Large Language Models for Mental Health: A Multilingual Evaluation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities across various NLP tasks, but their performance in multilingual settings is often underexplored. This study evaluates proprietary and open-source LLMs on eight mental health datasets of various languages. We compare their performance in zero-shot, few-shot, and fine-tuned settings against traditional methods. Results show that LLMs achieve competitive or superior F1 scores across several datasets, with fine-tuned models often surpassing state-of-the-art results. However, performance varies across languages, highlighting both the strengths and limitations of LLMs in this critical application. These findings provide actionable insights into the application of LLMs for multilingual mental health text classification.

1 Introduction

While LLMs have been transforming research in NLP, caution must be exercised when adopting these models in sensitive domains such as mental health (Hua et al., 2024). Due to the potential risks and ethical considerations, experts are cautious about the integration of LLMs in applications in sensitive domains. These concerns are further amplified in multilingual settings where studies have demonstrated that LLMs tend to perform worse when prompted in languages other than English (Jin et al., 2024; Raihan et al., 2024a).

Most mental health datasets are curated from social media mining platforms such as Reddit and X. A recent survey by Kumar et al. (2024) shows that the clear majority of such datasets (Mariappan et al., 2024; Turcan and Mckeown, 2019; Raihan et al., 2024b) are in English. Recent efforts are being made to curate similar resources in other languages such as Arabic (Baghdadi et al., 2022; Helmy et al., 2024), Russian (Narynov et al., 2020), and Thai (Hämäläinen et al., 2021). However, none of the aforementioned studies evaluate the performance of LLMs on non-English datasets, leaving an important gap in understanding their performance in multilingual mental health settings.

A few recent studies explore the performance of LLMs in English mental health datasets. Xu et al. (2024) examines LLMs performance on multiple datasets compared to statistical and traditional encoder-only models (Devlin et al., 2019; Alsentzer et al., 2019). Similarly, Kuzmin et al. (2024), Yang et al. (2023), and Wei et al. (2022) investigate LLM performance exploring various prompting strategies. Finally, Yang et al. (2024) presents a fine-tuning approach with the release of MentaLLaMA, a task-specific model for the domain. While these approaches achieve competitive results, they are limited to English, leaving a significant gap in research in other languages.

To address this gap, we present the first multilingual evaluation of state-of-the-art LLMs on mental health datasets. We consider mental health datasets in six languages - Arabic, Bengali, Spanish, Portuguese, Russian, and Thai - and two tasks, namely depression and suicide ideation detection. We address the following Research Questions (RQs):

- **RQ1**: How does the performance of LLMs compare to previously proposed models (e.g., statistical, neural, BERT-based)?
- **RQ2**: What are the best prompting strategies for LLMs on mental health?
- **RQ3**: What is the impact of instruction fine-tuning on the performance of open-source LLMs?

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.

2 Datasets

For our study, we acquire eight publicly available datasets presented in Table 1. The datasets include posts in Russian from VKontakte and posts in Thai from online blogs that have been annotated for depression (Narynov et al., 2020; Hämäläinen et al., 2021). Additionally, we acquire multiple datasets containing data sourced from X, annotated for depression in Bengali (Uddin et al., 2019) and Arabic (Helmy et al., 2024). Furthermore, there are posts from X annotated for suicidal ideation in Thai (Boonyarat et al., 2024), Portuguese (de Oliveira et al., 2022), Arabic (Baghdadi et al., 2022) and Spanish (Valeriano et al., 2020).

3 Experiments and Results

We evaluate a diverse set of LLMs, encompassing both proprietary and open-source architectures. Our evaluation includes multiple prompting strategies and also fine-tuning open-source models to gather better insights.

3.1 LLMs

We evaluate seven state-of-the-art LLMs spanning both proprietary and open-source architectures, as listed in Table 2.

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

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

Our selection includes three proprietary models—GPT-4 Omni, Claude 3.5 Sonnet, and Gemini 2 Flash—as well as 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. These 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.

3.2 Prompting

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. For CoT prompting, we adopt the SOTA prompting method for the mental health tasks, $CoT_{\text{Emo\_FS}}$, introduced by Yang et al. (2023).

Table 3 presents a comprehensive 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 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

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).

(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.

3.3 Fine-tuning

Due to the intrinsic black box nature of proprietary models and their high costs, we wanted to explore models that could be fully-customization to this task. Therefore, we experiment with fine-tuning the open-source models. The hyperparameters are chosen empirically as we run a set of experiments with different combinations of parameters and report the best results. The final selection of hyper-parameters is presented in Appendix A.

Table 4 presents a comparative analysis of F1 scores before and after fine-tuning on eight multilingual depression and suicide 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.

4 Observation and Analysis

We now revisit the 3 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)?*

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 4: 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.

> *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 Chain-of-Thought (CoT) prompting is the most effective strategy for mental health applications, consistently yielding higher F1 scores compared to zero-shot and few-shot methods. This structured approach to prompting enhances reasoning capabilities, enabling LLMs to better extract nuanced signals from text data, which is critical in sensitive domains such as mental health.*

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.*

5 Conclusion and Future Work

This work represents the first investigation of LLMs in the multilingual mental health domain. 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 our results demonstrate considerable promise, the variability in performance across languages and models indicates that further research is required to optimize these techniques for sensitive mental health applications.

Overall, this study lays an important foundation for future efforts aimed at refining LLM-based methodologies in complex, multilingual settings. In future work, we would like to include more tasks and languages to broaden our understanding and gain more insights. Additionally, we plan to adapt open-source models to the domain with methods like Continual Pretrainining and Synthetic Fine-tuning to potentially increase their performance.

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. These limitations notwithstanding, our work lays a foundational 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 engagement 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.

References

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 the 2nd Clinical Natural Language Processing Workshop*.

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*.

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*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

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

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 the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*.

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 the ACM on Web Conference 2024*.

Puneet Kumar, Alexander Vedernikov, and Xiaobai Li. 2024. Measuring non-typical emotions for mental health: A survey of computational approaches. *arXiv preprint arXiv:2403.08824*.

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*.

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 *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*.

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

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

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2024a. mhumaneval–a multilingual benchmark to evaluate large language models for code generation. *arXiv preprint arXiv:2410.15037*.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024b. Mentalhelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Elsbeth Turcan and Kathleen Mckeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*.

Abdul Hasib Uddin, Durjoy Bapery, and Abu Shamim Mohammad Arif. 2019. Depression analysis of bangla social media data using gated recurrent neural network. In *2019 1st International conference on advances in science, engineering and robotics technology (ICASERT)*, pages 1–6. IEEE.

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*.

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. *Advances in neural information processing systems*.

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 the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*.

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 the 2023 Conference on Empirical Methods in Natural Language Processing*.

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 the ACM on Web Conference 2024*.

A Experimental Details

The fine-tuning stage is performed on a single NVIDIA A100 GPU with 40 GB of memory, accessed via Google Colab[1]. The system is further equipped with 80 GB of RAM and 256 GB of disk storage to support computational efficiency.

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

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

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