STM JOURNALS

Recent Trends in
Programming Languages

ISSN: 2455-1821
Volume 12, Issue 3, 2025
September–December
DOI (Journal): 10.37591/RTPL

https://journals.stmjournals.com/rtpl

Research RTPL

MentaLLaMA: Advancing Mental Health Insights with Instruction-Finetuned Large Language Models

Himanshu Kumar Singh[1,*], Bhanu Prakash Lohani[2]

Abstract

The growing prevalence of mental health challenges in contemporary society has highlighted the urgent need for advanced, interpretable, and reliable artificial intelligence solutions that can support mental health assessment and intervention. In response to this need, this research introduces a novel collection of open-source, instruction-tuned large language models (LLMs) specifically designed to facilitate transparent and accurate mental health evaluations. Leveraging a newly developed dataset, which integrates multiple tasks and diverse sources and contains over 105,000 instances, the models were fine-tuned to deliver precise predictions while providing human-level explanations for their outputs across a variety of mental health-related tasks. Experimental results demonstrate that these models achieve high performance in terms of accuracy and consistency, while also displaying notable adaptability when applied to previously unseen tasks. Additionally, this study investigates both single-dataset and multi-dataset fine-tuning strategies to optimize LLMs for domain-specific applications, ensuring they outperform existing approaches in both efficiency and generalization. Despite these promising outcomes, the models are not without limitations, including gaps in domain-specific knowledge and the potential for biased outputs, underscoring the importance of ongoing pre-training and the development of more robust evaluation metrics for future research.

Keywords: Mental health analysis, large language models (LLMs), instruction tuning, multi-task learning, fine-tuning strategies, generalizability in AI, domain-specific AI models, open-source mental health tools

INTRODUCTION

Recent progress in Large Language Models (LLMs) such as GPT-4, PaLM, FLAN-T5, and Alpaca has demonstrated their capability to carry out a diverse range of tasks in zero-shot settings, ranging from question answering and logical reasoning to machine translation. With their vast parameter scales, these models exhibit an emergent capacity to interpret natural language, reason effectively, and even infer human common sense [1]. This has sparked interest in their potential applications across various domains, including mental health.

Mental health challenges represent a critical global concern, affecting people and societies on a significant scale. Data infers that over 24% of adults in the US experience mental health disorders, and 5.6% grapple with severe conditions that impair daily functioning. Furthermore, anxiety and depression collectively result in annual productivity losses of approximately $1 trillion worldwide. Given that natural language is central to mental health assessment and therapy, LLMs hold promise as tools for understanding mental states, detecting disorders, and assisting in interventions [2, 3].

***Author for Correspondence**
Himanshu Kumar Singh
E-mail: hksingh093@gmail.com

[1]Student, Department of Computer Science and Engineering, Amity School of Engineering & Technology Amity University, Greater Noida, Uttar Pradesh, India
[2]Assistant Professor, Department of Computer Science and Engineering, Amity School of Engineering & Technology Amity University, Greater Noida, Uttar Pradesh, India

Received Date: June 19, 2025
Accepted Date: July 14, 2025
Published Date: October 17, 2025

Citation: Himanshu Kumar Singh, Bhanu Prakash Lohani. MentaLLaMA: Advancing Mental Health Insights with Instruction-Finetuned Large Language Models. Recent Trends in Programming Languages. 2025; 12(3): 8–15p.

© STM Journals 2025. All Rights Reserved

Historically, research in the field of natural language processing (NLP) for mental health, there has been an emphasis on creating models that are specific to the domain, designed for tasks like detecting stress, predicting depression, or assessing the risk of suicide. However, these approaches require extensive fine-tuning for each task and lack flexibility. Similarly, traditional chatbots for mental health support, often rule-based, fail to leverage the sophisticated reasoning capabilities of modern LLMs. Despite these advancements, closed-source LLMs like ChatGPT still underperform supervised methods in mental health classification tasks, effectively in zero-shot or few-shot scenarios. This limitation often results in inaccurate reasoning and low-quality explanations, hindering their reliability for mental health applications [4].

To address these gaps, fine-tuning LLMs with task-specific data has emerged as an effective solution. However, this approach presents two major challenges. First, high-quality annotated datasets for mental health analysis remain scarce due to the sensitive nature of the topic and the high costs associated with obtaining expert-written explanations. Second, fine-tuning closed-source models like ChatGPT is expensive, time-intensive, and environmentally taxing, underscoring the need for accessible, open-source alternatives tailored to mental health tasks.

To overcome these barriers, researchers have begun modeling interpreting mental health assessments as a text generation challenge. This method not only identifies signs of mental health issues from social media content, but also generates human-like explanations for predictions. Leveraging instruction-tuned datasets like the Interpretable Mental Health Instruction (IMHI) dataset, which combines multi-task and multi-source annotations, has shown promise in bridging the gap between LLMs and domain-specific models [5, 6].

This effort has culminated in the development of MentaLLaMA which is the initial series of open-source large language models specially fine-tuned for analysis in mental health. Based on LLaMA2 foundation models, MentaLLaMA reaches top-tier performance in mental health detection while generating high-quality, human-like explanations.

LLMS IN MENTAL HEALTH

Large Language Models (LLMs) signify a significant leap in machine learning, showcasing expertise to generate and understand human-like text with remarkable precision. Their performance is typically assessed using specialized benchmarks that evaluate linguistic accuracy and contextual relevance. These measures include the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) for summarization tasks, and the Bilingual Evaluation Understudy (BLEU) for translation accuracy.

This progress is primarily attributed to the transformer architecture, which leverages a "self-attention" mechanism. This design enables parallel processing of information rather than sequential processing, resulting in enhanced speed and deeper contextual comprehension [7].

To qualify as an LLM, a model must utilize transformer-based architecture and typically comprise at least one billion parameters, categorizing models like GPT (from OpenAI) and BERT (from Google AI) which are prominent models in this area. Even though the typical BERT model has just 0.34 billion parameters, and falls short of the traditional "large" classification, its innovative bidirectional structure and role in developing new standards in (NLP) justifies its ranking among significant LLMs [8, 9].

The release of ChatGPT by OpenAI in 2022 marked a pivotal moment, drawing substantial attention from both academic and public spheres to the transformative capabilities of LLMs. Additional cutting-edge models consist of the Large Language Model Meta AI (LLaMA) developed by Meta AI and the Pathways Language Model (PaLM) created by Google AI. Together, these advancements underscore the ongoing evolution and expanding impact of LLMs within artificial intelligence research and applications (Table 1) [10].

© STM Journals 2025. All Rights Reserved

Table 1. Comparative analysis of large language models (LLMs) by parameter size and developer entity.

LLMs	Parameter Size (billion)	Developer
Generative Pretrained Transformer (GPT)	0.127	OpenAI
Bidirectional Encoder Representations from Transformers (BERT)	0.34	Google
PaLM 2	340	LAION
Meta (LLAMA)	65	Meta

METHODS

To advance interpretable mental health analysis, we developed a comprehensive experimental framework utilizing large language models (LLMs) across various mental health prediction tasks. This effort began with the collection of raw data from 10 diverse sources, covering eight distinct mental health challenges. The dataset includes annotated social media posts tailored to address these issues. Building on the effectiveness of approaches such as self-instruct and the ability of ChatGPT to provide human-like explanations, we utilized expertly designed few-shot Samples and collected annotations to guide ChatGPT in generating well-crafted explanations.

Experimental Framework for Mental Health Prediction

Our approach encompasses three key methods: zero-shot prompting, few-shot prompting, and instruction fine-tuning. These methodologies are designed to be model-agnostic, enabling their application across different LLM architectures. Below, we provide an overview of these approaches and the design of the experimental setups.

Few-shot Prompting

Few-shot prompting enhances domain-specific learning by incorporating a limited number of examples into the prompts. Unlike fine-tuning, this approach keeps the model parameters unchanged, instead relying on in-context learning from the examples provided [11]. Researchers have successfully applied this technique in specialized domains. In our experiments, we used additional randomly sampled prompt-label pairs (denoted as $Prompt_{ZS}$) to provide contextual guidance, enabling the model to dynamically acquire domain-specific knowledge without modifying its architecture.

$$Prompt_{ZS} = TextData + Prompt_{Part1\text{-}S} + Prompt_{Part2\text{-}Q} + OutputConstraint \quad (1)$$

Zero-shot Prompting

LLMs' robust language understanding and reasoning capabilities allow them to perform diverse tasks in a zero-shot framework, there is no need for training data specific to a particular domain [12, 13]. We created a universal prompt template for zero-shot applications (Prompt) for mental health tasks, comprising four components:

$$Prompt_{FS} = \{Sample\ Prompt_{ZS} - label\}_M + Prompt_{ZS} \quad (2)$$

Instruction Fine-tuning

Unlike the few-shot prompting strategy, this approach aligns more closely with traditional few-shot transfer learning. Here, the model is further trained using a small amount of domain-specific data. Multiple fine-tuning strategies are explored to enhance performance.

Single-Dataset Fine-tuning

This approach involves basic fine-tuning on a single dataset, typically the training set, as seen in prior work on mental health. The fine-tuned model's performance is assessed both on the same dataset's test set and on entirely different datasets to evaluate its generalizability.

Multi-Dataset Fine-tuning

We extend the scope by fine-tuning models across multiple datasets simultaneously. This is achieved using instruction fine-tuning, allowing large language models (LLMs) to address various tasks across different datasets [14].

© STM Journals 2025. All Rights Reserved

This method stands apart advanced models tailored for mental health (like Mental-RoBERTa), which are generally optimized for a specific task, such as identifying depression or predicting suicidal thoughts. After being fine-tuned for Task A, these models become task-specific and lack adaptability for other tasks.

In contrast, our approach trains LLMs on multiple mental health datasets concurrently. Each dataset contains unique instructions corresponding to its specific tasks. By incorporating diverse task instructions in a single training iteration, the model is equipped to handle multiple tasks without requiring additional fine-tuning for each specific task.

Fine-Tuning Workflow

For fine-tuning with either a single dataset or multiple datasets (Figure 1), the approach consists of the same two essential steps:

Step 1: Finetune with $\{\text{Prompt}_{zs} - \text{label} \sum \text{N}_{\text{Di-train}'}\}$

Step 2: Test with $[\text{Prompt}_{zs}] \sum \text{N}_{\text{Di-train}'}$ (3)

Evaluating the fine-tuned model's performance across tasks and datasets to assess its generalizability and effectiveness (Table 2).

Data Preparation and the IMHI Dataset

To ensure reliability, we conducted automated evaluations assessing prediction accuracy, annotation-explanation alignment, and explanation quality. Human assessments of a portion of the data were conducted using an annotation framework designed by domain experts [15].

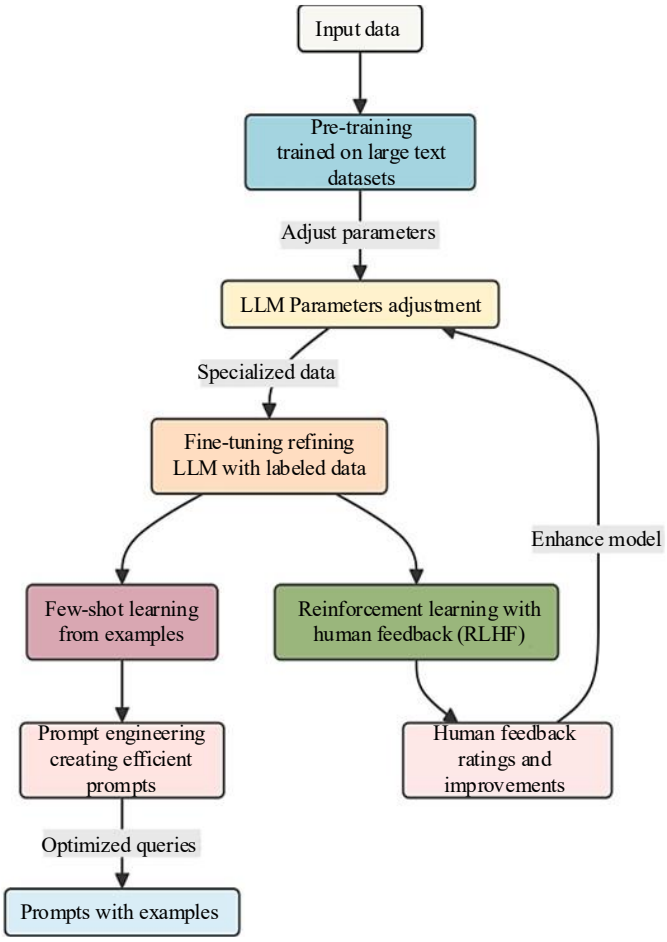

Figure 1. Flowchart for finetuning processes.

© STM Journals 2025. All Rights Reserved

Table 2. Correction evaluation findings on generalizability.

Model	CAMS	Dreaddit	IRF	T-SID
LLaMA2-13B$_{ZS}$	18.76	35.96	38.76	26.48
ChatGPT$_{ZS}$	32.69	73.89	57.37	37.36
MentaLLama-chat-7B	27.76	68.76	54.9	65.87
MentaLLama-chat-13B	36.29	74.54	67.87	75.87

The collected posts, annotations, and explanations were then transformed into instruction-based query-answer pairs using a rule-driven process. This effort resulted in the Interpretable Mental Health Instruction. The IMHI dataset is the inaugural resource that supports multiple tasks and sources for the training and evaluation of interpretable analysis in mental health [1].

Introduction of MentaLLaMA

Using the IMHI dataset, we created MentaLLaMA, an open-source series of language models built on the LLaMA2 foundation models, specifically designed for transparent mental health assessment. This series consists of three models of different sizes: MentaLLaMA-7B, MentaLLaMA-chat-7B, and MentaLLaMA-chat-13B. These models were evaluated on the IMHI benchmark, focusing on prediction accuracy and explanation quality.

Key Findings

1. *Accuracy and Generalization:* MentaLLaMA-chat-13B matched or outperformed state-of-the-art (SOTA) discriminative approaches in 7 out of 10 test sets. It also demonstrated superior generalization to unseen tasks.
2. *Explanation Quality:* Explanations generated by MentaLLaMA were at par with ChatGPT and consistently outperformed other generative pre-trained language models (PLMs).
3. *Model Enhancements:* Instruction fine-tuning, reinforcement learning driven by human feedback (RLHF), and increasing model sizes were pivotal in achieving these results.

Key Contributions

1. *IMHI Dataset:* This is the initial dataset designed for multi-task and multi-source instruction fine-tuning, designed to deliver insightful analysis of mental health through social media data.
2. *MentaLLaMA Models:* A pioneering open-source series of language models designed for understandable mental health applications, which includes instruction adherence capabilities.
3. *Holistic Evaluation Benchmark:* A robust benchmark with 19,000 test samples across eight tasks and ten test sets.
4. *Superior Performance:* MentaLLaMA models surpassed existing generative PLMs and even outperformed ChatGPT on several key metrics.

IMPLEMENTATION

Using the IMHI dataset, we refined LLaMA2 models to develop the MentaLLaMA series. We initiated this process by training the LLaMA2-7B model on the dataset for 10 epochs to produce MentaLLaMA-7B. The model that yielded the best performance was chosen based on the validation outcomes from the IMHI validation set. Important training parameters included a batch size of 32 and a gradient accumulation step of 7, leading to an effective batch size of 256, and the use of the AdamW optimizer. The maximum learning rate was set to 1e-5, with a warm-up ratio of 4%, and the input length was capped at 2096 tokens. To enhance training efficiency, Flash-Attention was employed.

We conducted fine-tuning on the LLaMA2-chat-7B and LLaMA2-chat-13B models to create the MentaLLaMA-chat-7B and MentaLLaMA-chat-13B versions, employing instruction tuning and RLHF. These models, tuned with the same IMHI dataset and experimental parameters, represent the LLMs in this domain with such optimization techniques.

© STM Journals 2025. All Rights Reserved

Experimentation and Analysis

In order to assess the effectiveness of our MentaLLaMA models, we conducted comparisons with a range of different established baseline models across different categories:

Discriminative Methods

Mental health analysis has traditionally been approached as a text classification task. For comparison, we included fine-tuned discriminative PLMs such as BERT, RoBERTa, and domain-specific models like MentalBERT and MentalRoBERTa. These domain-specific models are initially trained on extensive mental health datasets and subsequently refined using specific target datasets. While these models demonstrate state-of-the-art (SOTA) performance in 9 out of 10 test sets, their limitations lie in weak generalization and lack of interpretability, as their decisions are task-specific and not easily explainable [16].

Zero-shot and Few-shot Methods

With advancements in large language models (LLMs), zero-shot and few-shot approaches have become cost-effective solutions. We tested the 7B and 13B variants of LLaMA2, alongside proprietary models like ChatGPT and GPT-4, on benchmark data. ChatGPT significantly outperformed LLaMA2 models in zero-shot settings due to its emergent capabilities and scale. Few-shot prompting further improved performance for both ChatGPT and GPT-4, highlighting the effectiveness of in-context learning. However, GPT-4 demonstrated only marginal improvements over ChatGPT in most datasets.

Completion-based Fine-tuning Methods

Generative PLMs such as BART-large and T5-large were fine-tuned datasets to assess parameter efficiency. Surprisingly, these smaller models outperformed LLaMA2-7B on various test sets despite being only 18% of its size. This indicates that LLaMA2's performance is hindered by the unnatural format of the IMHI-completion dataset.

Key Observations

Fine-tuning with domain-specific instruction tuning (as in MentaLLaMA-7B) is more effective than completion-based methods, as MentaLLaMA-7B outperformed LLaMA2-7B on 8 out of 10 test sets. Enhanced models like MentaLLaMA-chat-7B and MentaLLaMA-chat-13B (fine-tuned with high-quality instructions) further improved performance, outperforming MentaLLaMA-7B on 9 out of 10 test sets. MentalRoBERTa demonstrates its nearly state-of-the-art performance in 9 out of 10 test sets.

Explanation Quality

To evaluate the quality of explanations, BART-score metrics were used. Completion-based fine-tuning improved explanation quality for LLaMA2-7B, but BART-large models demonstrated similar capabilities at a lower computational cost. While LLaMA2-7B slightly outperformed BART-large on most test sets, the margin of improvement was minimal [17].

EVALUATION AND ANALYSIS

Large language models (LLMs) are recognized not only for their exceptional text generation capabilities but also for their remarkable ability to generalize across unseen tasks. To assess the generalizability of MentaLLaMA, a modified training set, IMHI-general, was created by removing data from four specific tasks: stress detection, identification of mental disorders on Twitter (T-SID), the detection of causes related to depression and suicide (CAMS), and the identification of interpersonal risk factors (IRF) were conducted. Utilizing this updated dataset, the T5, BART, and MentaLLaMA-chat models underwent re-finetuning and were assessed against the test sets of these omitted tasks [10].

Correctness Evaluation

The performance results for correctness, demonstrate that MentaLLaMA models demonstrate a substantial advantage over LLaMA2-13B across all tested datasets. This underscores the model's

© STM Journals 2025. All Rights Reserved

capability to generalize to novel mental health analysis tasks. Furthermore, our models surpass ChatGPT on three datasets, illustrating their competitive strength in adjusting to new tasks.

Explanation Quality

The quality of explanations was assessed using the BART-score. On tasks like Dreaddit and CAMS, MentaLLaMA-chat models significantly surpassed T5 and BART, indicating their ability to produce higher-quality explanations for foundational mental health tasks. Their strong performance on IRF further demonstrates their deep understanding of complex mental health factors.

Interestingly, even when all Twitter-related data was excluded from training, MentaLLaMA-chat models performed better on the Twitter-derived test set T-SID, demonstrating their capacity to generalize effectively to data sources with varying characteristics. Furthermore, the larger MentaLLaMA-chat-13B model achieved higher explanation quality than its smaller counterpart, highlighting the benefits of scaling model size.

CONCLUSION AND FUTURE DIRECTION

This study presents a novel task focused on understandable mental health assessment and introduces the IMHI dataset, which is characterized by its multi-task and multi-source nature resource consisting of 106K data points, specifically designed for instruction tuning. To generate high-quality training data, we leverage ChatGPT and conduct rigorous evaluations using both automated and human assessments to guarantee the reliability of the data. We introduce MentaLLaMA, the inaugural open-source collection of large language models (LLMs) designed for interpretable mental health evaluation, concentrating on instruction-following capabilities. Results from evaluations on the IMHI benchmark show that MentaLLaMA is nearing (SOTA) performance in correctness while generating explanations at a human level. Additionally, MentaLLaMA demonstrates strong generalization abilities when faced with previously unseen tasks.

Despite these advancements, we observed that MentaLLaMA currently does not possess some specialized knowledge that more advanced models like ChatGPT have. To rectify this, upcoming efforts will aim at ongoing pre-training of MentaLLaMA with extensive and high-quality datasets to further enhance the professionalism and depth. Furthermore, the BART-score, used in this study as an automatic evaluation metric, shows only moderate correlation with human assessments, which restricts the dependability of our findings. Future research will aim to develop and incorporate more robust and reliable automated evaluation methods.

ETHICAL CONSIDERATION

The IMHI dataset was developed using raw data sourced from public social media platforms while adhering to stringent privacy regulations and ethical guidelines to ensure the protection of user privacy. All mental health-related content has been thoroughly anonymized. To mitigate potential misuse, the examples presented in this work have been paraphrased and obfuscated using a moderate disguising technique.

Despite the promising performance exhibited by MentaLLaMA, it is crucial to clarify that the predictions and explanations generated by the model are strictly intended for research purposes in non-clinical settings. Individuals in need of mental health support should seek assistance from licensed psychiatrists or clinical professionals.

Moreover, recent studies have identified inherent biases in large language models (LLMs), such as gender-based disparities. Issues such as inaccurate predictions, unsuitable explanations, and overgeneralization highlight the existing risks associated with current LLMs. These challenges underline the importance of further research to refine and responsibly apply LLMs in practical mental health monitoring scenarios.

© STM Journals 2025. All Rights Reserved

REFERENCES

1. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. arXiv preprint arXiv:2205.12689. 2022 May 25.
2. Markov AA. On a Remarkable Case of Samples Connected in a Chain. Appendix on the statistical investigation of a text by Aksakov. Sci Context. 2006 Dec; 19(4): 601–4.
3. Pakshina NA. Aleksandr Lyapunov: remembered by his contemporaries. IFAC-PapersOnLine. 2017 Jul 1; 50(1): 5208–18.
4. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res. 2021 Jan 13; 23(1): e17828.
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. Adv Neural Inf Process Syst. 2017 Dec 4; 30(1): 261–72.
6. Bommasani R. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. 2021.
7. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Yearb Med Inform. 1993; 2(01): 41–51.
8. Zhang Y, Sun S, Galley M, Chen YC, Brockett C, Gao X, Gao J, Liu J, Dolan B. Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911. 00536. 2019 Nov 1.
9. Liu H. Towards automated psychotherapy via language modeling. arXiv preprint arXiv:2104.10661. 2021 Apr 5.
10. Gillblad D. Language Models for Everyone—Responsible and Transparent Development of Open Large Language Models. Comput Sci Math Forum (MDPI). 2023 Sep 7; 8(1): 51.
11. Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. Nat Mach Intell. 2021 Jun; 3(6): 461–3.
12. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, Presser S. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027. 2020 Dec 31.
13. Devlin J, Chang MW, Lee K, Toutanova K. 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: human language technologies, volume 1 (long and short papers). 2019 Jun; 4171–4186.
14. Kenton JD, Toutanova LK. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT. 2019 Jun 2; 1(2).
15. Abid A, Farooqi M, Zou J. Persistent anti-Muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021 Jul 21; 298–306.
16. Das A, Selek S, Warner AR, Zuo X, Hu Y, Keloth VK, Li J, Zheng WJ, Xu H. Conversational bots for psychotherapy: a study of generative transformer models using domain-specific dialogues. In Proceedings of the 21st workshop on biomedical language processing. 2022 May; 285–297.
17. Ahmed A, Aziz S, Toro CT, Alzubaidi M, Irshaidat S, Serhan HA, Abd-Alrazaq AA, Househ M. Machine learning models to detect anxiety and depression through social media: A scoping review. Comput Methods Programs Biomed Update. 2022 Jan 1; 2: 100066.

© STM Journals 2025. All Rights Reserved