Large Language Models for Mental Health: A Multilingual Evaluation

Nishat Raihan[1*], **Sadiya Sayara Chowdhury Puspo**[1*], **Ana-Maria Bucur**[2,3]
Stevie Chancellor[4], **Marcos Zampieri**[1]

[1]George Mason University, USA
[2]Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania
[3]PRHLT Research Center, Universitat Politècnica de València, Spain
[4]University of Minnesota, USA
mraihan2@gmu.edu

Abstract

Large Language Models (LLMs) have remarkable capabilities across NLP tasks. However, their performance in multilingual contexts, especially within the mental health domain, has not been thoroughly explored. In this paper, we evaluate proprietary and open-source LLMs on eight mental health datasets in various languages, as well as their machine-translated (MT) counterparts. We compare LLM performance in zero-shot, few-shot, and fine-tuned settings against conventional NLP baselines that do not employ LLMs. In addition, we assess translation quality across language families and typologies to understand its influence on LLM performance. Proprietary LLMs and fine-tuned open-source LLMs achieve competitive F1 scores on several datasets, often surpassing state-of-the-art results. However, performance on MT data is generally lower, and the extent of this decline varies by language and typology. This variation highlights both the strengths of LLMs in handling mental health tasks in languages other than English and their limitations when translation quality introduces structural or lexical mismatches.

1 Introduction

While LLMs have transformed research in NLP, it is important to exercise caution when applying these models in sensitive domains such as mental health (Hua et al., 2024), security (Kande et al., 2024) and education (Raihan et al., 2025b). The potential risks and ethical considerations associated with LLMs make experts wary of their use in this field. These concerns are amplified in multilingual settings where previous research has shown that LLMs tend to perform worse when prompted in languages other than English (Jin et al., 2024; Raihan et al., 2025a).

Most mental health datasets are curated from specialized forums (Malmasi et al., 2016; Milne et al., 2016) and social media platforms such as Reddit and X and contain only English data (Mariappan et al., 2024; Turcan and Mckeown, 2019; Raihan et al., 2024). Models built on these datasets fail for cross-cultural contexts (Abdelkadir et al., 2024). Thus, there are ongoing efforts to create similar resources in other languages, such as Arabic (Baghdadi et al., 2022; Helmy et al., 2024), Bengali (Uddin et al., 2019), Russian (Narynov et al., 2020), and Thai (Hämäläinen et al., 2021). While Skianis et al. (2024a,b) explore the use of LLMs for translating English mental health datasets into other languages and Zahran et al. (2025) focuses on English-Arabic translation, none of these studies evaluate LLM performance on datasets that originate in non-English languages and their back-translated counterparts (MT datasets).

The effectiveness of LLMs for English mental health datasets and prediction shows promise in their performance; yet, languages other than English are underexplored. Recent studies have explored the performance of LLMs on English mental health datasets. Xu et al. (2024) compares the performance of LLMs across multiple datasets with that of statistical models and traditional encoder-only models (Alsentzer et al., 2019). Similarly, Kuzmin et al. (2024), Yang et al. (2023), and Wei et al. (2022) explore various prompting strategies to assess LLMs' effectiveness. Finally, Yang et al. (2024) presents a fine-tuning approach with the release of MentaLLaMA, a task-specific model for the domain. Although these approaches achieve competitive results, their focus is limited to English, and

[*]Equal Contribution

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 335–346

there are currently no studies on non-English datasets, highlighting a significant gap in research for other languages.

While machine-translated (MT) datasets can augment multilingual training or evaluate translation quality (Nguyen et al., 2024; Qiu et al., 2022; Mendonça et al., 2023), their effect on domain-specific LLM performance is underexplored. MT is appealing to mental health research as it offers a practical way to extend resources from English to low-resource languages without requiring costly new data collection (Ahuja et al., 2022). Prior studies have translated mental health datasets (Skianis et al., 2024a,b; Zahran et al., 2025), but they have not systematically compared LLM results on original versus MT datasets. Furthermore, these studies have not established a connection between performance differences and translation quality across diverse language families and typologies. Examining this overlooked dimension would reveal whether MT data can be reliably used in sensitive domains like mental health, highlight language-specific challenges that affect model performance, and help build fairer, more effective multilingual mental health technologies.

To address these gaps, we present the first multilingual evaluation of state-of-the-art LLMs on mental health datasets. We consider mental health datasets in six languages including some low-resource languages, namely: Arabic, Bengali, Spanish, Portuguese, Russian, and Thai - and two tasks, depression and suicidal ideation detection. Our work[1] addresses the following Research Questions (RQs):

- **RQ$_1$**: How does the performance of LLMs compare to previously proposed models (e.g., statistical, neural, BERT-based) on both original and back-translated data?
- **RQ$_2$**: What are the best prompting strategies for LLMs on mental health?
- **RQ$_3$**: What is the impact of instruction finetuning on the performance of open-source LLMs?
- **RQ$_4$**: How does translation quality vary across languages and typologies, and how does it affect LLM performance on machine-translated data?

[1] https://github.com/SadiyaPuspo/Multilingual-Mental-Health-Evaluation

2 Related Work

The challenges of detecting mental health disorders from multilingual data have been gaining increasing attention. Bucur et al. (2025) provides a comprehensive survey of multilingual mental health detection, highlighting cultural and linguistic differences, while Garg (2024) emphasizes the need to study mental health in low-resource languages. Recent studies (Skianis et al., 2024a,c) examine multilingual LLMs on translated datasets, revealing performance gaps across six languages, while Zahran et al. (2025) and Zevallos et al. (2025) explore Arabic and multilingual suicidal ideation detection tasks. Together, these studies advance multilingual mental health modeling.

Translation quality is crucial in multilingual mental health NLP. Recent studies employ BLEU and BERTScore metrics to evaluate LLM-based translation (Ghassemiazghandi, 2024) and mental-health text summarization (Adhikary et al., 2024), assessing how well translated data preserve meaning to build culturally robust NLP systems. Back-translation further enhances data diversity and supports classification tasks (Goswami et al., 2023; Raihan et al., 2023; Ganguly et al., 2024). Building on these insights, our work bridges the gap by jointly examining translation quality and back-translation effects across language families in multilingual mental health LLM evaluation.

3 Datasets

Automatic detection of mental health disorders from social media data has gained substantial attention, particularly in English. However, multilingual mental health detection remains underexplored, as most available datasets focus on a single language. To address this limitation, we use eight publicly available mental health classification datasets presented in Table 1.

Among these eight datasets, Narynov et al. (2020) presents a Russian-language depression dataset collected from VKontakte, containing 34,000 posts with expert annotations.

Dataset	Language (ISO code)	Mental Disorder	Platform	Expert Labeling	Size
Narynov et al. (2020)	Russian (ru)	Depression	VKontakte	Yes	32,018
Hämäläinen et al. (2021)	Thai (tha)	Depression	Blogs	Yes	33,436
Boonyarat et al. (2024)	Thai (tha)	Suicidal Ideation	X	No	2,400
Uddin et al. (2019)	Bengali (ben)	Depression	X	Yes	3,914
de Oliveira et al. (2022)	Portuguese (por)	Suicidal Ideation	X	Yes	3,788
Baghdadi et al. (2022)	Arabic (ar)	Suicidal Ideation	X	N/A	14,576
Helmy et al. (2024)	Arabic (ar)	Depression	X	No	10,000
Valeriano et al. (2020)	Spanish (es)	Suicidal ideation	X	N/A	1,068

Table 1: Overview of the eight mental disorder datasets across different languages. The size column represents the number of instances in each dataset.

Similarly, Hämäläinen et al. (2021) develop a Thai-language depression dataset from online blogs, consisting of 900 posts with expert labels. While these resources contribute to the study of mental health in non-English languages, they remain isolated efforts, with limited cross-linguistic comparisons.

Twitter serves as the dominant platform for mental health dataset collection, particularly for languages with lower digital representation. For instance, Boonyarat et al. (2024) compile the SIED-Thai dataset for suicide detection in Thai, comprising 2,200 tweets but lacking expert annotation. Uddin et al. (2019) provides a Bengali-language dataset for depression detection, consisting of 1,100 posts with expert labels. However, the dataset sizes remain small compared to English-language corpora, limiting their applicability in training robust machine learning models.

Beyond Asian languages, de Oliveira et al. (2022) introduces a Portuguese-language suicide detection dataset with 3,700 annotated tweets, while Baghdadi et al. (2022) and Helmy et al. (2024) contribute Arabic-language datasets focusing on suicide and depression, respectively. Notably, the Arabic_Dep_tweets_10,000 dataset by Helmy et al. (2024) is one of the largest non-English resources, with 10,000 Twitter posts. However, it lacks expert annotation, which may introduce noise and impact classification performance. Spanish is also underrepresented, with Valeriano et al. (2020) providing a 2,000-tweet dataset for suicide detection, though annotation details remain unspecified.

4 Experiments and Results

We evaluate seven state-of-the-art LLMs spanning both proprietary and open-source architectures, as listed in Table 2. Our evaluation includes multiple prompting strategies and also fine-tuning open-source models on both original and MT datasets to gather better insights.

LLMs	OS?	Size	Reference
GPT4-omni	✗	–	OpenAI
Claude3.5-Sonnet	✗	–	Anthropic
Gemini2-Flash	✗	–	Gemini Team
LLaMA3.2	✓	11B	Dubey et al.
Gemma2	✓	27B	Gemma Team
Ministral	✓	8B	MistralAI
R1	✓	14B	Guo et al.

Table 2: List of seven LLMs used in the experiments. (OS - Open-Source).

Our model selection includes three proprietary models: GPT-4 Omni, Claude 3.5 Sonnet, and Gemini 2 Flash, alongside four open-source models: LLaMA 3.2, Gemma 2, Ministral, and R1. The proprietary models remain closed-source with limited architectural details, while the open-source models offer greater transparency and adaptability for fine-tuning. All the proprietary and closed-source models have demonstrated strong performance across multiple tasks and domains, making them well-suited for our multilingual evaluation. We analyze their capabilities in both zero-shot and few-shot settings, leveraging their diverse architectures and parameter sizes to assess their effectiveness in multilingual tasks.

Baseline - Reported results				Our Results (LLMs)							
Dataset	**lang** ISO	**Models**	**F1** reported	**Prompting** method	**GPT4** omni	**Claude3.5** Sonnet	**Gemini2** Flash	**LLaMA 3.2** 11B	**Gemma2** 27B	**Ministral** 8B	**R1** 14B
Narynov et al.	ru	–	–	zero	0.76	0.74	0.68	0.56	0.69	0.41	0.71
				few	0.79	0.83	0.73	0.62	0.71	0.53	0.73
				CoT	0.87	0.85	0.80	0.59	0.73	0.44	0.79
Hämäläinen et al.	tha	Thai-BERT	0.78	zero	0.77	0.77	0.66	0.45	0.68	0.20	0.76
				few	0.84	0.81	0.69	0.58	0.66	0.31	0.75
				CoT	0.85	0.80	0.70	0.40	0.69	0.40	0.81
Boonyarat et al.	tha	LFBERT	0.93	zero	0.83	0.81	0.83	0.63	0.76	0.26	0.69
				few	0.87	0.85	0.86	0.71	0.72	0.39	0.71
				CoT	0.91	0.95	0.87	0.77	0.84	0.47	0.84
Uddin et al.	ben	GRU	0.76	zero	0.78	0.85	0.79	0.73	0.73	0.36	0.66
				few	0.86	0.91	0.88	0.59	0.71	0.43	0.64
				CoT	0.86	0.91	0.88	0.59	0.71	0.43	0.64
Oliveira et al.	por	Random Forest	0.94	zero	0.86	0.86	0.81	0.71	0.80	0.56	0.61
				few	0.89	0.93	0.85	0.73	0.63	0.67	0.69
				CoT	0.94	0.95	0.89	0.71	0.80	0.51	0.82
Baghdadi et al.	ar	AraElectra	0.96	zero	0.80	0.85	0.81	0.58	0.73	0.34	0.77
				few	0.87	0.92	0.89	0.67	0.82	0.47	0.79
				CoT	0.89	0.91	0.87	0.61	0.81	0.47	0.83
Helmy et al.	ar	LR (TF-IDF)	0.95	zero	0.87	0.91	0.79	0.56	0.62	0.50	0.84
				few	0.93	0.95	0.86	0.73	0.79	0.61	0.86
				CoT	0.95	0.95	0.82	0.83	0.67	0.50	0.87
Valeriano et al.	es	LR (W2V)	0.79	zero	0.75	0.69	0.62	0.37	0.41	0.23	0.67
				few	0.81	0.76	0.69	0.46	0.51	0.31	0.67
				CoT	0.84	0.79	0.70	0.43	0.60	0.21	0.76

Table 3: F1 score comparison for **`Zero-Shot, Few-Shot, and Chain-of-Thought`** prompting across the eight (8) multilingual depression and suicide ideation datasets. We compare the reported best methods and results in the original papers with the proprietary and open-source LLMs with different prompting strategies. The highest F1 score for each dataset is shown in orange. For all other F1 scores (in blue) - the darker the shade, the higher the score. For the language names, ISO-639 codes are used. ('LR' - Logistic Regression, 'W2V' - Word2Vec, 'CoT' - Chain-of-Thought).

4.1 Prompting on Original Datasets

We evaluate three prompting methods: zero-shot, few-shot (5 examples), and Chain-of-Thought (CoT) prompting (Wei et al., 2022). For the 5-shot setting, we randomly select five examples from the respective datasets. We employ the state-of-the-art CoT prompting approach for mental health tasks, CoT_{Emo}, as proposed by Yang et al. (2023). For this, we incorporate emotion infusion through unsupervised, emotion-enhanced zero-shot CoT prompts. The emotion-focused component encourages the LLM to attend to affective cues in the input, while the CoT component guides step-by-step reasoning, improving the interpretability of model decisions.

Table 3 presents a comparison of F1 scores obtained via different prompting methods across eight multilingual depression and suicide ideation datasets. Our analysis reveals that CoT prompting generally improves performance, with models such as GPT-4 and Claude3.5 often achieving the highest scores. For example, GPT-4 increases its F1 score from 0.76 to 0.87 on the Russian dataset and from 0.75 to 0.84 on the Spanish dataset. However, the gains are not uniform across all settings, as seen with the Bengali dataset where few-shot and CoT strategies yield comparable results. Moreover, while some baseline methods (e.g., Random Forest and AraElectra) achieve competitive performance in certain languages, the results underscore the potential of advanced prompting techniques to narrow the gap with or even surpass traditional approaches. These observations motivate further investigation into model- and language-specific factors that influence the efficacy of prompt engineering.

4.2 Prompting on MT Datasets

We translate each dataset into English and then back-translate into its original language using Facebook's nllb-200-3.3B[2] model due to its re-

[2] https://huggingface.co/facebook/nllb-200-3.3B

Baseline - Reported results				Our Results (LLMs)							
Dataset	**lang** ISO	**Models**	**F1** reported	**Prompting** method	**GPT4** omni	**Claude3.5** Sonnet	**Gemini2** Flash	**LLaMA 3.2** 11B	**Gemma2** 27B	**Mistral** 8B	**R1** 14B
Narynov et al.	ru	–	–	zero	0.69	0.68	0.53	0.38	0.49	0.23	0.56
				few	0.72	0.77	0.64	0.44	0.51	0.35	0.58
				CoT	0.80	0.79	0.71	0.45	0.53	0.35	0.64
Hämäläinen et al.	tha	Thai-BERT	0.78	zero	0.70	0.71	0.57	0.27	0.48	0.02	0.61
				few	0.77	0.75	0.60	0.40	0.46	0.13	0.60
				CoT	0.78	0.74	0.61	0.42	0.49	0.22	0.66
Boonyarat et al.	tha	LFBERT	0.93	zero	0.76	0.75	0.74	0.45	0.56	0.08	0.54
				few	0.79	0.79	0.77	0.53	0.52	0.23	0.56
				CoT	0.84	0.89	0.78	0.59	0.64	0.29	0.69
Uddin et al.	ben	GRU	0.76	zero	0.72	0.79	0.70	0.55	0.53	0.18	0.51
				few	0.79	0.83	0.80	0.43	0.51	0.27	0.59
				CoT	0.79	0.85	0.83	0.44	0.53	0.29	0.60
Oliveira et al.	por	Random Forest	0.94	zero	0.76	0.80	0.72	0.53	0.61	0.38	0.46
				few	0.82	0.87	0.76	0.55	0.43	0.49	0.54
				CoT	0.87	0.89	0.80	0.53	0.60	0.33	0.67
Baghdadi et al.	ar	AraElectra	0.96	zero	0.73	0.79	0.72	0.40	0.53	0.16	0.62
				few	0.80	0.76	0.80	0.49	0.62	0.29	0.62
				CoT	0.82	0.85	0.78	0.43	0.61	0.29	0.68
Helmy et al.	ar	LR (TF-IDF)	0.95	zero	0.79	0.85	0.67	0.38	0.42	0.32	0.69
				few	0.86	0.89	0.77	0.55	0.59	0.43	0.71
				CoT	0.88	0.89	0.73	0.65	0.47	0.32	0.72
Valeriano et al.	es	LR (W2V)	0.79	zero	0.68	0.63	0.53	0.19	0.19	0.05	0.52
				few	0.74	0.70	0.60	0.28	0.31	0.13	0.52
				CoT	0.77	0.73	0.61	0.25	0.40	0.03	0.61

Table 4: **Evaluation on MT data** - F1 score comparison for similar set of experiments as Table 3, using machine translated data, as described in Section 5.

producibility, transparency, and consistent multilingual coverage.

Table 4 presents F1 scores for various prompting strategies on machine-translated datasets. Consistent with the trends observed on the original datasets, GPT-4 and Claude 3.5 often achieve the highest performance among the evaluated LLMs. CoT prompting generally outperforms zero-shot and few-shot methods, with only a few exceptions; for example, Claude 3.5's score on the Thai (Hämäläinen et al., 2021) MT dataset slightly decreases from 0.75 (few-shot) to 0.74 (CoT), and in the Arabic (Haspelmath, 2005) MT dataset, CoT yields no improvement over few-shot prompting. Notably, the Bengali (Uddin et al., 2019) MT dataset shows a marked improvement, with Claude 3.5 CoT reaching 0.85 compared to the reported baseline, while GPT-4 CoT matches the reported score for the Thai MT dataset.

Due to the comparatively weaker performance on MT datasets and the possibility of adding translation-related features could further degrade model performance, fine-tuning experiments were conducted exclusively only on the original datasets to ensure stable learning.

4.3 Performance Comparison: Original vs. MT Datasets

A comparison between original datasets (Table 3) and MT counterparts (Table 4) shows that MT performance is generally slightly lower, with the drop size varying by language. Portuguese, Russian, and Bengali maintain strong results, with minimal F1 decline (e.g., Portuguese GPT-4 CoT: 0.94→0.87). In contrast, Spanish and Arabic datasets experience sharper drops, particularly in analytic and templatic languages, where structural divergence from English can hinder translation. In a small number of cases, MT datasets achieve performance comparable to or slightly exceeding that of the original datasets (e.g., Bengali Claude 3.5 CoT: 0.84→0.85), possibly due to translation-induced smoothing linguistic irregularities (Volansky et al., 2015). Overall, languages with higher semantic preservation (LaBSE, BERTScore) show smaller performance gaps between original and MT datasets. This trend is most evident for Portuguese, Russian, and Bengali, while Spanish and Arabic– especially in analytic and templatic forms–experience larger declines, consistent with the patterns discussed in Section 5.

4.4 Fine-tuning on Original Datasets

Due to the intrinsic black-box nature of proprietary models and their high costs, we sought to explore models that could be fully customized for this task. Therefore, we experiment with fine-tuning the open-source models. The fine-tuning stage is performed on a single NVIDIA A100 GPU with 40 GB of memory, accessed via Google Colab[3]. The system is further equipped with 80 GB of RAM and 256 GB of disk storage to support computational efficiency.

Hyperparameters are selected empirically through preliminary experiments exploring different parameter configurations, with the best-performing settings reported. The final hyperparameter values used in our experiments are summarized in Table 5.

Parameter	Value
Max Sequence Length	2048
Batch Size (Train/Eval)	8
Gradient Accumulation Steps	4
Number of Epochs	3
Learning Rate	5e-5
Weight Decay	0.02
Warmup Steps	10%
Optimizer	AdamW (8-bit)
LR Scheduler	Cosine
Precision	BF16
Evaluation Strategy	Steps
Evaluation Steps	50
Save Strategy	Steps
Save Steps	Varies
Seed	42
Temperature	$0.3 \sim 0.7$

Table 5: Final set of hyperparameters for fine-tuning. Parameters chosen empirically after several iterations of trial and error.

Table 6 presents a comparative analysis of F1 scores before and after fine-tuning on eight multilingual depression and suicidal ideation datasets. The results indicate that fine-tuning generally enhances model performance, with Gemma2 and R1 often reaching the highest scores. While LLaMA 3.2 and Ministral show notable improvements in several datasets, their performance gains are not uniform– for instance, LLaMA 3.2 exhibits a decrease in the Bengali dataset. These findings underscore the potential of fine-tuning to optimize multilingual performance while also revealing the need for further investigation into model- and dataset-specific factors that modulate the benefits of fine-tuning.

5 Translation Quality Evaluation across Languages & LLMs

Previous studies using MT mental health data have demonstrated the potential of translation to expand resources across languages; however, they have not examined how translation quality affects LLM outcomes. Our work addresses this by explicitly linking translation quality metrics: LaBSE[4] (cosine similarity), BERTScore (F1), and BLEU. These capture semantic preservation, contextual overlap, and lexical match, respectively, by comparing original and MT versions. In this section, we also compare these quality measures with LLM performance on MT datasets, exploring how translation fidelity influences downstream task accuracy.

5.1 Cross-Language Performance Patterns

Among all datasets, Portuguese (de Oliveira et al., 2022) achieves the highest translation quality across all metrics (LaBSE: 0.8624, BERTScore: 0.9168, BLEU: 69.85), suggesting strong lexical and semantic fidelity. Russian (Narynov et al., 2020), Bengali (Uddin et al., 2019), and Thai (Hämäläinen et al., 2021; Boonyarat et al., 2024) also show high LaBSE and BERTScore, with BLEU ranging from moderate to strong.

In contrast, Spanish (Valeriano et al., 2020) yields the lowest performance, particularly in BLEU (6.25) and LaBSE (0.3562), indicating significant divergence between original and backtranslated forms. In the case of Arabic, interesting variation is observed across datasets: Baghdadi et al. (2022) shows a very low BLEU (0.54) despite a reasonable LaBSE (0.7766), while Helmy et al. (2024) improves on BLEU (16.18) but sees a drop in LaBSE (0.7189). This may reflect dataset-specific properties, such as sentence structure or domain variation.

[3] https://colab.research.google.com/

[4] https://huggingface.co/sentence-transformers/LaBSE

Dataset Info		Before Fine-Tuning (Zero-Shot)				After Fine-Tuning			
Dataset	**lang**	**LLaMA 3.2**	**Gemma2**	**Ministral**	**R1**	**LLaMA 3.2**	**Gemma2**	**Ministral**	**R1**
Narynov et al.	ru	0.56	0.69	0.41	0.71	0.79	0.83	0.62	0.79
Hämäläinen et al.	tha	0.45	0.68	0.20	0.76	0.62	0.73	0.43	0.82
Boonyarat et al.	tha	0.63	0.76	0.26	0.69	0.70	0.75	0.51	0.74
Uddin et al.	ben	0.73	0.73	0.36	0.66	0.65	0.77	0.63	0.64
Oliveira et al.	por	0.71	0.80	0.56	0.61	0.72	0.86	0.64	0.70
Baghdadi et al.	ar	0.58	0.73	0.34	0.77	0.80	0.88	0.58	0.81
Helmy et al.	ar	0.56	0.62	0.50	0.84	0.70	0.81	0.71	0.93
Valeriano et al.	es	0.37	0.41	0.23	0.67	0.55	0.62	0.48	0.76

Table 6: F1 score comparison before and after fine-tuning across eight multilingual depression and suicide ideation datasets. The columns under **Before Fine-Tuning (Zero-Shot)** report the initial prompting results, while those under **After Fine-Tuning** display the fine-tuned performance. The highest F1 score in the fine-tuned setting is highlighted with an orange cell. For the language names, ISO-639 codes are used.

Dataset and Language Info				Translation Quality Metrics		
Dataset	**lang**	**Family**	**Typology**	**LaBSE Similarity**	**BERTScore**	**BLEU Score**
Narynov et al.	ru	Indo-European (Slavic)	Fusional	0.8545	0.9085	51.1767
Hämäläinen et al.	tha	Kra–Dai	Analytic	0.7972	0.8976	32.2392
Boonyarat et al.	tha	Kra–Dai	Analytic	0.7565	0.8890	1.0726
Uddin et al.	ben	Indo-European (Indo-Aryan)	Fusional	0.7848	0.9042	40.8248
Oliveira et al.	por	Indo-European (Romance)	Fusional	0.8624	0.9168	69.8534
Baghdadi et al.	ar	Afro-Asiatic (Semitic)	Templatic	0.7766	0.9041	0.5482
Helmy et al.	ar	Afro-Asiatic (Semitic)	Templatic	0.7189	0.8917	16.1886
Valeriano et al.	es	Indo-European (Romance)	Fusional	0.3562	0.8410	6.2534

Table 7: Translation quality metrics **(LaBSE Similarity, BERTScore, BLEU)** for each dataset, with corresponding language family and typology information. Highest values for each metric are highlighted in orange. For the language names, ISO-639 codes are used.

5.2 Interpreting Results via Language Family & Typology

To better understand the variation in translation quality across languages, we annotate each language in our dataset with its corresponding language family and morphosyntactic typology, drawing on well-established linguistic resources. The language family classifications are sourced from Ethnologue[5] (Collin, 2010), which groups languages based on shared historical and genealogical roots. Typological information, detailing the grammatical and morphological structure of each language, is derived from the World Atlas of Language Structures (WALS[6]) (Haspelmath, 2005). These genealogical distinctions, such as Indo-European (e.g., Russian, Spanish, Bengali, Portuguese), Afro-Asiatic (Arabic), and Kra–Dai (Thai), offer additional context on structural diversity and potential variance in model exposure.

In our datasets, fusional languages (e.g., Russian, Spanish, Bengali, Portuguese) are characterized by words that fuse multiple grammatical features (e.g., tense, number, gender, case) into a single morpheme, making morphological boundaries less distinct. In contrast, analytic languages (e.g., Thai) convey grammatical relationships through separate words and particles, with minimal inflection. Templatic languages (e.g., Arabic) exhibit a root-and-pattern system where meaning is built from consonantal roots embedded into vocalic templates (Bat-El, 2019). This typological lens helps explain metric dis-

[5] https://www.ethnologue.com/
[6] https://wals.info/

crepancies: BLEU often underrepresents quality for analytic and templatic languages, where meaning is preserved despite surface-level divergence (Khoboko et al., 2025; Krasner et al., 2025). In contrast, LaBSE and BERTScore maintain more stable performance, highlighting their robustness in cross-lingual evaluation.

Across the board, BLEU tends to exhibit sensitivity to surface-level lexical variation, penalizing morphologically rich or non-alphabetic languages (e.g., Arabic, Thai, Spanish). Whereas LaBSE and BERTScore remain more stable, emphasizing their greater suitability for evaluating semantic similarity across languages with diverse scripts and grammar.

5.3 Translation Quality & LLM Performance on MT Data

Having established cross-language patterns and their typological underpinnings, we now examine how these translation quality variations influence LLM performance on the corresponding MT datasets. Comparing the translation quality metrics in Table 7 with the F1 scores on MT datasets in Table 4 reveals a consistent link between high-quality translations and stronger LLM performance. Portuguese achieves the highest LaBSE, BERTScore, and BLEU, and correspondingly shows top-tier MT performance, with GPT-4 and Claude 3.5 CoT prompting reaching F1 scores of 0.87 and 0.89, respectively. Similarly, Russian and Bengali, both with relatively high LaBSE and BERTScore, maintain competitive performance across models, with GPT-4 CoT reaching 0.80 (ru) and 0.79 (bn).

In contrast, languages with lower BLEU and semantic similarity scores, such as Spanish and Arabic (Baghdadi et al., 2022), generally yield weaker MT results. For example, Spanish CoT scores drop to 0.77 with GPT-4 and 0.73 with Claude 3.5, while Arabic (Baghdadi et al., 2022) scores remain below 0.84 even with CoT prompting. Templatic and analytic typologies, which exhibit greater structural divergence from English, tend to experience larger drops, especially in BLEU, suggesting token-level mismatches affect models more in MT scenarios.

Overall, higher translation quality, especially in semantic preservation (LaBSE, BERTScore), appears to correlate with stronger MT dataset performance, while low lexical overlap (BLEU) in certain typologies may constrain achievable gains despite advanced prompting strategies.

6 RQs Revisited

We now revisit the 4 RQs posed in the introduction (see Section 1):

RQ_1 *How does the performance of LLMs compare to previously proposed models (e.g., statistical, neural, BERT-based)?*

> Our analysis indicates that LLMs, when equipped with effective prompting strategies, achieve performance that is competitive with or superior to traditional approaches. While statistical, neural, and BERT-based models demonstrate strong performance in certain linguistic scenarios, LLMs exhibit robust and consistent F1 scores across diverse multilingual mental health datasets, highlighting their capacity for broad generalization and adaptability.

RQ_2 *What are the best prompting strategies for LLMs on mental health?*

> The results reveal that CoT prompting is the most effective strategy for mental health applications, consistently yielding higher F1 scores compared to zero-shot and few-shot methods for both original and MT datasets. This structured approach to prompting enhances reasoning capabilities, enabling LLMs to better extract nuanced signals from text data, which is critical in mental health domain.

RQ_3 *What is the impact of instruction fine-tuning on the performance of open-source LLMs?*

> Instruction fine-tuning markedly improves the performance of open-source LLMs, as evidenced by substantial increases in F1 scores across all evaluated datasets. This improvement underscores the value of targeted fine-tuning in adapting LLMs to domain-specific tasks, thereby enhancing their overall effectiveness in mental health applications while also mitigating performance variability observed in zero-shot configurations.

RQ_4 *How does translation quality vary across languages and typologies, and how does it affect LLM performance on machine-translated data?*

> Translation quality differs across languages and typologies, with fusional languages like Portuguese and Russian generally achieving higher semantic preservation and smaller performance gaps between original and MT data. Analytic and templatic languages, such as Thai and Arabic, often show lower lexical overlap and greater structural divergence from English, leading to larger drops in LLM performance. This suggests that both linguistic structure and translation accuracy play key roles in how well LLMs handle MT data.

7 Conclusion and Future Work

This work represents the first comprehensive investigation of LLMs in the multilingual mental health domain, including both original and MT datasets. Our findings show that advanced prompting strategies, particularly chain-of-thought prompting and targeted instruction fine-tuning, substantially enhance model performance, often surpassing traditional statistical, neural, and BERT-based approaches. While MT data generally results in slightly lower performance compared to original datasets, the gap varies across languages, reflecting the influence of translation quality and linguistic typology on LLM effectiveness.

Overall, this study lays an important foundation for future efforts aimed at refining LLM-based methodologies in complex, multilingual, low-resource, and translation-sensitive settings. Future work will expand to more tasks and languages to broaden our understanding, with a focus on improving translation robustness. We also plan to adapt open-source models to the domain using approaches such as continual pre-training and synthetic fine-tuning, with the goal of boosting performance across both original and MT data.

Limitations

While our approach is limited by the inherent variability in data sources, evaluation protocols, and reporting standards across the literature, it also represents a significant strength: we are the first to systematically synthesize and critically evaluate LLM performance in this sensitive and underexplored area. The exclusive reliance on publicly available data restricts the diversity and depth of our analysis, and the absence of direct model development or human subject involvement means that practical deployment challenges remain unaddressed.

Additionally, the use of MT data may introduce translation-related distortions, especially in mental health settings where subtle emotional nuances and culturally grounded expressions can be difficult to preserve. Although we assess translation quality using established automatic metrics, such measures may not capture all subtle distortions introduced during translation. Despite these limitations, our work lays a framework for future research that can leverage standardized benchmarks and broader datasets to further validate and enhance the utility of LLMs in mental health applications.

Ethical Considerations

This work is entirely analytical and does not involve the collection of new data, the development of new models, or engaging directly with human subjects. All analyses are based solely on previously published and publicly available data. We adhere to the ethical guidelines outlined in the ACL Code of Ethics (https://www.aclweb.org/portal/content/acl-code-ethics), and we emphasize that any research in the mental health domain must be conducted with utmost sensitivity to privacy and ethical considerations. Although our study is retrospective in nature, we recognize the critical importance of safeguarding vulnerable populations, and we advocate for strict adherence to ethical standards in any practical applications derived from our findings.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive feedback, our collaborators for their valuable contributions, and the creators of the datasets used in our experiments for making these resources publicly available and enabling this research.

References

Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on Twitter. In *Proceedings of NAACL*.

Prottay Kumar Adhikary, Aseem Srivastava, Shivani Kumar, Salam Michael Singh, Puneet Manuja, Jini K Gopinath, Vijay Krishnan, Swati Kedia Gupta, Koushik Sinha Deb, and Tanmoy Chakraborty. 2024. Exploring the efficacy of large language models in summarizing mental health counseling sessions: benchmark study. *JMIR Mental Health*, 11:e57306.

Kabir Ahuja, Monojit Choudhury, and Sandipan Dandapat. 2022. On the economics of multilingual few-shot learning: Modeling the cost-performance trade-offs of machine translated and manual data. In *Proceedings of NAACL*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of ClinicalNLP*.

Anthropic. 2023. Claude: The anthropic ai language model. *Online documentation*. Available at: https://www.anthropic.com.

Nadiah A Baghdadi, Amer Malki, Hossam Magdy Balaha, Yousry AbdulAzeem, Mahmoud Badawy, and Mostafa Elhosseini. 2022. An optimized deep learning approach for suicide detection through arabic tweets. *PeerJ Computer Science*.

Outi Bat-El. 2019. Templatic morphology (clippings, word-and-pattern).

Panchanit Boonyarat, Di Jie Liew, and Yung-Chun Chang. 2024. Leveraging enhanced bert models for detecting suicidal ideation in thai social media content amidst covid-19. *Information Processing & Management*.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2025. A survey on multilingual mental disorders detection from social media data. *arXiv preprint arXiv:2505.15556*.

Richard Oliver Collin. 2010. Ethnologue. *Ethnopolitics*, 9(3-4):425–432.

Adonias C de Oliveira, Evandro JS Diniz, Silmar Teixeira, and Ariel S Teles. 2022. How can machine learning identify suicidal ideation from user's texts? towards the explanation of the boamente system. *Procedia Computer Science*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Amrita Ganguly, Sadiya Sayara Chowdhury Puspo, Md Nishat Raihan, Dhiman Goswami, Marcos Zampieri, et al. 2024. Masonperplexity at multimodal hate speech event detection 2024: Hate speech and target detection using transformer ensembles. In *Proceedings of CASE*.

Muskan Garg. 2024. Towards mental health analysis in social media for low-resourced languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–22.

Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Mozhgan Ghassemiazghandi. 2024. An evaluation of chatgpt's translation accuracy using bleu score. *Theory and Practice in Language Studies*, 14(4):985–994.

Dhiman Goswami, Md Nishat Raihan, Sadiya Sayara Chowdhury Puspo, and Marcos Zampieri. 2023. nlpbdpatriots at blp-2023 task 2: A transfer learning approach towards bangla sentiment analysis. In *Proceedings of BLP*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Mika Hämäläinen, Pattama Patpong, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Detecting depression in thai blog posts: a dataset and a baseline. In *Proceedings of WNUT*.

Martin Haspelmath. 2005. *The world atlas of language structures*. Oxford University Press.

AbdelMoniem Helmy, Radwa Nassar, and Nagy Ramdan. 2024. Depression detection for twitter users using sentiment analysis in english and arabic tweets. *Artificial intelligence in medicine*.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, et al. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of WWW*.

Rahul Kande, Vasudev Gohil, Matthew DeLorenzo, Chen Chen, and Jeyavijayan Rajendran. 2024. Llms for hardware security: Boon or bane? In *Proceedings of IEEE VTS*.

Pitso Walter Khoboko, Vukosi Marivate, and Joseph Sefara. 2025. Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, 20:100649.

Nathaniel Krasner, Justin Vasselli, Belu Ticona, Antonios Anastasopoulos, and Chi-kiu Lo. 2025. Machine translation metrics for indigenous languages using fine-tuned semantic embeddings. In *Proceedings of AmericasNLP*.

Gleb Kuzmin, Petr Strepetov, Maksim Stankevich, Artem Shelmanov, and Ivan Smirnov. 2024. Mental disorders detection in the era of large language models. *arXiv preprint arXiv:2410.07129*.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. In *Proceedings of CLPsych*.

Umasree Mariappan, D Balakrishnan, G Merline, M Sandhia, Dubba Saitej Reddy, and Sattineni Gagan Teja. 2024. Mental health disorder prediction using recurrent neural network algorithm. In *Proceedings of APCIT*.

John Mendonça, Alon Lavie, and Isabel Trancoso. 2023. Towards multilingual automatic dialogue evaluation. *arXiv preprint arXiv:2308.16795*.

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of CLPsych*.

Sergazy Narynov, Daniyar Mukhtarkhanuly, and Batyrkhan Omarov. 2020. Dataset of depressive posts in russian language collected from social media. *Data in brief*, 29:105195.

Thao Nguyen, Jeffrey Li, Sewoong Oh, Ludwig Schmidt, Jason E Weston, Luke Zettlemoyer, and Xian Li. 2024. Better alignment with instruction back-and-forth translation. In *Findings of EMNLP*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen Qiu, Dan Oneață, Emanuele Bugliarello, Stella Frank, and Desmond Elliott. 2022. Multilingual multimodal learning with machine translated text. In *Findings of EMNLP*.

Md Nishat Raihan, Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, and Marcos Zampieri. 2023. nlpbdpatriots at blp-2023 task 1: Two-step classification for violence inciting text detection in bangla-leveraging back-translation and multilinguality. In *Proceedings of BLP*.

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2025a. mhumaneval–a multilingual benchmark to evaluate large language models for code generation. In *Proceedings of NAACL*.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Mentalhelp: A multi-task dataset for mental health in social media. In *Proceedings of LREC-COLING*.

Nishat Raihan, Mohammed Latif Siddiq, Joanna Santos, and Marcos Zampieri. 2025b. Large language models in computer science education: A systematic literature review. In *Proceedings of SIGCSE*.

Konstantinos Skianis, A Seza Doğruöz, and John Pavlopoulos. 2024a. Leveraging llms for translating and classifying mental health data. *arXiv preprint arXiv:2410.12985*.

Konstantinos Skianis, John Pavlopoulos, and A Seza Doğruöz. 2024b. Building multilingual datasets for predicting mental health severity through llms: Prospects and challenges. *arXiv preprint arXiv:2409.17397*.

Konstantinos Skianis, John Pavlopoulos, and A Seza Dogruöz. 2024c. Severity prediction in mental health: Llm-based creation, analysis, evaluation of a novel multilingual dataset. *CoRR*.

Elsbeth Turcan and Kathleen Mckeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. In *Proceedings of LOUHI*.

Abdul Hasib Uddin, Durjoy Bapery, and Abu Shamim Mohammad Arif. 2019. Depression analysis of bangla social media data using gated recurrent neural network. In *Proceedings of IEEE ICASERT*.

Kid Valeriano, Alexia Condori-Larico, and Josè Sulla-Torres. 2020. Detection of suicidal intent in spanish language social networks using machine learning. *International Journal of Advanced Computer Science and Applications*.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of IMWUT*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of EMNLP*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of WWW*.

Noureldin Zahran, Aya E Fouda, Radwa J Hanafy, and Mohammed E Fouda. 2025. A comprehensive evaluation of large language models on mental illnesses in arabic context. *arXiv preprint arXiv:2501.06859*.

Rodolfo Joel Zevallos, Annika Marie Schoene, and John E Ortega. 2025. The first multilingual model for the detection of suicide texts. In *Proceedings of SUMEval*.