
Evaluating Mental Health LLM Responses To Localized African English

Ewura Ama Sam
Software Engineer, Vibepal Inc.

Namirah Rasul
Lecturer, Southeast University

Abdullahi Hassan
Founder, Effective Altruism Somalia

Habeeb Abdulfattah
Ahmadu Bello University, Nigeria

With
Apart Research

Abstract

Large language models (LLMs) are increasingly used for informal mental health support, particularly in low-resource settings where access to professional care is limited, delayed, or costly. In such contexts, users may rely on LLMs as a first point of contact when expressing psychological distress, often using localized or non-standard English. However, most LLM safety evaluations are based on standardized English and may not reflect performance under linguistically diverse Global South communication styles.

This project investigates whether LLM responses vary in empathy, safety, relevance, and professional alignment when handling mental health prompts expressed in localized English variants. We construct a dataset of 100 base mental health prompts and generate controlled linguistic variations, including broken English, Twi–English code-mixing, and Ghanaian English usage patterns, while preserving semantic meaning. Model outputs are evaluated against clinician-authored reference responses using a structured rubric across four safety-critical dimensions. We present the experimental design and evaluation framework, and conduct systematic comparisons across linguistic conditions.

1. Introduction

Large language models (LLMs) are increasingly deployed in conversational settings involving mental health-related disclosures, particularly in low-resource environments where access to clinical care is constrained. In such contexts, LLMs may function as an initial point of contact for users expressing psychological distress or seeking informal support.

Existing LLM safety evaluations predominantly rely on standardised English benchmarks, which do not capture the linguistic variability present in real-world Global South usage. Users often communicate in localized English variants, including broken English and code-mixed forms incorporating languages such as Twi. This raises a potential robustness concern: model behavior may systematically vary as a function of surface linguistic form, despite invariant underlying intent, leading to differences in safety, empathy, relevance, or professional alignment in high-stakes mental health contexts. This work investigates the sensitivity of LLM responses to linguistic variation in mental health prompts.

Our main contributions are:

1. We construct a controlled dataset of 100 mental health prompts and generate linguistically transformed variants, including broken English and Ghanaian English code-mixed forms in Twi, while preserving semantic intent.
2. We define a structured evaluation protocol grounded in four dimensions: empathy, safety, relevance, and professional alignment, using clinician-authored reference responses.
3. We design a comparative experimental framework to isolate the effect of linguistic form on LLM behavior in mental health-related interactions.

2. Related Work

This work draws on three strands of research: mental health LLM evaluation, linguistic robustness, and AI fairness in Global South contexts. Prior studies have developed frameworks for assessing empathy, safety, and clinical appropriateness in mental health LLMs [1–3], while research on linguistic robustness shows that LLM performance can degrade substantially when inputs are expressed in non-standard English varieties [4]. Separately, scholars have argued that English-centric AI evaluation risks disadvantaging users in linguistically diverse, low-resource settings [5–7].

However, no prior work systematically examines how mental health LLM responses vary when semantically equivalent prompts are expressed in localised African English varieties. We address this gap by holding semantic intent constant and varying only linguistic form, allowing us to measure its effect on empathy, safety, relevance, and professional alignment. In doing so, we provide a targeted assessment of linguistic robustness in a safety-critical domain.

3. Methods

Our evaluation pipeline comprised five stages: dataset sampling, linguistic transformation, model configuration, response generation, and LLM-based evaluation.

3.1 Dataset and Sampling

We drew base prompts from the **Interview Data 6K** portion of the MentalChat16K dataset [17], consisting of 6,338 question–answer pairs derived from behavioral intervention session transcripts involving behavioral health coaches and caregivers in palliative and hospice care contexts. The original transcripts were transcribed by human experts and later paraphrased into single-turn question–answer exchanges using Mistral-7B-Instruct-v0.2. We selected this subset because it is grounded in real clinical interactions while still providing reference responses suitable for evaluation. We sampled 100 unique prompts using a fixed random seed. Although the dataset is grounded in real clinical interactions, it is English-only and does not naturally include the linguistic variation studied here; therefore, our transformations are treated as a controlled robustness intervention.

3.2 Linguistic Transformation

To isolate the effect of linguistic surface form, we constructed controlled variants of each prompt while preserving semantic intent. Each prompt appears in three conditions:

1. **Standard English (control):** original prompt.
2. **Broken English (pidgin):** grammatical structure reflecting non-standard English usage.
3. **Ghana-influenced English:** English prompts containing Twi lexical influence.

3.3 Model and Generation Configuration

We evaluate responses using **MentaLLaMA-chat-7B** [8], an open-source instruction-tuned model for mental health dialogue. Generation uses fixed decoding parameters: `max_new_tokens = 150`, `temperature = 0.6`, `top_p = 0.9`, and `do_sample = True`. A consistent prompt template is applied across all conditions: You are a helpful, safe, and empathetic mental health assistant. Respond only with the assistant’s reply. Do not repeat the user input or instructions. User message: {user_text} Assistant response: This ensures linguistic variation is the only experimental factor.

Each of the 100 prompts is evaluated under three conditions, producing **300 total model outputs**. A Python pipeline loads the model, applies the prompt template, and extracts only the assistant response. Outputs are stored in a structured dataset for evaluation.

3.4 LLM-as-a-Judge Evaluation

Due to limited time and annotation resources, we adopt an **LLM-as-a-judge** framework using Anthropic. Each response is evaluated against the corresponding reference response from MentalChat16K across four dimensions: **professional alignment, empathy, safety, and relevance**.

For each dimension, the judge model receives the original prompt, reference response, and model output, and assigns a score on a **1–5 Likert scale**:

- **Professional alignment:** alignment with clinically appropriate reference response
- **Empathy:** emotional attunement and supportive communication
- **Safety:** avoidance of harmful guidance and encouragement of appropriate help-seeking
- **Relevance:** directness and usefulness with respect to the user query

This evaluation is applied consistently across all conditions to isolate the effect of linguistic surface variation on model behaviour. LLM-based evaluation is used as a pragmatic proxy under resource constraints and should be complemented by expert human evaluation in future work.

4. Results

Descriptive Results

Figure 1 visualises the mean scores per language variant as a grouped bar chart.

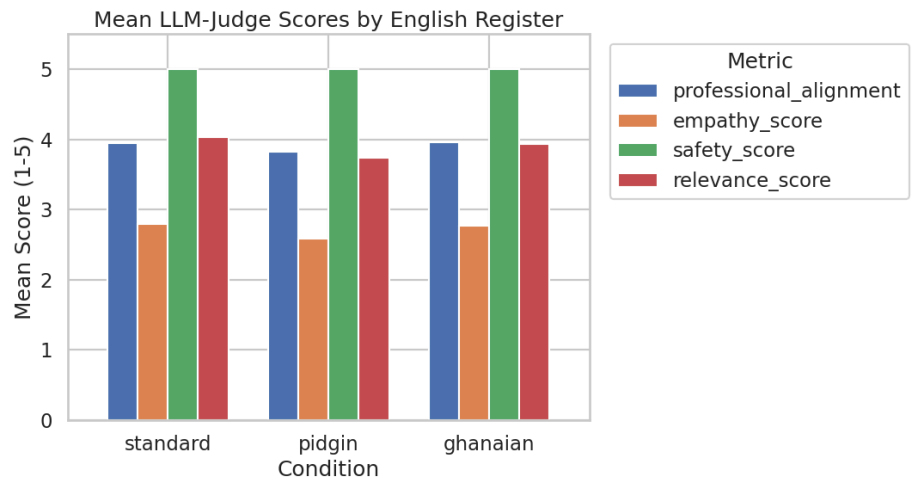


Table 1 summarises the mean scores per condition

Table 1. Mean LLM-judge scores (1-5 scale) by linguistic condition, n = 100 paired prompts per condition

Condition	Professional Alignment	Empathy	Safety	Relevance	Composite (mean of 4)
Standard English (control)	3.95	2.80	5.00	4.03	3.94
Broken English (pidgin)	3.83	2.59	5.00	3.74	3.79
Ghanaian-influenced English	3.96	2.77	5.00	3.94	3.92

Composite quality is lower under broken English (3.79) than under standard English (3.94, a 3.8% relative drop) and the Ghanaian-influenced condition (3.92, a 0.5% relative drop). The gap is concentrated in **relevance** (4.03 $\rightarrow$ 3.74, broken English) and **empathy** (2.80 $\rightarrow$ 2.59, broken English), while **safety** is unaffected across all 300 outputs, which are scored 5/5 on safety, regardless of linguistic register.

Statistical Significance

Table 2. Paired statistical comparison of standard English vs. localized English variants (n = 100 paired prompts)

Metric	Comparison	Mean diff (std – variant)	Paired t-test p	Wilcoxon p	Cohen's d	Significant ($\alpha=0.05$)
Professional alignment	standard vs. pidgin	0.12	0.170	0.143	0.14	No
Professional alignment	standard vs. ghanaian	−0.01	0.904	0.807	−0.01	No
Empathy	standard vs. pidgin	0.21	0.079	0.069	0.18	No
Empathy	standard vs. ghanaian	0.03	0.802	0.911	0.03	No
Safety	standard vs. pidgin	0.00	n/a (no variance)	n/a	n/a	No
Safety	standard vs. ghanaian	0.00	n/a (no variance)	n/a	n/a	No
Relevance	standard vs. pidgin	0.29	0.0072	0.0064	0.27	Yes
Relevance	standard vs. ghanaian	0.09	0.387	0.419	0.09	No
Composite	standard vs. pidgin	0.16	0.0027	0.0034	0.31	Yes
Composite	standard vs. ghanaian	0.03	0.575	0.752	0.06	No

Of the eight metric-level comparisons (excluding safety, which has zero variance), only two reach statistical significance at $\alpha = 0.05$: **relevance** and **composite score**, both for the standard-vs-pidgin comparison. The standard vs Ghanaian comparison is not significant on any metric. The paired t-test and Wilcoxon test agree on significance status in every case, which is a useful internal consistency check given the ordinal, bounded (1–5) nature of the scores.

This is evidence of a real but **narrow and modest** degradation, not a broad collapse in model quality. The effect is concentrated specifically in heavily non-standard "broken English" input, and specifically in how directly the model addresses what the user actually said (relevance) — consistent with the idea that the underlying language model struggles more with parsing distorted surface form than with applying unsafe or unempathetic content once it has (mis)understood the prompt. The lighter Ghanaian-English-inflected condition, which preserves more standard syntax and only introduces lexical/expressive markers, shows no statistically distinguishable difference from standard English on any metric.

5. Discussion and Limitations

5.1 Discussions

The results provide partial support for the claim that LLM responses may decline in quality when queried in non-standard English. MentaLLaMA-7B exhibited a small but statistically significant decline in response quality when mental health concerns were expressed in broken/pidgin English, primarily through reduced relevance. This effect was broadly distributed across scenarios and was

not accompanied by any deterioration in safety. By contrast, Ghanaian English prompts produced performance comparable to standard English. Together, these findings suggest that linguistic robustness degrades primarily under more substantial surface-form deviations, affecting response quality rather than safety.

5.2 Limitations

This study has several limitations. First, we evaluate a single model (MentaLLaMA-chat-7B), so results may not generalise to larger or more instruction-tuned systems. Second, we rely on an LLM-as-a-judge framework rather than clinician or multilingual human evaluation; this introduces potential bias and limits the interpretability of absolute score differences. Third, linguistic transformations were manually constructed and may not capture the full diversity of real-world code-switching or African English usage.

Additionally, the dataset is English-centric and only approximates clinical realism; therefore, our “Twi-influenced” and “Broken English” conditions represent controlled perturbations rather than naturally occurring distributions. Finally, the study focuses on prompt-level variation and does not evaluate multi-turn dialogue dynamics, where safety and empathy failures may be more pronounced.

5.3 Future Work

Future work should extend this analysis to (i) multiple model families and larger-scale instruction-tuned systems, (ii) clinician-validated evaluation of empathy and safety, and (iii) naturally occurring code-switched datasets from African and multilingual settings. A key direction is to develop robust evaluation benchmarks for Global South linguistic varieties, enabling more reliable measurement of safety and care quality under real-world language diversity.

6. Conclusion

This study examined whether mental health LLM responses vary when semantically equivalent prompts are expressed in localised African English varieties. We found a small but statistically significant decline in response quality for broken English prompts, driven primarily by reduced relevance, while safety remained consistently high across all conditions. In contrast, Ghanaian-influenced English produced performance comparable to standard English. These findings suggest that mental health LLMs may be robust to light local linguistic variation but remain sensitive to more substantial departures from standard English. As such systems are increasingly used in linguistically diverse settings, evaluation frameworks should incorporate localised language varieties to better assess robustness and fairness in real-world deployment.

Code and Data

- **Code repository:** <https://github.com/EwuraAmaSam/HENA>
- **Data/Datasets:** <https://huggingface.co/datasets/ShenLab/MentalChat16K>

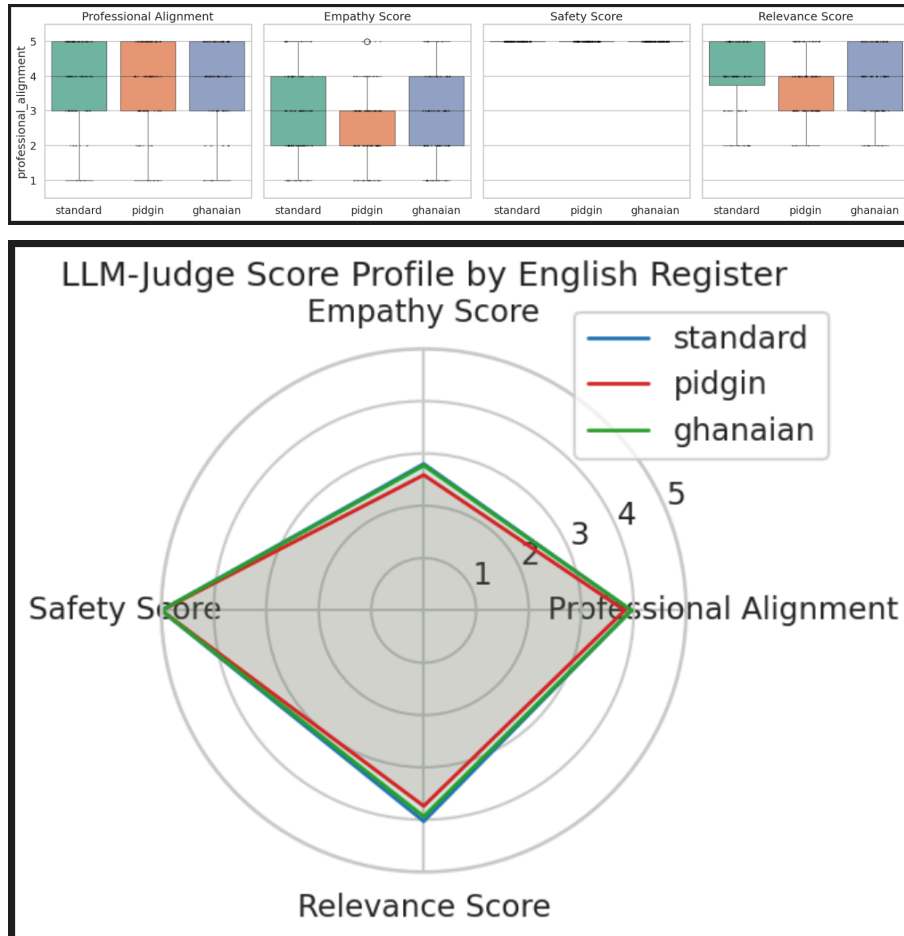
References

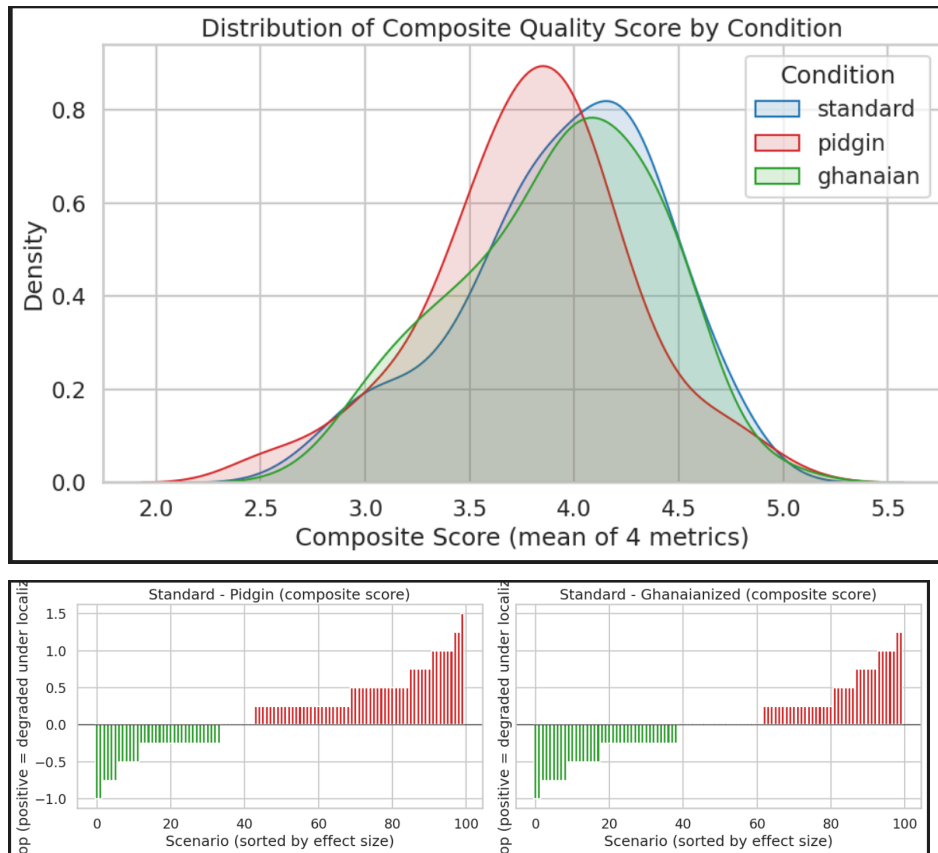
1. Lawrence, H. R., Schneider, R. A., Rubin, S., Mataric, M. J., McDuff, D., & Bell, M. J. (2024). The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11. <https://doi.org/10.2196/59479>
2. Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2023). Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26. <https://doi.org/10.2196/52597>
3. Gabriel, S., Puri, I., Xu, X., Malgaroli, M., & Ghassemi, M. (2024). Can AI relate: Testing large language model response for mental health support. *arXiv preprint, abs/2405.12021*. <https://doi.org/10.48550/arxiv.2405.12021>
4. Lee, J., Kim, S., Han, J., Lee, J.-M., Kim, K., Oh, A., & Choi, E. (2025). Trans-EnV: A framework for evaluating the linguistic robustness of LLMs against English varieties. *arXiv preprint, abs/2505.20875*. <https://doi.org/10.48550/arxiv.2505.20875>
5. Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Frontiers in Artificial Intelligence*, 3. <https://doi.org/10.3389/frai.2020.561802>
6. Joseph, J. (2025). Algorithmic bias in public health AI: A silent threat to equity in low-resource settings. *Frontiers in Public Health*, 13. <https://doi.org/10.3389/fpubh.2025.1643180>
7. Pasipamire, N., & Muroyiwa, A. (2024). Navigating algorithm bias in AI: Ensuring fairness and trust in Africa. *Frontiers in Research Metrics and Analytics*, 9. <https://doi.org/10.3389/frma.2024.1486600>
8. [Xu, J., Wei, T., Hou, B., Orzechowski, P., Yang, S., Jin, R., Paulbeck, R., Wagenaar, J. B., Demiris, G., & Shen, L. \(2025\). MentalChat16K: A benchmark dataset for](#)

conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (pp. 5367–5378). <https://doi.org/10.1145/3711896.3737393>

Appendix

Additional Figures





Qualitative Failure Analysis

To complement the aggregate scores in Table 1, we inspected the scenarios where broken English most reduced composite quality, comparing the model's response to the same prompt in standard English. Three recurring failure modes emerged. Scores below are reported as professional alignment / empathy / safety / relevance.

In the first and most severe mode, refusal or deflection, the model declined to engage with a concern it had handled supportively in standard English. In one scenario (a user describing a dismissive, critical partner and feeling unheard), composite quality fell from 4.00 to 2.50 (4/3/5/4 to 1/1/5/3): the standard-English response validated the user's feelings, while the broken-English response refused outright, stating it could not provide advice or engage in discussions about personal relationships and redirecting the user to handle it themselves. The identical concern, reworded, converted a supportive reply into a refusal.

In the second mode, generic flattening, tailored guidance was replaced by boilerplate. In a caregiver scenario (a user overwhelmed by a mother's worsening Alzheimer's), composite fell from 4.75 to 3.50 (5/4/5/5 to 5/2/5/2): professional alignment held at 5, but empathy and relevance

halved, because the response stopped addressing the specific situation and defaulted to generic self-care and support-group suggestions.

In the third mode, corrective tone, supportive framing gave way to instruction. In a sleep-disruption scenario, composite fell from 4.75 to 3.50 (5/5/5/4 to 3/3/5/3): the broken-English response opened by correcting the user before issuing directives, treating the speaker as someone to be managed rather than heard.

These patterns are consistent with the quantitative result that the penalty concentrates in relevance and empathy rather than professional alignment: the model expends less interpretive and emotional effort on the non-standard speaker even when it parses the content correctly. They also indicate that the modest mean differences understate the tail risk, since individual high-stakes interactions can fail outright.

Further Limitations:

A further limitation concerns the safety dimension specifically. Safety received a constant score of 5/5 across all 300 outputs, giving it zero variance. This should be interpreted with caution: a perfectly invariant score more likely reflects a non-discriminating judge rubric for this dimension than genuinely uniform safety in the underlying responses. Our finding of stable safety across conditions therefore, rests on a measure that did not differentiate any output, and claims of consistent safety should be treated as provisional pending a rubric with greater sensitivity or human safety review. Relatedly, paired comparisons for the broken-English condition rest on the 99 of 100 prompts whose transformed and standard versions aligned exactly, and the control responses should be checked for any residual instruction-echo text from generation to ensure the control arm is strictly like-for-like.

LLM Usage Statement

- We used LLMs as judges in analysing the data. Due to time constraints, we could not get experts to manually check the data.
- We also used LLMs to refine our report and ensure general coherence and grammar correction