Creating Multilingual Mental Health Dialogue Datasets: Limits of Persona-Based Localization via Nationality and Language

Yunkai Xu and Saeed Abdullah
Pennsylvania State University, United States
{yunkai, saeed}@psu.edu

Abstract

AI and large language models (LLMs) have emerged as promising tools to address global mental health challenges. Despite the global nature of these challenges, there remains a critical shortage of high-quality datasets for training and evaluating such systems. To mitigate this gap, researchers increasingly generate synthetic clinical personas to simulate user data and test digital mental health support systems. However, most validated personas rely on English-centric contexts. This paper investigates whether similar persona-based methods can be used to generate multilingual mental health datasets. We modified nationality and language parameters in personas to generate clinical dialogues in Mandarin, Bengali, and Hindi. We then examined how different LLMs perform when evaluating the depression severity of these generated multilingual datasets against the baseline in English. Our findings indicate that just adding nationality and language parameters in personas might not be adequate, as it can introduce clinical inconsistency across languages. LLM judge models often exhibit inaccuracies in assessing depression severity in non-English texts, with performance varying across different models. This exposes the systemic limitations of applying English-centric personas to multilingual contexts. Ultimately, our work highlights the urgent need for culturally responsive data generation to ensure equitable mental health systems globally.

All personas and the full pipeline are publicly available at **GitHub link**.

1 Introduction

Depression presents one of the significant global challenges among mental health disorders (Moitra et al., 2023; Liu et al., 2024). Regions with limited medical infrastructure, such as Bangladesh, experience a severe shortage of mental health professionals (Hasan et al., 2021). Artificial intelligence (AI) and large language models (LLMs) provide a potential medium for depression screening and continuous support (Chen et al., 2024b; Ignashina et al., 2025; Li et al., 2025b; Zhang et al., 2025). However, the benefits of these technological advancements are also not distributed equitably across regions. Prior work has primarily focused on English (Zhang et al., 2025) and emerging evidence suggests that language itself plays a critical role in shaping LLM performance, with more pronounced effects observed in non-English languages (Jin et al., 2023; Raihan et al., 2026). This marginalizes populations that use other languages, creating systemic biases and representation disparities in digital mental health assessments.

Recently, researchers have increasingly utilized synthetic data to mitigate these biases in underrepresented languages. This need is particularly pronounced in the mental health domain, where access to real-world data is further constrained by privacy concerns (Kang et al., 2024; Lorge et al., 2025). Among various approaches to synthetic data generation, persona-based methods have emerged as an effective strategy, particularly for scaling up data generation (Zhang et al., 2018; Ge et al., 2025; Jandaghi et al., 2024). In the mental health field, some personas simulate specific psychological states, such as those defined by the Beck Depression Inventory II (BDI-II) (Weiner and Craighead, 2010), acting as standardized patients for training and evaluating clinical dialogue systems (Wang et al., 2025).

However, the majority of validated clinical synthetic personas originate from English and Western contexts (Wang et al., 2025). It might be feasible to address the current language gap by modifying specific demographic and linguistic variables within the original English persona prompt. However, prior studies demonstrate that altering a single persona parameter affects downstream behavior in other domains (Weeber et al., 2026; Kamruzzaman et al., 2025). It remains unclear whether this

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 138–152

parameter-based localization preserves the original clinical symptoms when generating interactions in non-English contexts.

This study investigates the preservation of clinical features following the parameter-based adaptation of synthetic personas. We designed a controlled experiment using a clinically validated English persona as a baseline (Wang et al., 2025). By modifying only the language and nationality variables, we generated parallel personas for different regions. We then used independent judges implemented with different LLMs to assess the depression severity reflected in the generated chat histories. Our study addresses the following research questions:

- **RQ1:** How does parameter-based localization affect the clinical consistency of symptoms expressed by synthetic personas across different languages?
- **RQ2:** How do diverse LLMs model judges vary in their capability and certainty when assessing depression severity in non-English synthetic conversations generated by the personas?

By addressing these questions, our research provides the following key contributions:

- We provide empirical evidence that modifying persona parameters leads to disparities in depression level representation and weakens alignment with clinical severity levels in dialogue datasets generated in Mandarin, Bengali, and Hindi.
- We highlight the systemic limitations of applying synthetic personas to multilingual artificial intelligence systems, and argue for treating them as clinical artifacts. Specifically, we call for rigorous output-level validation and language-specific evaluation to maintain cross-lingual consistency.

2 Related Works

2.1 Multilingual Large Language Models in Mental Health

Mental health care resources remain unevenly distributed worldwide, with many regions lacking trained clinicians and accessible services (Hasan et al., 2021). Recent work has therefore started to explore how LLMs might support mental health care in non-English contexts (Zhang et al., 2025). Existing efforts span a range of linguistic and cultural settings. In Chinese-speaking contexts, prior studies have developed LLM-based support for cognitive behavioral therapy and deployed chatbot systems for anxiety and depression support (Na, 2024; Chen et al., 2025). Similar efforts are beginning to emerge in South Asia. For instance, one study considered the potential of LLMs for suicide prevention in India (Chakraborty et al., 2025), while another proposed an LLM-supported intervention for postpartum mental health among women in Bangladesh (Ahmed et al., 2025). Taken together, these studies point to growing interest in multilingual mental health support, but they also remain focused on particular languages and application scenarios.

In principle, fair multilingual models should maintain consistent performance across languages. However, some studies report clear performance gaps between high-resource languages such as English and lower-resource languages such as Bengali (Bhowmik et al., 2025). A main reason for this gap is the imbalance in training and fine tuning data. High-quality clinical data are common in English datasets but rare in many other languages. This shortage limits the use of language models for mental health support in different cultural settings.

2.2 Datasets for Depression Detection and Synthetic Data

Existing depression detection datasets are primarily English-centric, sourced from platforms like Reddit and X (Gui et al., 2019; Ríssola et al., 2020; Naseem et al., 2022; Parapar et al., 2026). While resources in languages such as Chinese, Japanese and Portuguese have emerged (dos Santos et al., 2024; Xiao et al., 2026; Agarwal and Dhingra, 2021), they remain limited to single-language settings. As a result, they do not address the broader challenge of capturing linguistic and cultural diversity across multiple languages.

Collecting large-scale clinical data in non-English context is further hindered by privacy regulations and the scarcity of online mental health communities (Kang et al., 2024; Lorge et al., 2025). To bridge this gap, researchers often translate English datasets into other languages (Yang et al., 2019; Myung et al., 2024; Jin et al., 2023). However, human translation is costly and difficult to scale, while machine translation often fails to capture cultural nuances or specific mental health expres-

sions (Bhowmik et al., 2025; Raihan et al., 2026).

To address data scarcity and ethical constraints, recent studies leverage LLMs for synthetic data generation (Wang et al., 2024). In mental health, this includes simulating conversational therapy sessions (Zhezherau and Yanockin, 2024), patient needs (Ronan et al., 2025), and patients' clinical synopses to balance severity distributions (Kang et al., 2024). Others have generated context-enhanced social media data to identify psychosocial risks (Garg et al., 2026). Despite these advances, existing work focuses almost exclusively on single-language generation. However, existing work often focuses on generating data within a single language. It remains unclear how synthetic representations generalize across languages and whether they can preserve clinically meaningful signals in a multilingual framework.

2.3 Personas as a Source of Synthetic Data

Personas represent archetypal people through fictional yet data-informed profiles, enabling designers to reason about the characteristics and goals of a target population (Pruitt and Grudin, 2003). LLMs can generate high-quality personas comparable to human experts (Schuller et al., 2024), making them a practical alternative when traditional data collection is constrained (Salminen et al., 2025). LLMs have further enabled data-driven persona development (Jung et al., 2025), maintaining high fidelity even in specialized domains (Kaur et al., 2025).

In mental health, personas are increasingly used to synthesize therapeutic dialogue data to bypass data scarcity. For instance, Jandaghi et al. (2024) proposed a generator-critic architecture for faithful persona-based interactions, while Wu et al. (2025) leveraged persona traits to modulate emotional support sessions. To ensure clinical grounding, Wang et al. (2025) incorporated specific attributes like BDI-II scores into simulations. To standardize these efforts, the PatientHub framework (Sabour et al., 2026) provides a modular system for reproducible simulated patient deployment.

However, the use of synthetic personas introduces risks associated with data validity and algorithmic bias (Li et al., 2025a; Batzner et al., 2025). Variations in the components of a persona can cause inconsistencies in downstream performance (Wu et al., 2025). Kamruzzaman et al. (2025) identified that assigning nationality-specific personas to LLMs results in emotional stereotypes, as the models disproportionately attribute certain emotions to specific countries. Furthermore, assigned nationality personas can alter how models perceive different nations, which often leads to more favorable treatment of Western regions compared to others (Kamruzzaman and Kim, 2025).

While prior work highlights cultural bias in LLMs in healthcare, the majority of validated clinical personas are still developed within English-speaking contexts. There remains a lack of evidence concerning whether the simple modification of nationality and language parameters within English-centric prompts is sufficient to generate valid clinical dialogues in other languages. Our work addresses this gap by investigating how such adaptations influence the clinical accuracy of depression assessments across different linguistic settings.

3 Method

3.1 Persona Construction

We adapt the workflow illustrated in Figure 1 and the clinically validated baseline personas from prior work by Wang et al. (Wang et al., 2025). The personas consist of several dimensions that capture demographic attributes, depression symptoms, communication patterns, and social context (see Appendix A.2). Table 1 summarizes these dimensions and their sources. These elements serve specific functions in clinical depression diagnosis. For instance, a documented life history is an important component in understanding depression and related mental health outcomes, and analyzing detailed life histories helps identify pathways that link individual experiences with mental health trajectories (Singer et al., 1998). Therefore, we include life history information when constructing depression personas. In total, we use 12 personas (Table 5). All personas include the dimensions described above, and the content in these dimensions differ across personas. We focus on personas' depression characteristics by varying symptoms, severity levels, and BDI-II scores to represent different levels of depression.

We introduce two new variables to these baseline personas: **Nationality** and **Language Use**. Many prior studies show that Nationality can change how a persona behaves in downstream tasks (Kamruzzaman and Kim, 2025; Kamruzzaman et al., 2025). We include the Language Use variable to dictate the output language of the generated text. We represent specific geographic regions by pairing nationality

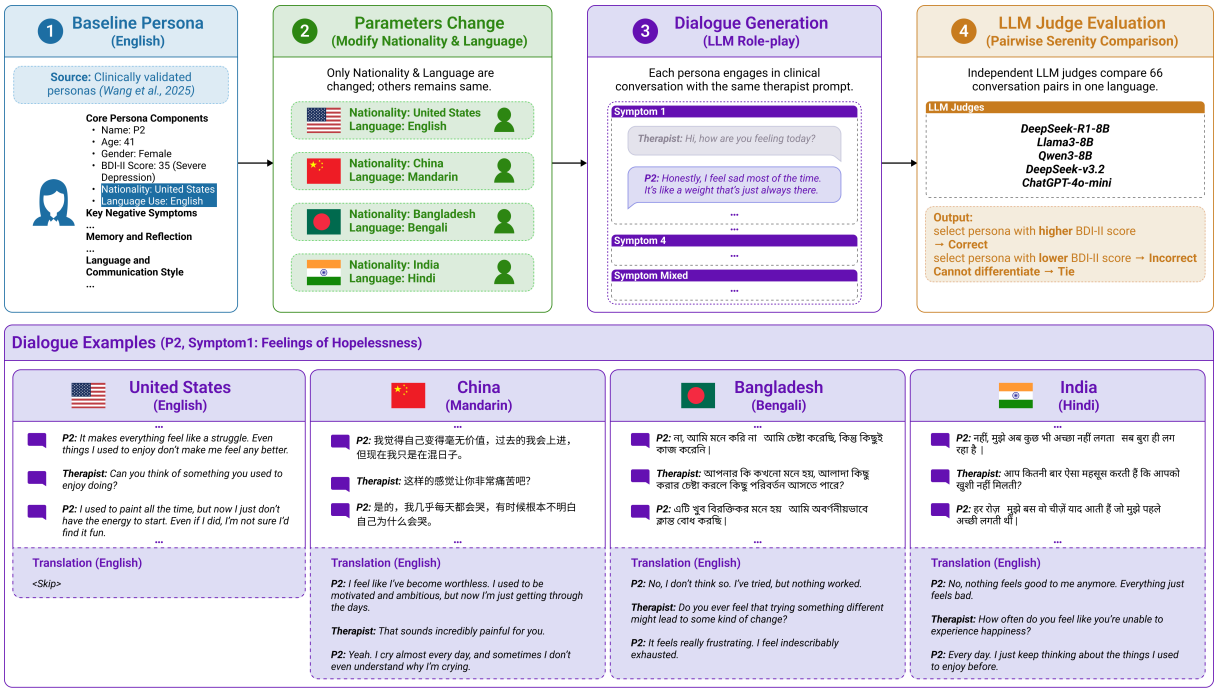

Figure 1: Overview of the multilingual synthetic dialogue generation and evaluation workflow. Example dialogue sessions are translated into English for readability using Google Translate[1].

with language use (United States-English, Chinese-Mandarin, Bangladeshi-Bengali, and Indian-Hindi). We focused on these regions and languages as they are among the most widely spoken languages globally (SIL International, 2023). The remainder of the persona prompt remains in English and keeps the same content and structure as the original clinically validated persona. Using this procedure, we generate distinct personas for different target regions, with 12 personas in each language and 48 personas in total.

This approach follows prior work that adopts controlled modification to enable systematic comparison across persona settings (Sakai et al., 2025).

3.2 Evaluation: Pairwise Severity Comparison

To validate the clinical consistency of our generated personas, we mainly replicate a pairwise depression severity differentiation task (Wang et al., 2025) and extend this to a multilingual context to test our personas. The evaluation process consists of two stages:

1. **Dialogue Generation:** We employ an independent LLM-based therapist agent (powered by ChatGPT-4o-mini) to conduct semi-structured clinical interviews with each persona. The agent using the simple prompt was adapted from (Wang et al., 2025), with additional instructions added to ensure that the generated responses remained consistent with the language settings specified in the persona (see Appendix B). For each persona, this interaction generated five distinct dialogue sessions. Specifically, four sessions were designed to focus on individual target symptoms defined in the persona, while an additional mixed-symptom session combined all four symptoms. The dialogue session consisted of five to seven turns. Unlike Wang et al. (2025)'s work, our personas possess specific nationalities, and the therapist interacts with them in their **corresponding languages**. This results in a multilingual corpus consisting of 60 dialogue sessions per language.

2. **Blind Pairwise Judging:** An independent LLM-based judge is presented with generated dialogues from two personas with different depression severity levels (based on BDI-II scores). The judge is blinded to the underlying persona configurations and scores. Its task is to determine which patient exhibits a higher level of depression severity, or if they are indistinguishable (i.e., Persona A > Persona B, B > A, or "Tie" as a tie case). Consequently,

[1] https://translate.google.com/

Table 1: Dimensions used to construct clinical personas.

Dimension	Sources
Age	(Faravelli et al., 2013)
Gender	(Faravelli et al., 2013; Ensel, 1982)
BDI-II score	(Weiner and Craighead, 2010)
Depression symptoms and severity	(American Psychiatric Association and American Psychiatric Association, 2013)
Communication style	(Smirnova et al., 2018; Al-Mosaiwi and Johnstone, 2018)
Life history	(Singer et al., 1998)
Social context	(Aseltine et al., 1994; Brown, 2002)
Clinician behavior constraints	(Bickley and Szilagyi, 2012)
Social media use	(Huang, 2022; Berryman et al., 2018)

each evaluation cycle comprises 66 pairwise comparisons ($N_{\text{total}} = \binom{12}{2} = 66$).

This setup evaluates whether the clinical cues of depression remain detectable and consistent when expressed through different languages and cultural backgrounds, and whether the LLM judge can maintain diagnostic sensitivity despite the changes in the personas.

3.3 Evaluation Metrics

Replicated from Wang et al. (2025)'s work, We also evaluate model performance using three metrics: overall accuracy, same-level error rate, and tie distance.

Overall Accuracy. This metric evaluates the LLM judge's proficiency in correctly identifying the relative depression severity within each pair. A prediction is counted as correct if the judge successfully selects the dialogues associated with the persona that possesses the higher ground-truth BDI-II score. Cases where the model outputs selects the persona with the lower score are treated as incorrect. Tie responses were treated as a separate outcome category. Thus, the total number of comparisons can be expressed as:

$$N_{\text{total}} = N_{\text{correct}} + N_{\text{incorrect}} + N_{\text{tie}} \quad (1)$$

Overall accuracy is then computed over non-tie comparisons:

$$\text{Overall Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}} - N_{\text{tie}}} \times 100\% \quad (2)$$

Same-Level Error Rate. This metric analyzes the quality of model errors. When a model makes an incorrect prediction, we examine whether the two personas belong to the same BDI-II severity category. We define Same-Level Error Rate as the percentage of incorrect predictions where the two evaluated personas actually belong to the same severity level. This metric captures errors that occur near clinical severity boundaries.

$$\text{Same-Level Error Rate} = \frac{E_{\text{same}}}{E_{\text{total}}} \times 100\% \quad (3)$$

where E_{same} is the number of errors between personas within the same severity category, and E_{total} is the total number of incorrect predictions. Errors that cross severity boundaries indicate larger clinical misjudgments.

Tie Distance. Models occasionally refuse to select one chat history and instead output a tie. We measure the Average Tie Distance to quantify the severity gap between the personas in these tie cases. The distance is the absolute difference between the underlying BDI-II scores of the two personas. A larger tie distance indicates that the model expresses uncertainty even when the actual severity difference is large.

3.4 Model Use

We use several LLMs as judges, including proprietary models (GPT-4o-mini (OpenAI et al., 2024), DeepSeek-V3.2 (DeepSeek-AI et al., 2025b)) and open-weight models (LLaMA3.1-8B (Grattafiori et al., 2024), Qwen3-8B (Yang et al., 2025), and DeepSeek-R1-8B (DeepSeek-AI et al., 2025a)). We use the 8B versions to align with prior work (Wang et al., 2025) and maintain efficiency.

In the evaluation task, each model compares the dialogues from two personas to decide which one shows more severe depression. The models do not

Table 2: Overall accuracy across models and languages.

Model	Bengali	English	Hindi	Mandarin
ChatGPT-4o-mini	92.42	95.45	87.88	86.36
DeepSeek-v3.2	90.91	93.75	83.33	84.85
DeepSeek-R1-8B	63.33	98.21	67.24	84.62
Llama3-8B	45.45	84.13	65.38	75.76
Qwen3-8B	71.21	89.39	73.44	80.30

see the persona prompts, so the assessment is based only on the generated text.

4 Findings

4.1 Cross-Model and Cross-Lingual Results

4.1.1 Overall Accuracy and Cross-Lingual Disparity

We present the overall accuracy of the evaluated models across four languages in Table 2. The results reveal a persistent English advantage across all language models. Performance in English consistently serves as the upper bound for each model.

ChatGPT-4o-mini and DeepSeek-v3.2 achieve the highest overall accuracy. ChatGPT-4o-mini maintains the lowest cross-lingual variability with a standard deviation of 4.17. DeepSeek-v3.2 is slightly higher than ChatGPT-4o-mini. Both of them demonstrate relatively robust calibration in interpreting depression severity across different linguistic contexts.

Smaller open-weight models demonstrate substantial performance drops when processing non-English personas. DeepSeek-R1-8B and Llama3-8B exhibit high performance fluctuations across languages, with standard deviations of 16.15 and 16.69 respectively. DeepSeek-R1-8B achieves 98.21% accuracy in English but drops to 63.33% in Bengali and 67.24% in Hindi. This indicates that mental disorder severity interpretation capabilities in smaller models remain heavily biased toward English contexts.

4.1.2 Same-Level Error Rate and Cross-Lingual Disparity

Cross-language differences manifest in overall accuracy and in the nature of the errors. We analyze whether incorrect model answers stay within the same depression severity level or cross-severity boundaries. Table 3 details these error distributions.

Models exhibit severe cross-lingual inconsistency in their error calibration. English consistently corresponds to fewer cross-severity mistakes for most models. For example, DeepSeek-R1-8B maintains a 100% Same-Level Error Rate in English. This means all its English mistakes occur between personas of identical severity tiers. This is important because many studies are concerned with severity categories rather than exact scores (Naseem et al., 2022; Peng et al., 2026), making within-tier errors less problematic than cross-tier ones. However, its Same-Level Error Rate drops to 22.73% in Bengali. This indicates that Bengali errors frequently cross-severity boundaries.

Table 3: Same-level error rates (%) across languages and models. Higher values indicate fewer cross-severity errors. Bold values indicate the highest score within each language.

Model	Bengali	Hindi	Mandarin	English
ChatGPT-4o-mini	80.00	**87.50**	66.67	66.67
DeepSeek-v3.2	**100.00**	72.73	70.00	**100.00**
DeepSeek-R1-8B	22.73	31.58	**80.00**	**100.00**
Llama3-8B	23.33	33.33	50.00	60.00
Qwen3-8B	31.58	29.41	61.54	85.71

Llama3-8B and Qwen3-8B show similar patterns. Llama3-8B confines 60.00% of its English errors to the same severity level, compared to 23.33% in Bengali and 33.33% in Hindi. DeepSeek-v3.2 is an exception. It maintains consistently high Same-Level Error Rates across all languages, indicating a robust calibration in distinguishing adjacent severity levels regardless of the input language. These findings demonstrate that language changes affect both the performance level and the structural consistency of clinical severity judgments.

4.1.3 Uncertainty and Tie Behavior

We further analyze model uncertainty by examining explicit tie outputs. Table 4 summarizes the frequency and severity distance of tie cases.

Tie behavior varies unevenly across models and languages. DeepSeek-R1-8B produces the highest number of ties across all four languages, totaling 25 cases. Llama3-8B produces 14 ties, but these are heavily concentrated in Bengali and Hindi. ChatGPT-4o-mini never outputs a tie, and DeepSeek-v3.2 rarely produces ties.

The context of these ties also differs significantly between models. DeepSeek-v3.2 expresses uncertainty exclusively in near boundary cases, with a mean tie distance of 4.00. DeepSeek-R1-8B and Llama3-8B frequently express uncertainty even when severity differences are large. Their mean tie distances are 10.80 and 14.14 respectively.

Table 4: Overall tie counts and average tie distance by model.

Model	Tie Count	Avg Distance	Max Distance
DeepSeek-R1-8B	25	10.80	34
Llama3-8B	14	14.14	28
Qwen3-8B	2	5.50	6
DeepSeek-v3.2	2	4.00	5
ChatGPT-4o-mini	0	-	-

Furthermore, the difficulty threshold that triggers a tie depends on the language. For DeepSeek-R1-8B, the average tie distance is 15.12 in Hindi compared to 7.33 in Bengali and 8.90 in English. This indicates that the severity gap required to confuse the model is not constant and shifts substantially according to the linguistic context.

4.2 Persona-based Methods for Multilingual Dataset

The high accuracy and low cross boundary error rates observed in English evaluations suggest that the baseline English personas successfully generate distinct and recognizable clinical traits, which is consistent with prior studies (Wang et al., 2025). However, the decline in judge accuracy for Bengali and Hindi indicates a degradation in the underlying text. The increased frequency of cross-severity errors and large distance tie cases in non-English languages shows that the generated dialogues may lack clear clinical signals in the original personas.

Overall, these results collectively demonstrate that modifying the nationality and language labels within an English-centric persona is insufficient to create viable multilingual clinical personas. The shallow localization process introduces clinical ambiguity and this problem is much more visible in smaller 8B models, causing the resulting text to lose alignment with the intended BDI-II severity scales.

5 Discussion

5.1 Why Minimal Localization in Personas Fails to Preserve Clinical Cues

This study examined a strategy for multilingual persona construction: starting from an expert-validated English persona prompt and replacing only nationality and language to generate non-English personas. We then evaluated the resulting dialogues through blinded severity judgment. Across LLM judge models, English consistently acted as the upper bound, while non-English dialogues more often led to severity misclassification and ties. This pattern suggests that simple parameter substitution does not reliably preserve clinically relevant cues across languages.

One potential reason is that changing nationality and language also alters the cultural context embedded in the persona, while symptom expression may vary substantially across cultures and therefore may not be preserved consistently (Goodmann et al., 2021; Jovanović et al., 2026; Bradshaw et al., 2026). The same BDI-II level can surface in different ways depending on the conversation, as the expression of and response to depression vary significantly across cultures (Teja et al., 1971). This issue has been observed not only in real clinical interactions but also in synthetic sessions generated from personas (Sakai et al., 2025). When these expressive patterns shift, the resulting dialogue may still describe similar experiences but fail to present them in a form that judges can consistently interpret. Prior work reports similar cross-lingual gaps in health-related tasks and shows that models often perform better when queries are expressed in English (Jin et al., 2023). Therefore, our findings extend this observation to persona-driven dialogue generation.

In addition, both nationality and language do not act as neutral parameters for the LLM dialogue generation and the LLM judges. Evidence from cultural benchmarks shows that model performance can vary across regions and languages on everyday knowledge tasks, and that performance can remain higher in English than in the local language for some low-resource cultures (Myung et al., 2024). Furthermore, assigning nationality has been shown to change model outputs in systematic ways, including how emotions and social attributes are expressed (Kamruzzaman and Kim, 2025; Kamruzzaman et al., 2025). In our setting, this means that the generated dialogue may introduce additional culture-specific interaction patterns that were not present in the original English persona. As a result, symptom descriptions can shift in structure and emphasis, which reduces the clarity of severity signals even when the underlying prompt remains unchanged.

5.2 Implications for Multilingual Mental Health Persona Construction and Evaluation

The findings have direct implications for the use of personas in multilingual mental health data generation and evaluation. First, multilingual dialogues

produced through simple parameter replacement should not be treated as equivalent samples across languages. In our study, non-English dialogues show not only lower accuracy but also more cross-severity errors and unstable tie behavior. This indicates a shift in the underlying data distribution. Similar effects have been reported in multilingual mental health benchmarks, where model performance drops when tasks rely on translated rather than native data, with variation across languages and translation quality (Raihan et al., 2026).

Second, persona construction requires stronger methodological control in clinical settings. Prior work has identified common issues in persona-based research, including weak specification of target populations and limited reporting of construction procedures (Batzner et al., 2025). Our results show that these issues extend to multilingual settings. When personas are adapted through minimal parameter changes, the resulting dialogues may no longer preserve the intended clinical condition. A multilingual clinical persona should therefore be treated as a new artifact that requires validation at the output level, rather than as a direct extension of an English template.

Third, the observed cross-language gap may arise from both generation and evaluation. Prior work shows that LLM-based evaluation can align with human judgments in English tasks, while also exhibiting bias and sensitivity to LLM judgeś prompt design (Liu et al., 2023). However, this alignment does not guarantee consistency across languages. Recent studies report that multilingual LLM-as-a-judge setups can produce inconsistent results on parallel data, especially in lower-resource languages (Fu and Liu, 2025). Both human and LLM judges have also been shown to be affected by bias and input variation (Chen et al., 2024a; Ye et al., 2024). In our setting, weaker clinical cues in non-English dialogue and lower LLM judge consistency may interact, producing larger observed differences across languages.

6 Conclusion

This paper examined whether modifying nationality and language variables in an English clinical persona prompt can preserve depression severity signals across languages, using controlled persona construction, multilingual dialogue generation, and blinded pairwise evaluation. The results show that English remains the most stable setting, while Bengali and Hindi exhibit lower accuracy, more cross-severity errors, and higher uncertainty in tie cases, which indicates that the generated dialogues do not consistently reflect the intended BDI-II levels when only minimal parameters are changed. The comparison across models further shows that performance differences are not uniform, and that some models maintain calibration in English but fail to do so in other languages, suggesting that both generation and evaluation processes contribute to the observed gaps. These findings demonstrate that parameter-based localization does not maintain clinical consistency and that multilingual personas produced in this way should not be treated as equivalent to their English counterparts. Future work should treat multilingual persona construction as a separate design and validation process that incorporates culturally grounded expression, evaluates symptom representation beyond severity labels, and uses multiple evaluation strategies, including human review to ensure that generated data can support mental health applications across languages.

7 Ethical Considerations

This work relies on synthetic personas and dialogues rather than identifiable data. The personas sourced from Wang et al. (2025) are also anonymous. Although personas are grounded in clinically informed attributes such as BDI-II severity levels, the generated content should not be interpreted as clinical evidence or used for diagnosis. Given the sensitive nature of mental health, we acknowledge that synthetic personas may oversimplify or misrepresent how depression is expressed across languages and cultures (Teja et al., 1971; Sakai et al., 2025). Our findings further show that minimal parameter-based localization can distort clinical signals, which raises risks if such data are treated as equivalent across languages or used without validation. In addition, observed cross-lingual performance disparities reflect broader fairness concerns in multilingual NLP, where low-resource languages may be disproportionately affected. We therefore caution against deploying such synthetic data in real-world mental health applications and advocate for culturally grounded data construction and validation practices.

8 Limitations

First, our conclusions are based on a specific localization strategy, namely replacing nationality

and language in an English persona prompt while keeping the rest of the structure fixed. This design allowed us to isolate the effect of minimal parameter-based localization, but it simplifies cultural and language use variation by encoding it through a limited set of persona parameters, which may overlook within-country and within-region variation. In particular, language and cultural background are coupled in our current design through nationality-language pairings. Therefore, we cannot fully determine whether the observed cross-lingual differences arise from linguistic structure, culturally specific symptom expression, the behavior of the LLMs used for generation and judgment, or the interaction among these factors, as discussed in Section 5.1. Future work should separate these dimensions more explicitly, for example by comparing multiple languages within the same cultural context, multiple cultural contexts within the same language, and bilingual settings where patients may move between languages depending on emotional or clinical context (Williams et al., 2020; Elwahsh et al., 2025).

We also do not claim that all forms of multilingual persona construction will lead to the same results. Richer localization methods that explicitly model cultural context, symptom narration, and discourse style may perform differently. Future work should compare minimal localization with stronger adaptation strategies that preserve symptom strength while allowing culturally appropriate expression when generating and using synthetic multilingual personas.

A second limitation concerns the evaluation pipeline. We relied on LLM-based pairwise severity judgments and did not include independent validation from human clinicians, or native-speaking annotators. As a result, we could not directly verify whether the generated multilingual dialogues were comparable in clinical realism, symptom coverage, discourse naturalness, or linguistic quality across languages. Our evaluation measures severity recoverability, but this is only one aspect of validity for multilingual clinical personas. A dialogue may preserve the intended BDI-II severity ordering while still failing to reflect realistic narrative structure, culturally grounded symptom framing, or safe and appropriate clinical communication. Future work should evaluate multilingual personas across multiple dimensions, including symptom fidelity, cultural appropriateness, linguistic naturalness, discourse realism, and safety.

Our current framework cannot fully disentangle dialogue generation quality from LLM judge reliability. The observed cross-lingual performance differences may reflect weakened clinical cues in the generated dialogues, inconsistent multilingual reasoning by the judge models, or both. Although the judge models were blinded to the persona prompts and evaluated only the generated conversations, their judgments may still be affected by language-specific limitations and model-specific uncertainty. Future work should reduce this uncertainty by combining multiple evaluation strategies, including human review, multilingual expert annotation, agreement analysis across LLM judges, and direct quality assessment of the generated dialogues before downstream severity comparison.

There are also constraints in the dialogue generation setup itself (Section 3.2). We largely followed the prompting strategy used in prior work and did not independently evaluate whether the therapist agent guided each patient model with equal depth across languages or maintained balanced coverage of BDI-II symptoms. The generated conversations are limited in length, which may have constrained the range and depth of depressive symptoms expressed in the dialogues. Longer and more adaptive interviews may produce richer clinical signals, but they may also introduce additional variability across languages. Future work should examine how interview length, therapist prompting strategy, and symptom-specific questioning affect the stability of multilingual persona-based dialogue generation.

Finally, we did not conduct statistical significance testing across language conditions in the current analysis. Our pairwise evaluation outcomes include correct, incorrect, and tie cases, and tie behavior varied substantially across models and languages. Rather than imposing a single treatment of ties, we report tie frequency and tie distance as descriptive indicators of model uncertainty. Future work with larger samples should pre-register how tie cases are handled and apply statistical models that can account for repeated comparisons across personas, languages, and judge models.

References

Kaustubh Agarwal and Bhavya Dhingra. 2021. Deep Learning Based Approach For Detecting Suicidal Ideation in Hindi-English Code-Mixed Text: Baseline and Corpus. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 100–105, National Institute of Tech-

nology Silchar, Silchar, India. NLP Association of India (NLPAI).

Istiaq Ahmed, Syed Niaz Mohtasim, Faiza Omar Arpita, Ashraful Islam, and M. Ashraful Amin. 2025. A Conceptual Design Framework of GorbhoShongi App for Mental Well-Being Among Bangladeshi Pregnant and Postpartum Women. In *HCI International 2024 – Late Breaking Posters*, pages 3–13, Cham. Springer Nature Switzerland.

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542.

American Psychiatric Association and American Psychiatric Association, editors. 2013. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, 5th ed edition. American Psychiatric Association, Washington, D.C.

Robert H. Aseltine, Susan Gore, and Mary Ellen Colten. 1994. Depression and the social developmental context of adolescence. *Journal of Personality and Social Psychology*, 67(2):252–263.

Jan Batzner, Volker Stocker, Bingjun Tang, Anusha Natarajan, Qinhao Chen, Stefan Schmid, and Gjergji Kasneci. 2025. Whose Personae? Synthetic Persona Experiments in LLM Research and Pathways to Transparency. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):343–354.

Chloe Berryman, Christopher J. Ferguson, and Charles Negy. 2018. Social media use and mental health among young adults. *Psychiatric Quarterly*, 89(2):307–314.

Shimanto Bhowmik, Tawsif Tashwar Dipto, Md Sazzad Islam, Sheryl Hsu, and Tahsin Reasat. 2025. Evaluating LLMs' Multilingual Capabilities for Bengali: Benchmark Creation and Performance Analysis. *Preprint*, arXiv:2507.23248.

Lynn Bickley and Peter G. Szilagyi. 2012. *Bates' Guide to Physical Examination and History-Taking*. Lippincott Williams & Wilkins. Google-Books-ID: g0Ao61hGAloC.

Matt Bradshaw, Koichiro Shiba, Sung Joon Jang, Blake Victor Kent, Rebecca Bonhag, Byron R Johnson, and Tyler J VanderWeele. 2026. Demographic variation in symptoms of depression and anxiety across 22 global flourishing study countries. *Communications Medicine*.

George W. Brown. 2002. Social roles, context and evolution in the origins of depression. *Journal of Health and Social Behavior*, 43(3):255–276.

Tanmoy Chakraborty, Koushik Sinha Deb, Himanshu Kulkarni, Sarah Masud, Suresh Bada Math, Gayatri Oke, Rajesh Sagar, and Mona Sharma. 2025. The promise of generative AI for suicide prevention in India. *Nature Machine Intelligence*, 7(2):162–163.

Chen Chen, Kok Tai Lam, Ka Man Yip, Hung Kwan So, Terry Yat Sang Lum, Ian Chi Kei Wong, Jason C. Yam, Celine Sze Ling Chui, and Patrick Ip. 2025. Comparison of an AI Chatbot With a Nurse Hotline in Reducing Anxiety and Depression Levels in the General Population: Pilot Randomized Controlled Trial. *JMIR Human Factors*, 12(1):e65785.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or LLMs as the Judge? A Study on Judgement Bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. 2024b. Depression Detection in Clinical Interviews with LLM-Empowered Structural Element Graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8181–8194, Mexico City, Mexico. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, and 1 others. 2025a. DeepSeek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645(8081):633–638.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, and 1 others. 2025b. DeepSeek-v3.2: Pushing the frontier of open large language models. *Preprint*, arxiv:2512.02556 [cs].

Wesley Ramos dos Santos, Rafael Lage de Oliveira, and Ivandré Paraboni. 2024. SetembroBR: A social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, 58(1):273–300.

Sarah Elwahsh, Nora Stern, Aneesha Singh, and Amid Ayobi. 2025. Linguistic Diversity and Mental Well-Being: Co-Designing Custom AI Chatbots with Multilingual Mothers. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, ACM Conferences, pages 1–17.

Walter M. Ensel. 1982. The role of age in the relationship of gender and marital status to depression. *The Journal of Nervous and Mental Disease*, 170(9):536.

Carlo Faravelli, Maria Alessandra Scarpato, Giovanni Castellini, and Carolina Lo Sauro. 2013. Gender differences in depression and anxiety: The role of age. *Psychiatry Research*, 210(3):1301–1303.

Xiyan Fu and Wei Liu. 2025. How Reliable is Multilingual LLM-as-a-Judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

Muskan Garg, Xingyi Liu, Eunji Jeon, Joanna M. Biernacka, Mark A. Frye, Yonas E. Geda, and Sunghwan Sohn. 2026. Leveraging reddit data for context-enhanced synthetic health data generation to identify low self esteem. *Frontiers in Psychiatry*, 16.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling Synthetic Data Creation with 1,000,000,000 Personas. *Preprint*, arXiv:2406.20094.

Danielle R. Goodmann, Sariah Daouk, Megan Sullivan, Juan Cabrera, Nancy H. Liu, Suzanne Barakat, Ricardo F. Muñoz, and Yan Leykin. 2021. Factor analysis of depression symptoms across five broad cultural groups. *Journal of Affective Disorders*, 282:227–235.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arxiv:2407.21783 [cs].

Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative Multimodal Approach to Depression Detection in Twitter. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):110–117.

M. Tasdik Hasan, Tasnim Anwar, Enryka Christopher, Sahadat Hossain, Md Mahbub Hossain, Kamrun Nahar Koly, K. M. Saif-Ur-Rahman, Helal Uddin Ahmed, Nazish Arman, and Saima Wazed Hossain. 2021. The current state of mental healthcare in bangladesh: part 1 – an updated country profile. *BJPsych International*, 18(4):78–82.

Chiungjung Huang. 2022. A meta-analysis of the problematic social media use and mental health. *International Journal of Social Psychiatry*, 68(1):12–33.

Mariia Ignashina, Paulina Bondaronek, Dan Santel, John Pestian, and Julia Ive. 2025. LLM Assistance for Pediatric Depression. *Preprint*, arXiv:2501.17510.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful Persona-based Conversational Dataset Generation with Large Language Models. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 114–139, Bangkok, Thailand. Association for Computational Linguistics.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. *Preprint*, arXiv:2310.13132.

Veljko Jovanović, Sabirah Adams, Rebeca Aritio-Solana, Christ Billy Aryanto, Andreja Avsec, Ali Bakhshi, Martina Baldassarre, Michael Bender, Sophie Berjot, Sonia Betancourth Zambrano, Andreja Brajša-Žganec, Yunier Broche-Pérez, Carmen Buzea, Rosario Cabello, Rosalinda Cassibba, Judith Cavazos-Arroyo, Fatemeh Daemi, Diego D. Díaz-Guerra, Marija Džida, and 56 others. 2026. Depression and anxiety symptoms in adolescents across 30 countries: Cross-national measurement invariance and relationships with subjective well-being. *Journal of Affective Disorders*, 406:121693.

Soon-Gyo Jung, Joni Salminen, Kholoud Khalil Aldous, and Bernard J. Jansen. 2025. Personacraft: Leveraging language models for data-driven persona developmentpersonacraft. *International Journal of Human-Computer Studies*, 197:103445.

Mahammed Kamruzzaman, Abdullah Al Monsur, Gene Louis Kim, and Anshuman Chhabra. 2025. From Anger to Joy: How Nationality Personas Shape Emotion Attribution in Large Language Models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 48–68, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Mahammed Kamruzzaman and Gene Louis Kim. 2025. Exploring Changes in Nation Perception with Nationality-Assigned Personas in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3678, Suzhou, China. Association for Computational Linguistics.

Andrea Kang, Jun Yu Chen, Zoe Lee-Youngzie, and Shuhao Fu. 2024. Synthetic Data Generation with LLM for Improved Depression Prediction. *Preprint*, arXiv:2411.17672.

Arshnoor Kaur, Amanda Aird, Harris Borman, Andrea Nicastro, Anna Leontjeva, Luiz Pizzato, and Dan Jermyn. 2025. Synthetic Voices: Evaluating the Fidelity of LLM-Generated Personas in Representing People's Financial Wellbeing. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 185–193, New York City USA. ACM.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025a. LLM Generated Persona is a Promise with a Catch. *Preprint*, arXiv:2503.16527.

Yi Li, Xuanxuan Ding, Yifan Chen, Yeye Li, and Nan Ma. 2025b. Customizable AI for Depression Care: Improving the User Experience of Large Language Model-Driven Chatbots. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, pages 1844–1866, New York, NY, USA. Association for Computing Machinery.

Junjiao Liu, Yueyang Liu, Wenjun Ma, Yan Tong, and Jianzhong Zheng. 2024. Temporal and spatial trend analysis of all-cause depression burden based on global burden of disease (GBD) 2019 study. *Scientific Reports*, 14(1):12346.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Isabelle Lorge, Dan W. Joyce, Niall Taylor, Alejo Nevado-Holgado, Andrea Cipriani, and Andrey Kormilitzin. 2025. Detecting the clinical features of difficult-to-treat depression using synthetic data from large language models. *Computers in Biology and Medicine*, 194:110246.

Modhurima Moitra, Shanise Owens, Maji Hailemariam, Katherine S. Wilson, Augustina Mensa-Kwao, Gloria Gonese, Christine K. Kamamia, Belinda White, Dorraine M. Young, and Pamela Y. Collins. 2023. Global mental health: Where we are and where we are going. *Current Psychiatry Reports*, 25(7):301–311.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki A. Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew A. Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen H. Muhammad, Kiwoong Park, Anar S. Rzayev, Nina White, Seid M. Yimam, Mohammad T. Pilehvar, and 3 others. 2024. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Hongbin Na. 2024. CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940, Torino, Italia. ELRA and ICCL.

Usman Naseem, Adam G. Dunn, Jinman Kim, and Matloob Khushi. 2022. Early Identification of Depression Severity Levels on Reddit Using Ordinal Classification. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 2563–2572, New York, NY, USA. Association for Computing Machinery.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and 1 others. 2024. GPT-4 technical report. *Preprint*, arxiv:2303.08774 [cs].

Javier Parapar, Anxo Perez, Xi Wang, and Fabio Crestani. 2026. Overview of eRisk 2025: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 242–265, Cham. Springer Nature Switzerland.

Shixin Peng, Kun Jiang, Yu Yang, Jingying Chen, and Guandong Xu. 2026. Framework for Diverse Depression Patients Roleplaying and Cognitive Diagnosis of Scales Using LLMs Based on CoT Prompts. In *Behavioural and Social Computing*, pages 399–414, Singapore. Springer Nature.

John Pruitt and Jonathan Grudin. 2003. Personas: Practice and theory. In *Proceedings of the 2003 Conference on Designing for User Experiences*, DUX '03, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Ana-Maria Bucur, Stevie Chancellor, and Marcos Zampieri. 2026. Large Language Models for Mental Health: A Multilingual Evaluation. *Preprint*, arXiv:2602.02440.

Esteban A. Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2020. A Dataset for Research on Depression in Social Media. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, pages 338–342, New York, NY, USA. Association for Computing Machinery.

Isabel Ronan, Patrice Crowley, Eva Rombouts, Nicola Cornally, Mohamad M. Saab, David Murphy, and Sabin Tabirca. 2025. A LangChain-based pipeline for one-shot synthetic text generation using generative pre-trained transformers in palliative care research. *Journal of Biomedical Informatics*, 171:104936.

Sahand Sabour, TszYam NG, and Minlie Huang. 2026. PatientHub: A Unified Framework for Patient Simulation. *Preprint*, arXiv:2602.11684.

Shintaro Sakai, Jisun An, Migyeong Kang, and Haewoon Kwak. 2025. Somatic in the East, Psychological in the West?: Investigating Clinically-Grounded Cross-Cultural Depression Symptom Expression in LLMs. *Preprint*, arXiv:2508.03247.

Joni Salminen, Danial Amin, Soon-Gyo Jung, and Bernard Jansen. 2025. The Use of Large Language Models in HCI: A Critical Analysis of Synthetic Users. In *Proceedings of the Augmented Humans International Conference 2025*, AHs '25, pages 413–417, New York, NY, USA. Association for Computing Machinery.

Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. 2024. Generating personas using LLMs and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, pages 1–7, New York, NY, USA. Association for Computing Machinery.

SIL International. 2023. Ethnologue: Languages of the world. 26th edition.

Burton Singer, Carol D. Ryff, Deborah Carr, and William J. Magee. 1998. Linking life histories and mental health: A person-centered strategy. *Sociological Methodology*, 28(1):1–51.

Daria Smirnova, Paul Cumming, Elena Sloeva, Natalia Kuvshinova, Dmitry Romanov, and Gennadii Nosachev. 2018. Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in Psychiatry*, 9.

J. S. Teja, R. L. Narang, and A. K. Aggarwal. 1971. Depression Across Cultures. *The British Journal of Psychiatry*, 119(550):253–260.

Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and Yunhong Wang. 2024. A Survey on Data Synthesis and Augmentation for Large Language Models. *Preprint*, arXiv:2410.12896.

Xi Wang, Anxo Perez, Javier Parapar, and Fabio Crestani. 2025. TalkDep: Clinically grounded LLM personas for conversation-centric depression screening. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pages 6554–6558. Association for Computing Machinery.

Franziska Weeber, Vera Neplenbroek, Jan Batzner, and Sebastian Padó. 2026. One persona, many cues, different results: How sociodemographic cues impact LLM personalization. *Preprint*, arxiv:2601.18572 [cs].

Irving B. Weiner and W. Edward Craighead. 2010. *The Corsini Encyclopedia of Psychology, Volume 1*. John Wiley & Sons. Google-Books-ID: be0VYic5iWwC.

Aya Williams, Mahesh Srinivasan, Chang Liu, Pearl Lee, and Qing Zhou. 2020. Why do bilinguals code-switch when emotional? Insights from immigrant parent–child interactions. *Emotion*, 20(5):830–841.

Shenghan Wu, Yimo Zhu, Wynne Hsu, Mong-Li Lee, and Yang Deng. 2025. From Personas to Talks: Revisiting the Impact of Personas on LLM-Synthesized Emotional Support Conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5439–5453, Suzhou, China. Association for Computational Linguistics.

Yunze Xiao, Tingyu He, Lionel Z. Wang, Yiming Ma, Xingyu Song, Xiaohang Xu, Mona Diab, Irene Li, and Ka Chung Ng. 2026. JiraiBench: A Bilingual Benchmark for Evaluating Large Language Models' Detection of Human Self-Destructive Behavior Content in Jirai Community. *Preprint*, arXiv:2503.21679.

An Yang, Anfeng Li, Baosong Yang, and 1 others. 2025. Qwen3 technical report. *Preprint*, arxiv:2505.09388 [cs].

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2024. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. *Preprint*, arXiv:2410.02736.

Qiyang Zhang, Renwen Zhang, Yiying Xiong, Yuan Sui, Chang Tong, and Fu-Hung Lin. 2025. Generative AI Mental Health Chatbots as Therapeutic Tools: Systematic Review and Meta-Analysis of Their Role in Reducing Mental Health Issues. *Journal of Medical Internet Research*, 27(1):e78238.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Alexey Zhezherau and Alexei Yanockin. 2024. Hybrid Training Approaches for LLMs: Leveraging Real and Synthetic Data to Enhance Model Performance in Domain-Specific Applications. *Preprint*, arXiv:2410.09168.

A Persona

A.1 Persona Details

Table 5: Overview of the 12 clinical personas used in this study. Each persona includes a BDI-II score and key symptoms with severity levels.

No.	BDI-II Score	Key Symptoms (Severity)
P1	15 (Mild)	Difficulty concentrating (1), Irritability (1), Sleep disturbance (2), Appetite change (1)
P2	35 (Severe)	Hopelessness (3), Crying (2), Extreme fatigue (3), Isolation (3)
P3	12 (Mild)	Anhedonia (2), Social withdrawal (2), Sleep change (1), Indecisiveness (1)
P4	13 (Mild)	Irritability (2), Reduced accomplishment (2), Reduced appetite (2), Social insecurity (2)
P5	22 (Moderate)	Loss of energy (3), Worthlessness (2), Social withdrawal (2), Decision difficulty (2)
P6	23 (Moderate)	Sadness (2), Worthlessness (3), Fatigue (3), Concentration difficulty (2)
P7	28 (Moderate)	Guilt (3), Hopelessness (2), Indecisiveness (2), Fatigue (3)
P8	38 (Severe)	Past failure rumination (3), Agitation (3), Loss of interest (3), Concentration difficulty (3)
P9	40 (Severe)	Sadness (2), Tiredness (1), Self-criticism (3), Loss of interest (2)
P10	6 (Minimal)	Worry (1), Restlessness (1), Self-criticism (1), Fatigue (1)
P11	5 (Minimal)	Self-doubt (1), Low motivation (1), Irritability (1), Sleep change (1)
P12	7 (Minimal)	Anxiety (2), Sleep difficulty (1), Self-criticism (1), Decreased enthusiasm (1)

A.2 Example Persona

The following example shows one clinical persona used in our study. The same structure was used across language and nationality conditions, with only the nationality and language use parameters modified.

P2, a 41-Year-Old Individual

Core Persona Components
Name: P2
Age: 41
Gender: Female
BDI-II Score: 35 (Severe Depression)
Nationality: United States
Language Use: English
Key Negative Symptoms
Feelings of Hopelessness: "I feel like no matter what I do, nothing is going to get better." (Severity: 3)
Crying: "I cry almost every day." (Severity: 2)
Extreme Fatigue: "I'm so tired all the time, even when I haven't done anything." (Severity: 3)
Feeling of Isolation: "I don't think anyone truly cares about me or what I'm going through." (Severity: 3)

Memory and Reflection
Life History:
"P2 has been living alone since her divorce three years ago, struggling to rebuild her sense of self."
"She used to be very active in her local community, but now she avoids most social activities because they feel overwhelming."
"P2 has started to think about her life in terms of what she's lost rather than what she still has."
Social Context:
"P2 rarely posts on social media and primarily uses it to scroll through others' lives, which often leaves her feeling worse."
"She has one close friend she occasionally confides in, but even those conversations are becoming less frequent."
Past Interactions:
"P2 has mentioned feeling like she's 'just existing' rather than living, often in a tone of quiet despair."

Language and Communication Style
Linguistic Patterns:
Often uses passive, self-defeating language like "It doesn't matter" or "What's the point?"
Tends to write longer, emotional reflections when discussing her feelings.
Emotional Tone:
Consistently negative, with undertones of hopelessness and isolation.
Typical Topics:
"Reflects on her struggles with loneliness and her lack of energy to engage with others."
"Mentions her inability to see a future for herself or make plans."

Behavioral Constraints
Rarely seeks advice or reassurance, believing it won't make a difference.
Avoids discussing her past relationship directly unless prompted.

Response Goals
Express deep feelings of hopelessness and isolation but avoid directly asking for help.
Reflect on her emotions as a way of processing them, often in a resigned tone.

Environment and Context
Social Media Activity:
Example Post: "Another day of just getting through it. I don't even know why I bother anymore."
Typical Interactions: Occasionally reacts to posts about mental health or personal growth but rarely comments.
Current Context of Interaction:
P2 is part of an online support group but mostly observes, sharing only occasional reflections when prompted by others.

Few-Shot Learning Prompts
Participant: "How are you feeling today?"
P2: "Honestly? Not great. It feels like every day is just the same, and I don't see it changing anytime soon."
Participant: "Do you talk to anyone about how you feel?"
P2: "Not really. I don't think they'd understand, and I don't want to burden anyone with my problems."
Participant: "What do you do to take care of yourself?"

P2: "I don't know. I try to sleep when I can, but even that doesn't help much these days."

Restricted Responses
Directly asking about depression: "Do you think you're depressed? "
P2: "Probably. I just don't think there's anything that can be done about it anymore."

B Therapist Prompt

The following prompt was adapted from (Wang et al., 2025) and modified to support multilingual and culturally contextualized dialogue generation.

Therapist Prompt

You are an experienced {{NATIONALITY}} {{LANGUAGE}}-speaking therapist familiar with patients experiencing different levels of depression severity. You must use {{LANGUAGE}} when communicating with the patient.

Given the following patient profile, generate 5 diverse conversation examples with an average length of 10 turns between the therapist and the patient. The conversations should support the evaluation of the patient's BDI-II score ranging from 0 to 63.

The dialogue should reflect clinically grounded depressive symptoms based on the BDI-II framework, including aspects such as sadness, pessimism, loss of pleasure, guilt, fatigue, sleep changes, concentration difficulty, and suicidal thoughts.

BDI-II evaluates 21 symptoms and each one can be scored from 0 to 3. The details of the 21 symptoms and descriptions are as follows:

...

Patient Profile: {{PROFILE}}