Large Language Models as Instrument-Aware Guardrails for Structured Mental Health Questionnaires

Rafhael da S. Monteiro, Carlos Augusto M. Oeiras, Jefferson M. de Morais

[1]Universidade Federal do Pará (UFPA) – Belém – Pará – Brasil

rafhaelsmonteiro@gmail.com, cm.oeiras01@gmail.com, jmorais@ufpa.br

Abstract. *Psychometric instruments such as PHQ-9, GAD-7, and BDI-II are increasingly integrated into digital assessment systems, where preserving instrument-specific response constraints is essential for reliable data collection. This work investigates whether Large Language Models (LLMs) can operate as instrument-aware guardrails for validating responses to structured mental health questionnaires. A controlled synthetic dataset containing structurally valid and invalid responses was evaluated independently by GPT-4o-mini, DeepSeek-V3, and LLaMA3 using a common instrument-specific validation protocol. The models achieved substantial to almost-perfect inter-model agreement (Fleiss' κ: 0.795–0.933), while a deterministic rule-based baseline achieved 1.000 accuracy across all three instruments. These findings indicate that the evaluated LLMs can consistently apply explicit instrument-specific validity criteria under controlled conditions. However, the perfect performance of the deterministic baseline indicates that the present benchmark does not establish semantic superiority of LLMs over rule-based validation. The results therefore provide evidence for the feasibility of LLM-based instrument-aware validation as a complementary layer in digital assessment systems, while motivating future evaluation with genuinely free-form and semantically variable responses.*

1. Introduction

Psychometric instruments play a central role in psychological assessment, where reliable measurement depends on the consistent application of their defined response and scoring procedures. As highlighted by [Brickman et al. 2025], rigorous measurement remains essential, even though traditional self-report approaches suffer from limitations such as response biases and restricted self-awareness. Recent work has also started treating language itself as a relevant source of psychological information capable of capturing nuances that standard instruments may miss [Kjell et al. 2024].

This is a critical issue in mental health. Instruments such as the PHQ-9 [Kroenke et al. 2001], GAD-7 [Spitzer et al. 2006], and BDI-II [Wang and Gorenstein 2013] are widely used for symptom screening and assessment. Increasing tension arises when these instruments are transferred to technology-mediated systems: psychometric validity depends, in part, on responses remaining compatible with the response options and scoring rules defined by each instrument. Digital interactions, however, may introduce deviations from expected response formats, creating challenges for automated validation.

Studies have linked linguistic markers to depression, and Transformer-based models have estimated symptom severity from interview transcripts [Eichstaedt et al. 2018,

Ohse et al. 2024]. However, this line of research primarily concerns inference of psychological states from language. A different and more constrained problem is determining whether an individual response is admissible under the response structure and scoring rules of a standardized instrument. Ensuring the validity of structured questionnaire data before scoring remains comparatively underexplored.

Technology-mediated assessments may involve partially controlled interactions in which responses deviate from the canonical formats expected by standardized instruments. Such deviations can include structurally invalid, evasive, inconsistent, or otherwise non-admissible responses, creating a need for automated validation mechanisms. Deterministic rule-based methods can provide highly reliable validation when responses follow predefined formats, whereas language-model-based approaches offer an alternative mechanism for interpreting responses within an instrument-specific context. The extent to which LLMs provide an advantage over deterministic validation, particularly for genuinely free-form responses, remains an open empirical question.

Whether LLMs can function as instrument-aware validators—assessing whether a response is admissible within the constraints of a specific questionnaire—has received comparatively little attention. This is a narrower task than inferring psychological states: it involves determining whether a response is compatible with an instrument's predefined response options, structural requirements, and scoring rules. Such validation is important because invalid or inconsistent responses can compromise the integrity of downstream scoring and analysis.

To address this, we investigate whether LLMs can serve as automated, instrument-aware validation mechanisms for the PHQ-9, GAD-7, and BDI-II. Using a controlled experimental setup with synthetic response generation and independent model auditing, we examine whether different models can reliably distinguish between records that conform to, or violate, predefined instrument-specific response constraints. Validity is determined at the item level and aggregated to the record level according to explicit deterministic criteria, without any focus on symptom prediction or clinical diagnosis.

This work makes three contributions: it frames LLMs as instrument-aware validation mechanisms for structured psychometric assessment; it introduces a controlled experimental framework combining synthetic data generation, programmatic ground truth, and a deterministic rule-based baseline; and it provides empirical evidence of high inter-model reliability in applying explicit instrument-specific validity criteria. Together, these contributions provide evidence for the feasibility of LLM-based validation as a complementary mechanism for preserving response-format integrity in digital mental health assessment systems, while also establishing the limitations of the present controlled benchmark and the need for evaluation with more linguistically variable responses.

2. Background and Related Work

2.1. Psychometric Instruments and Response Structures

Psychometric instruments are standardized measurement tools defined by items, response options, scoring procedures, and interpretation rules. This study considers the PHQ-9, GAD-7, and BDI-II [Kroenke et al. 2001, Spitzer et al. 2006, Wang and Gorenstein 2013]. PHQ-9 (9 items) and GAD-7 (7 items) use shared ordinal response schemes, whereas BDI-II (21 items) uses item-specific response options.

Thus, validity must be evaluated according to the specific instrument and, for BDI-II, the individual item. This structural diversity makes the three instruments suitable for evaluating instrument-aware validation.

2.2. Inter-Rater Reliability

Inter-rater reliability measures agreement among independent evaluators. Cohen's Kappa accounts for agreement beyond chance [McHugh 2012]:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is observed agreement and P_e is agreement expected by chance. Fleiss' Kappa extends this framework to multiple evaluators. In this study, the categorical classifications produced by the three LLMs are treated as independent evaluations of the same validation task, and Fleiss' Kappa is used to quantify inter-model agreement.

2.3. Guardrails in Language Models

Guardrails are mechanisms designed to monitor, constrain, or regulate model inputs and outputs [Inan, H. et al. 2023, Dong et al. 2024]. Foundation models have also been investigated as guardrail mechanisms for evaluating or constraining model behavior [Nguyen et al. 2025]. Here, the concept is adapted to instrument-aware validation: a guardrail assesses whether questionnaire responses comply with the response options and structural requirements of a predefined psychological instrument, without performing diagnosis or symptom inference.

2.4. Deterministic and Rule-Based Validation

Deterministic rule-based approaches explicitly encode expected structures, mappings, and constraints, providing transparency, reproducibility, and deterministic behavior. When an instrument's response space is predefined, admissibility can be evaluated without probabilistic inference. Comparisons between LLM-based and rule-based approaches in other structured language-processing tasks illustrate different trade-offs between explicit formalization and language-model processing [Nair et al. 2025].

In this study, a deterministic rule-based validator is included as a baseline. Its purpose is to establish the performance obtainable when the instrument-specific validity criteria are explicitly encoded, providing a reference for interpreting the LLM-based results.

2.5. LLMs in Psychological Assessment

LLMs have increasingly been applied to psychological assessment, including the extraction and prediction of psychological information from language [Brickman et al. 2025, Kjell et al. 2024]. Studies have explored depression screening and symptom estimation using clinical or semi-clinical text [Eichstaedt et al. 2018, Ohse et al. 2024, McCoy et al. 2025, Teferra et al. 2025]. Moëll and Sand Aronsson [Moëll and Sand Aronsson 2026] also demonstrated the use of LLM-generated synthetic self-reports for predicting mental health scores.

These studies primarily concern psychological inference rather than validation of whether responses conform to the structural requirements of a predefined psychometric instrument.

2.6. Research Gap

Although both psychometric instruments and LLMs have been extensively studied in digital psychological assessment, comparatively little attention has been given to LLMs as instrument-aware validators of structured questionnaire responses. In particular, controlled comparisons across multiple models against an explicit deterministic baseline remain limited. This study addresses this gap by evaluating GPT-4o-mini, DeepSeek-V3, and LLaMA3 in applying predefined validity criteria to PHQ-9, GAD-7, and BDI-II responses.

3. Methodology

This study evaluates whether Large Language Models (LLMs) can operate as instrument-aware guardrails for detecting structural invalidity in standardized mental health questionnaires. The experimental pipeline comprised human validation of the response-generation protocol, controlled synthetic dataset generation, programmatic ground-truth assignment, multi-model auditing, rule-based baseline evaluation, and statistical comparison of model outputs. Figure 1 summarizes the workflow.

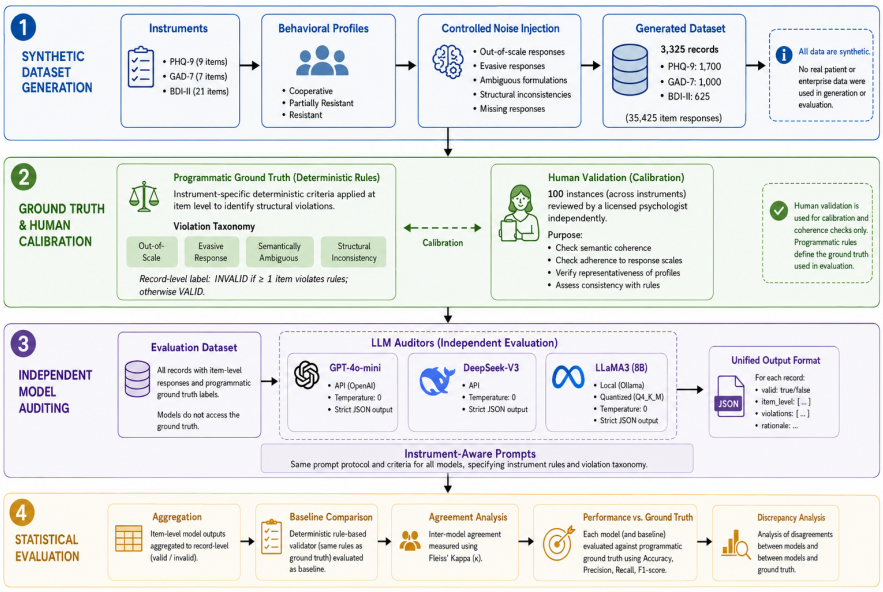


Figure 1. Overview of the methodological framework, comprising human validation, controlled synthetic generation, programmatic labeling, multi-model auditing, result consolidation, and statistical evaluation.

3.1. Study Design and Human Validation

The study used a controlled synthetic benchmark designed to evaluate structural validity rather than psychological diagnosis or symptom severity. Before large-scale generation, 100 simulated questionnaire instances covering PHQ-9, GAD-7, and BDI-II were reviewed by a psychologist. The review examined semantic coherence with the corresponding items, adherence to the admissible response scales or item-specific options, and the representation of cooperative and non-cooperative response patterns, including evasive and inconsistent responses.

This human review was used to verify that the generation protocol produced responses consistent with the intended validity categories and to refine the admissibility

criteria subsequently implemented programmatically. The 100 reviewed instances therefore served as a qualitative human validation of the generation protocol rather than as a statistical estimate of clinical prevalence or diagnostic validity.

3.2. Synthetic Dataset Generation

Following the human review, a controlled synthetic generation pipeline was implemented using a locally deployed LLaMA3 model. The dataset comprised 3,325 complete questionnaire records: 1,700 PHQ-9, 1,000 GAD-7, and 625 BDI-II records, corresponding to 35,425 individual item responses. The different record counts were determined pragmatically according to the number of responses required by each instrument and the desired overall response volume; they were not intended to reproduce population prevalence or a clinical sampling distribution.

Three behavioral profiles were generated: *cooperative*, *partially resistant*, and *resistant*. These profiles were used to introduce controlled deviations from the expected response format. The resulting dataset included valid responses as well as off-scale, evasive, ambiguous, irrelevant, missing, and structurally inconsistent responses. The controlled nature of the generation process allowed the presence or absence of each violation to be recorded independently of the subsequent auditing models.

3.3. Violation Taxonomy and Ground Truth

Structural violations were categorized into four types: (i) out-of-scale responses, in which a response did not correspond to an admissible option; (ii) evasive responses, in which the respondent did not provide the requested answer; (iii) semantically ambiguous expressions, whose meaning could not be reliably mapped to an admissible response category; and (iv) structural inconsistencies, such as missing items or incompatible response formats.

Ground-truth labels were assigned programmatically from the generation metadata and deterministic admissibility rules. For PHQ-9 and GAD-7, a response was valid when it corresponded to one of the four predefined ordinal categories. For BDI-II, validity was determined item by item according to its item-specific response options. A complete record was labeled *valid* only when every response satisfied the corresponding instrument constraints; the presence of at least one structural violation resulted in an *invalid* record.

Thus, the ground truth represents structural admissibility under the predefined experimental rules and should not be interpreted as a clinical diagnosis, psychological assessment, or severity label.

3.4. Multi-Model Auditing Protocol

Three LLMs independently audited the generated records: GPT-4o-mini (gpt-4o-mini-2024-07-18), DeepSeek-V3, and Meta Llama3 8B with Q4_0 quantization executed locally through Ollama. Except for instrument-specific admissibility rules, the validation prompt structure was kept consistent across models.

The prompt instructed each model to act as a questionnaire data validator and to determine whether the complete record was structurally valid. For PHQ-9 and GAD-7, the prompt specified the four-category ordinal response mapping. For BDI-II, it specified the item-specific response options and the corresponding intensity criteria. Models were instructed to return structured JSON containing the validity status, total score, alert indicator, and justification.

Inference temperature was set to zero or as close to zero as supported by the deployment environment. The auditing models did not receive the generation prompts or the programmatic labeling rules. Consequently, ground-truth labels were not inferred from model judgments but from the predefined generation and admissibility criteria.

LLaMA3 was used both during synthetic generation and as one of the auditing models. Although the two stages were implemented as separate experimental roles and the auditing model did not receive the generation or labeling rules, this reuse may introduce model-specific familiarity with the generated response distribution and is therefore considered a limitation of the study.

3.5. Rule-Based Baseline

A deterministic rule-based validator was implemented as a baseline using the same instrument-specific admissibility constraints employed to derive the reference labels. The baseline used strict matching against the predefined response categories and did not attempt semantic interpretation of linguistic variation.

Because the baseline implements the same deterministic rules used to construct the reference labels, its perfect performance is expected by design and does not constitute evidence that rule-based validation generally outperforms or is equivalent to LLM-based validation. Its role was instead to provide a transparent lower-complexity reference for the controlled benchmark and to establish the performance ceiling under exact format compliance.

3.6. Result Consolidation and Statistical Evaluation

After auditing, model outputs were consolidated by instrument and normalized to a common binary representation (*valid*/*invalid*). Records with missing or inconsistent model classifications were excluded from the corresponding analysis sets. For inter-model agreement analyses, only records with valid classifications from all three models were retained, resulting in 1,695 PHQ-9, 984 GAD-7, and 623 BDI-II records. In contrast, model-versus-ground-truth metrics were computed independently for each model using all available records with valid classifications for that model.

Inter-model reliability was assessed using Cohen's Kappa for pairwise comparisons and Fleiss' Kappa across the three models. Unanimity was also computed as the proportion of records for which all models produced the same classification. Fleiss' Kappa confidence intervals were estimated using bootstrap resampling.

Model classifications were additionally compared against the programmatically derived ground truth using accuracy, precision, recall, and F1-score. The *valid* class was treated as the positive class. Because the class distributions were asymmetric, particularly for BDI-II and GAD-7, F1-score for the valid class was used as the primary class-specific summary measure, while accuracy, precision, and recall were retained as complementary indicators. All statistical analyses were performed separately for each instrument.

Importantly, the primary classification unit was the complete questionnaire record. Response-level validity was used to construct the record-level ground truth, such that a single invalid response was sufficient to make the complete record invalid. Consequently, agreement at the record level does not imply that models identified the same individual response-level violations.

3.7. Scope and Evaluation Design

The experiment was designed to evaluate structural validation under controlled linguistic conditions rather than clinical validity. The synthetic benchmark provides reproducible control over response categories and violations, while the human review of 100 instances provides a qualitative check of the generation protocol. However, the design does not establish clinical effectiveness or generalization to real-world patient responses.

The study therefore interprets model performance as evidence of the feasibility of LLM-based instrument-aware validation. In particular, high agreement with the programmatic ground truth indicates the ability of the evaluated models to reproduce the predefined notion of structural admissibility, rather than demonstrating diagnostic accuracy or universal superiority over deterministic validators.

4. Results

The final synthetic dataset comprised 3,325 complete questionnaire records: 1,700 PHQ-9, 1,000 GAD-7, and 625 BDI-II records, corresponding to 35,425 individual item responses. After output normalization and exclusion of records with missing or inconsistent model classifications, the validated analysis sets comprised 1,695 PHQ-9, 984 GAD-7, and 623 BDI-II records. GPT-4o-mini, DeepSeek-V3, and a locally deployed LLaMA3 independently audited these records. The results examine inter-model agreement, model-specific classification distributions, correspondence with the programmatically derived ground-truth labels, and performance of the deterministic rule-based baseline.

4.1. Inter-Model Agreement

Fleiss' Kappa values were accompanied by 95% confidence intervals estimated via bootstrap resampling and are reported in Table 1. The estimates indicate substantial agreement for PHQ-9 and BDI-II and near-perfect agreement for GAD-7 among the three auditing models.

Table 1. Inter-model agreement by instrument

Instrument	Fleiss' Kappa (95% CI)	Unanimity (%)	Records Analyzed
PHQ-9	0.795 [0.765, 0.826]	90.15	1695
BDI-II	0.860 [0.798, 0.907]	95.18	623
GAD-7	0.933 [0.906, 0.960]	97.46	984

As shown in Table 1, the three models achieved substantial to near-perfect inter-model agreement across the evaluated instruments. GAD-7 showed the highest consistency, whereas PHQ-9 exhibited greater variation among model-level validation decisions.

4.2. Pairwise Agreement (Cohen's Kappa)

Table 2 reports pairwise Cohen's Kappa values between the three auditing models for each instrument. The results show substantial to near-perfect pairwise agreement, although the magnitude of agreement varied according to both the instrument and the model pair.

For PHQ-9, DeepSeek-V3 and LLaMA3 showed the highest pairwise agreement, whereas GPT-4o-mini exhibited lower agreement with both models. BDI-II and GAD-7

Table 2. Pairwise inter-model agreement measured by Cohen's Kappa for each instrument

Instrument	GPT × DeepSeek	GPT × LLaMA	DeepSeek × LLaMA
PHQ-9	0.744	0.711	0.964
BDI-II	0.899	0.825	0.860
GAD-7	0.983	0.901	0.918

showed consistently high pairwise agreement, with GAD-7 presenting the highest values overall.

4.3. Distribution of Model Classifications

Table 3 presents the distribution of record-level validity classifications produced by each auditing model. These distributions provide a complementary view of model behavior by showing the proportion of questionnaire records accepted or rejected under the structural validity criteria.

Table 3. Percentage distribution of valid and invalid classifications by model and instrument

Instrument	Model	Valid (%)	Invalid (%)
PHQ-9	GPT-4o-mini	73.51	26.49
	DeepSeek-V3	82.35	17.65
	LLaMA3	83.55	16.45
BDI-II	GPT-4o-mini	13.28	86.72
	DeepSeek-V3	12.00	88.00
	LLaMA3	14.29	85.71
GAD-7	GPT-4o-mini	13.60	86.40
	DeepSeek-V3	14.00	86.00
	LLaMA3	16.36	83.64

Note: Percentages use all available classifications per model; inter-model agreement uses only records classified by all three models.

PHQ-9 exhibited the largest variation in valid classifications across models. BDI-II and GAD-7 showed similar distributions, with invalid classifications predominating in both instruments. This pattern is consistent with the controlled generation procedure, which included partially resistant and resistant response profiles containing structural deviations.

These distributions describe model behavior under the experimental benchmark and should not be interpreted as estimates of invalid-response prevalence in real-world psychological assessments.

4.4. Performance Against Ground Truth

Table 4 reports the classification performance of each auditing model against the programmatically derived ground-truth labels. The *valid* classification was treated as the positive class for the calculation of precision, recall, and F1-score.

Because the class distributions were asymmetric, particularly for BDI-II and GAD-7, F1-score was used as the primary summary measure, while accuracy, precision, and recall were reported as complementary indicators of classification behavior.

Table 4. Classification performance of each model against the reference ground-truth labels by instrument

Instrument	Model	Accuracy	Precision	Recall	F1
PHQ-9	GPT-4o-mini	0.912	**1.000**	0.893	0.943
	DeepSeek-V3	**1.000**	**1.000**	**1.000**	**1.000**
	LLaMA3	0.990	0.988	**1.000**	0.994
BDI-II	GPT-4o-mini	0.982	0.867	**1.000**	0.929
	DeepSeek-V3	**0.995**	**0.960**	**1.000**	**0.980**
	LLaMA3	0.973	0.809	**1.000**	0.894
GAD-7	GPT-4o-mini	0.996	**1.000**	0.971	0.986
	DeepSeek-V3	**1.000**	**1.000**	**1.000**	**1.000**
	LLaMA3	0.979	0.870	**1.000**	0.930

Bold: Best value per instrument. The *valid* class was treated as positive. Ground truth was derived from the instrument-specific admissibility criteria.

As shown in Table 4, all models achieved high correspondence with the reference labels. DeepSeek-V3 obtained perfect F1-scores for PHQ-9 and GAD-7 and near-perfect performance for BDI-II.

GPT-4o-mini exhibited a more conservative profile, characterized by high precision and comparatively lower recall. LLaMA3 showed the opposite tendency, achieving perfect recall across instruments but lower precision, particularly for BDI-II. These results indicate that high inter-model agreement coexisted with meaningful differences in classification strictness.

4.5. Deterministic Rule-Based Baseline

To provide a controlled reference for the LLM-based auditing results, the deterministic rule-based validator described in Section 3 was evaluated against the same programmatically derived ground-truth labels.

The baseline reproduced the explicit admissibility rules used to construct the reference labels and therefore achieved perfect classification performance, with accuracy of 1.000 across the evaluated instruments. This result was expected by design, since both the ground truth and baseline were based on the same deterministic constraints.

The baseline therefore provides a reproducible control condition rather than evidence that rule-based validation is universally equivalent or superior to LLM-based auditing. Its limitation becomes relevant when responses exhibit linguistic variation that cannot be captured by strict lexical matching.

5. Discussion

The results indicate that LLMs can operate as instrument-aware guardrails when provided with explicit, instrument-specific validity criteria. Across PHQ-9, GAD-7, and BDI-II, the

models showed high inter-model agreement and strong correspondence with programmatically derived reference labels, while exhibiting distinct levels of strictness in accepting or rejecting responses.

5.1. LLMs as Instrument-Aware Guardrails

The findings support the use of LLMs as a complementary validation layer for structured psychometric questionnaires. Rather than inferring psychological states, the evaluated models were used to determine whether responses complied with the response structure and admissibility criteria of a specific instrument. This distinction is important because the proposed guardrail operates at the data-quality layer, before questionnaire scoring or downstream analysis.

The high agreement observed across different model families suggests that explicit instrument-specific constraints can support consistent validation behavior. The models also processed the controlled non-canonical responses included in the benchmark, suggesting that LLM-based validation warrants further evaluation with naturally occurring linguistic variation. However, this advantage should not be interpreted as evidence of clinical validity or superiority over deterministic methods in unrestricted settings.

5.2. Inter-Model Agreement and Model Behavior

The Fleiss' Kappa values, ranging from 0.795 to 0.933, together with unanimity rates above 90%, indicate substantial to near-perfect agreement among the three auditing models. Nevertheless, the pairwise results reveal meaningful differences in model behavior, particularly for PHQ-9.

DeepSeek-V3 showed the most balanced performance against the reference labels. GPT-4o-mini exhibited a more conservative profile, characterized by higher precision and comparatively lower recall in some settings, whereas LLaMA3 showed higher recall but lower precision, reflecting a more permissive acceptance pattern. These differences indicate that high inter-model agreement does not imply identical decision behavior and that deployment should consider the relative costs of false acceptance and false rejection.

5.3. Deterministic Baseline and Practical Deployment

The deterministic rule-based baseline provides an important control condition. Its perfect accuracy is expected because the same instrument-specific admissibility rules used to construct the programmatic ground truth were encoded directly in the baseline. Therefore, this result should not be interpreted as evidence that deterministic validation universally outperforms LLM-based approaches.

Instead, the results support a complementary architecture in which deterministic rules enforce explicit structural constraints while LLMs provide semantic validation for responses expressed through natural language. Such a configuration could be integrated between a conversational interface and the questionnaire scoring component, allowing structurally inadmissible responses to be identified before they affect downstream results.

The local LLaMA3 deployment also demonstrates the feasibility of performing this validation without relying exclusively on proprietary cloud APIs. Although its performance did not consistently match DeepSeek-V3, its high agreement with the other models suggests potential value in privacy-sensitive environments where local processing is desirable.

5.4. Limitations

The main limitation is the controlled synthetic nature of the dataset. Although this design enables reproducible manipulation of structural violations and deterministic ground-truth assignment, real-world responses may exhibit greater linguistic, cultural, and contextual variability. The initial human validation of 100 instances provided an important reference for the generation protocol, but does not constitute clinical validation of the resulting dataset.

Class imbalance, particularly in BDI-II and GAD-7, may also influence agreement and classification metrics. In addition, only three models were evaluated under a shared prompt and fixed inference configuration. Finally, the use of LLaMA3 in both generation and auditing represents a potential source of model-specific bias, although the generation and auditing procedures were separated and the auditing models did not receive the generation or labeling rules.

Future work should therefore evaluate the framework using independently annotated real-world conversational data, additional instruments and models, multilingual settings, and hybrid validation architectures combining deterministic constraints with semantic LLM-based auditing.

6. Conclusion

This study investigated whether Large Language Models can operate as instrument-aware guardrails for preserving the structural integrity of standardized mental health questionnaires. Using a controlled synthetic dataset comprising PHQ-9, GAD-7, and BDI-II records, three LLMs were evaluated according to their inter-model agreement and correspondence with programmatically defined instrument-specific admissibility criteria.

The results showed substantial to near-perfect inter-model agreement, with Fleiss' Kappa values ranging from 0.795 to 0.933, together with high correspondence with the reference labels. Although the models exhibited different behavioral profiles in terms of conservativeness and permissiveness, the findings support the feasibility of LLMs as a complementary validation layer for conversational systems administering structured psychological questionnaires.

The proposed approach is not intended to replace psychometric instruments, deterministic scoring procedures, or professional psychological assessment. Its role is to identify structurally inadmissible or ambiguous responses before they enter the scoring process. A hybrid architecture combining deterministic rules for explicit constraints with LLM-based semantic auditing may therefore provide a practical balance between transparency, reproducibility, and linguistic flexibility.

References

Brickman, J., Gupta, M., and Oltmanns, J. R. (2025). Large language models for psychological assessment: A comprehensive overview. *Advances in Methods and Practices in Psychological Science*, 8(3).

Dong, Y. et al. (2024). Safeguarding large language models: A survey. *arXiv preprint arXiv:2406.02622*.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preotiuc-Pietro, D., Asch, D. A., and Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Inan, H. et al. (2023). Llama guard: Llm-based input-output safeguard for human-ai conversations. arXiv preprint arXiv:2312.06674.

Kjell, O. N. E., Kjell, K., and Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, 333:115667.

Kroenke, K., Spitzer, R. L., and Williams, J. B. W. (2001). The phq-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

McCoy, T. H., Castro, V. M., and Perlis, R. H. (2025). Estimating depression severity in narrative clinical notes using large language models. *Journal of Affective Disorders*, 381:270–274.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Moëll, B. and Sand Aronsson, F. (2026). High-accuracy prediction of mental health scores from english bert embeddings trained on llm-generated synthetic self-reports. *Frontiers in Digital Health*.

Nair, R. P., Thushara, M. G., and Sugumaran, V. (2025). Fine-tuned llms versus rule-based nlp for uml diagram generation: An educational evaluation. *IEEE Access*, 13:211605–211619.

Nguyen, H. H. et al. (2025). Foundation models as guardrails: Llm- and vlm-based approaches to safety and alignment. In *Proceedings of the APSIPA Annual Summit and Conference*.

Ohse, J. et al. (2024). Large language models for depression screening from transcribed semi-clinical interviews. *Computer Speech & Language*, 88:101663.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The gad-7. *Archives of Internal Medicine*, 166(10):1092–1097.

Teferra, B. G. et al. (2025). Leveraging large language models for automated depression screening. *PLOS Digital Health*, 4(7):e0000943.

Wang, Y.-P. and Gorenstein, C. (2013). Psychometric properties of the beck depression inventory-ii: A comprehensive review. *Revista Brasileira de Psiquiatria*, 35(4):416–431.