. (2025). Eeyore: Realistic depression simulation via expert-in-the-loop supervised and preference optimization. Findings of the Association for Computational Linguistics: ACL 2025, 13750–13770. <https://doi.org/10.18653/v1/2025.findings-acl.707>
- Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), 599–627. <https://doi.org/10.1145/3715275.3732039>
- Qiu, J., He, Y., Juan, X., Wang, Y., Liu, Y., Yao, Z., Wu, Y., Jiang, X., Yang, L., & Wang, M. (2025). EmoAgent: Assessing and safeguarding human-AI interaction for mental health safety. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 11741–11756. <https://doi.org/10.18653/v1/2025.emnlp-main.594>
- Rahsepar Meadi, M., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., & Batelaan, N. (2025). Exploring the ethical challenges of conversational AI in mental health care: Scoping review. JMIR Mental Health, 12, e60432. <https://doi.org/10.2196/60432>
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21–33.
- Sabour, S., Ng, T., & Huang, M. (2026). PatientHub: A unified framework for patient simulation. arXiv preprint arXiv:2602.11684. <https://doi.org/10.48550/arXiv.2602.11684>
- Sedlakova, J., & Trachsel, M. (2023). Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? The American Journal of Bioethics, 23(5), 4–13. <https://doi.org/10.1080/15265161.2022.2048739>
- Shawon, M. T. R., Irbaz, M. S., Elyazori, H. R. A., Resapu, K. R., Lin, Y., Cardenas, V. F., Alemi, F., & Lybarger, K. (2026). Advancing AI trustworthiness through patient simulation: Risk assessment of conversational agents for antidepressant selection. arXiv preprint arXiv:2602.11391. <https://doi.org/10.48550/arXiv.2602.11391>
- Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. npj Mental Health Research, 3(1), 12. <https://doi.org/10.1038/s44184-024-00056-z>
- Wang, R., Milani, S., Chiu, J. C., Zhi, J., Eack, S. M., Labrum, T., Murphy, S. M., Jones, N., Hardy, K. V., Shen, H., Fang, F., & Chen, Z. (2024). PATIENT-Ψ: Using large language models to simulate patients for training mental health professionals. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 12772–12797. <https://doi.org/10.18653/v1/2024.emnlp-main.711>



This work is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/)

Research and Advancement in Psychological Science and Outcomes (RASO)

Vol. 01, No. 1, May 2026

- Wang, X., Pérez, A., Parapar, J., & Crestani, F. (2025). TalkDep: Clinically grounded LLM personas for conversation-centric depression screening. Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), 6554–6558. <https://doi.org/10.1145/3746252.3761617>
- Xu, Y., Fang, Z., Lin, W., Jiang, Y., Jin, W., Balaji, P., Wang, J., & Xia, T. (2025). Evaluation of large language models on mental health: From knowledge test to illness diagnosis. Frontiers in Psychiatry, 16, 1646974. <https://doi.org/10.3389/fpsy.2025.1646974>