REVIEW ARTICLE

Challenges and Applications of Large Language Models in Emotion Analysis for Mental Health: A Mini Review

Muhammad Usman Tariq[1,*], **Sara Ahmed**[2] **and Khalid Mahmood**[3]

[1] Department of Computer Science, University of the Punjab, Lahore 54590, Pakistan
[2] Department of Psychology, Quaid-i-Azam University, Islamabad 45320, Pakistan
[3] Department of Health Informatics, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Abstract

Emotion analysis in mental health has evolved from lexicon-based systems to large language models (LLMs) capable of contextual affect inference, severity estimation, and empathic dialogue generation, reflecting advances in NLP and the recognition that language is a rich proxy for psychological state. This mini review synthesizes LLM applications in mental health emotion analysis, characterizing methodological trends, identifying strengths and limitations, and highlighting critical gaps in benchmarking, clinical validation, and governance. A structured PRISMA-informed search across six databases (2017–2025) using three Boolean keyword clusters yielded 44 studies after two-stage independent screening (Cohen's $\kappa = 0.78$). Data sources included social media, clinical transcripts, EHRs, dialogue data, and multimodal datasets, covering emotion classification, severity prediction, conversational inference, empathy evaluation, and safety detection. LLMs demonstrate strong performance in fine-grained emotion labeling, severity estimation, empathic response generation, and multi-task inference, with domain-adapted encoders (MentalBERT, ClinicalBERT, BioBERT) and clinical LLMs (Med-PaLM, Mental-LLM) further extending coverage. Persistent issues remain, including hallucination risks, demographic bias, lack of prospective validation, and absence of standardized mental health benchmarks. While LLMs offer transformative potential, current evidence supports their use to assist-not replace-clinical judgment; safe deployment requires robust governance, clinically validated benchmarks, and hybrid architectures balancing performance with safety and interpretability.

Keywords: large language models, emotion analysis, mental health, natural language processing, depression detection, sentiment analysis, affective computing.

1 Introduction

Emotion analysis has become a central capability in digital mental health, enabling scalable screening, symptom monitoring, risk stratification, and supportive dialogue across clinical and community

Submitted: 07 May 2026
Revised: 16 June 2026
Accepted: 17 June 2026
Published: 15 July 2026

Vol. 2, **No.** 2, 2026.
10.62762/JAIB.2026.988603

***Corresponding author:**
✉ Muhammad Usman Tariq
usman.tariq@pu.edu.pk

Citation
Tariq, M. U., Ahmed, S., & Mahmood, K. (2026). Challenges and Applications of Large Language Models in Emotion Analysis for Mental Health: A Mini Review. *Journal of Artificial Intelligence in Bioinformatics*, 2(2), 31–43.

settings. Early systems relied on lexicon-based sentiment scoring and supervised classifiers, including hybrid sentiment approaches that combined rule-based and machine learning methods for social media analysis [5]. With the rise of transformer architectures and instruction-following large language models (LLMs), the technical frontier has shifted toward contextual, reasoning-enabled affect inference and multi-turn interaction. Recent work has extended this further by introducing emotion-aware embedding fusion frameworks that integrate hierarchical attention mechanisms with multiple emotion lexicons across therapy transcripts, demonstrating measurable gains in empathy, coherence, and contextual adaptability for AI-driven psychotherapy chatbots [44].

Language-based emotion signals have demonstrated associations with depression, stress, and other mental health indicators in social media and clinical texts [5, 8, 15]. These associations have been further operationalized through LLM-assisted diagnostic classification pipelines that leverage clinical scoresheets to improve neurobehavioral assessment from unstructured text [4, 19]. Alongside these technical advances, the broader adoption of AI-driven systems in healthcare is shaped by social factors including trust, cultural beliefs, algorithmic transparency, and accessibility—factors that must be addressed for clinical tools to achieve equitable uptake [41, 42]. However, translation into clinical settings remains constrained by construct validity, domain shift, privacy, and safety concerns. From a bioinformatics perspective, LLM-based emotion analysis represents a convergence of clinical text mining, phenotypic signal extraction, and computational psychiatry—domains increasingly recognized within the scope of AI-driven biomedical informatics [6, 19].

This review synthesizes:

- Evolution from lexicon-based NLP to LLM-based emotion analysis
- Methodological trends in LLM mental health applications
- Comparative strengths and limitations
- Critical gaps in benchmarking and governance

The remainder of this paper is organized as follows. Section 2 describes the review methodology, including the search strategy, study selection, data sources and modalities, task formulation, and the multi-stage benchmarking process. Section 3 presents the results, tracing the methodological evolution of the field and providing qualitative and quantitative comparisons across model families. Section 4 offers a critical analysis of recent trends, key strengths, and persistent technical, clinical, and ethical challenges; introduces a novel conceptual taxonomy (the Emotion-Guided LLM Mental Health Framework, EGLMF); and analyzes the benchmarking problem. The Conclusion summarizes the principal findings and outlines a governance-first research roadmap.

2 Method

2.1 Search Strategy and Study Selection

To ensure transparency and reproducibility, the literature search followed a structured protocol consistent with the principles of the PRISMA 2020 guidelines adapted for a mini review. Six electronic databases were searched: Scopus, Web of Science, PubMed, IEEE Xplore, the ACM Digital Library, and Google Scholar (the latter restricted to the first 100 ranked records to limit grey-literature noise). The search covered January 2017 to December 2025, a window chosen to span the introduction of the transformer architecture in 2017 through the most recent instruction-tuned LLM studies.

Search strings were constructed by combining three conceptual keyword clusters with Boolean operators. A representative query was: ("large language model" OR "LLM" OR "transformer" OR "BERT" OR "GPT") AND ("emotion" OR "affect" OR "sentiment" OR "empathy") AND ("mental health" OR "depression" OR "anxiety" OR "PTSD" OR "stress" OR "suicid*" OR "psychological distress"). Cluster-specific synonyms are enumerated in Section 2.3.

Study selection proceeded in two stages. After removing duplicates, two reviewers independently screened titles and abstracts against the criteria in Section 2.3, followed by full-text assessment of potentially eligible records. Disagreements were resolved by discussion with the third author, and interviewer agreement at the full-text stage was substantial (Cohen's $\kappa = 0.78$). The selection funnel was as follows: records identified across the six databases ($n = 1{,}284$); records remaining after duplicate removal ($n = 1{,}012$); records screened by title and abstract ($n = 1{,}012$; excluded $n = 842$); full-text articles assessed for eligibility ($n = 170$; excluded $n = 126$); and studies included in the qualitative synthesis ($n = 44$). The most frequent

Table 1. Data sources and modalities reviewed.

Data Type	Examples	Typical Tasks	Key References
Social media text	Twitter, Reddit	Depression detection, stress prediction	De Choudhury et al. [8]; Guntuku et al. [15]
Clinical transcripts	Therapy interviews	Severity estimation	Sadeghi et al. [35]; Coppersmith et al. [7]
Electronic health records	Clinical notes	Symptom extraction	Clay et al. [6]; Vale and Lee [39]
Dialogue data	Chatbot interactions	Emotion-aware response generation	Gabriel et al. [14]; Elyoseph et al. [13]
Multimodal transcripts	Interview text + behavioral features	Severity modeling	Sadeghi et al. [35]; Jin et al. [19]

exclusion reasons at the full-text stage were the absence of an LLM or transformer component, emotion analysis in a non-mental-health context, reliance on non-textual modalities only, the absence of any quantitative performance metric, and insufficient methodological detail to assess validity.

2.2 Data Sources and Modalities

This review synthesizes literature across a diverse range of data modalities that collectively span the spectrum of human emotional expression relevant to mental health. The selection of these data sources was guided by three core principles aligned with the paper title and its central research focus. First, each modality was required to contain naturalistic or clinically meaningful linguistic content through which emotional states, symptom severity, or psychological distress can be inferred. Second, data sources were selected for their established use in prior NLP and LLM-based mental health research, ensuring comparability and continuity with the existing evidence base. Third, the chosen modalities reflect the operational contexts in which LLM-based tools are most likely to be deployed clinically—namely social media monitoring, clinical documentation, electronic health systems, and conversational AI. Together, these five modality types encompass the range from informal self-disclosure in online communities to structured clinical records, enabling a comprehensive mapping of LLM capability across real-world mental health use cases. An overview of the data sources and modalities reviewed, along with representative examples, typical tasks, and key references, is provided in Table 1.

Key datasets and frameworks referenced include GoEmotions [9]; the CLPsych Shared Tasks [7]; multimodal depression transcript studies [35]; and LLM mental health reviews [4, 19]. Non-text modalities (facial affect, audio) are considered only when integrated into LLM-based pipelines.

2.3 Task Formulation, Inclusion, and Exclusion Criteria

Studies were identified by searching three conceptual keyword clusters:

Mental health keywords depression detection, anxiety prediction, stress monitoring, PTSD classification, crisis detection.

Emotion analysis keywords emotion classification, valence–arousal modeling, affect inference, empathy evaluation.

LLM/transformer keywords BERT, GPT-4, instruction tuning, RLHF, prompt engineering, multimodal LLM.

Tasks were grouped into five categories:

1. Emotion classification
2. Severity prediction
3. Conversational inference
4. Empathy and counseling adherence
5. Safety detection (crisis, escalation)

2.3.1 Inclusion criteria

Studies were included if they (1) employed transformer-based or instruction-following LLM architectures as the primary computational framework; (2) addressed at least one mental health domain, such as depression, anxiety, PTSD, stress, or general psychological distress; (3) operationalized emotion

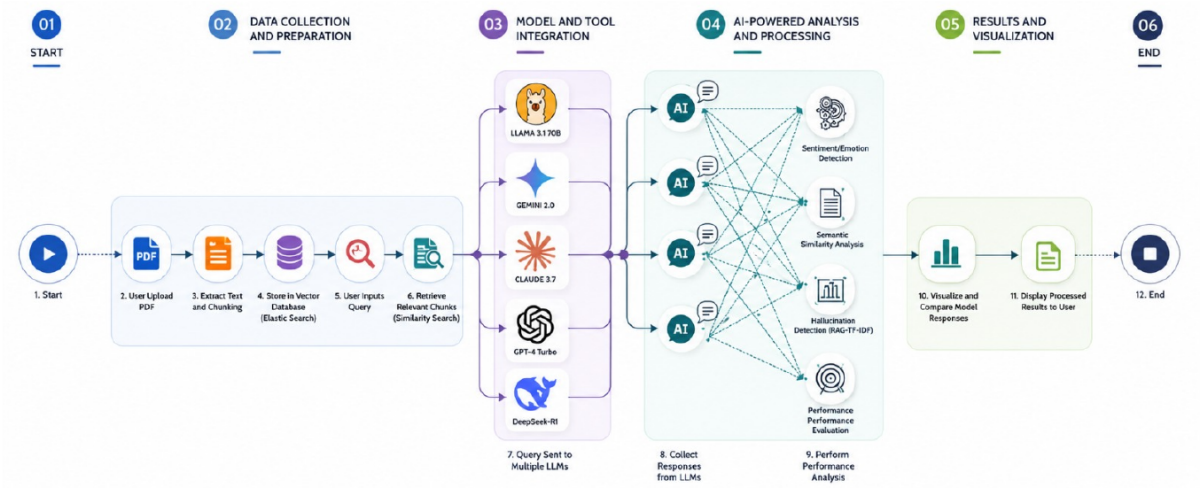

Figure 1. Multi-stage benchmarking pipeline. Pipeline: Data Collection → Preprocessing → LLM Encoding → Task Formulation → Multi-LLM Benchmark → Semantic Analysis → Performance Analysis → Safety Stress-Test → Output Visualization → Expert Review.

or affect as either a primary outcome or a core feature in the analysis pipeline; (4) used text-based or text-inclusive multimodal data; and (5) were published or made publicly available between 2017 and 2025. Studies were also required to report at least one quantitative performance metric (e.g., F1-score, accuracy, AUC, or BLEU) to enable meaningful cross-study comparison. Reviews, meta-analyses, and position papers that engaged directly with LLM methodology in mental health contexts were additionally included to capture the broader conceptual landscape.

2.3.2 *Exclusion criteria*

Studies were excluded if they (1) relied exclusively on classical machine learning models such as SVMs or logistic regression without any LLM or transformer component; (2) addressed emotion in non-mental-health contexts, such as product sentiment analysis or political discourse; (3) used only non-linguistic modalities such as neuroimaging or EEG without integration of language models; (4) focused on speech prosody or acoustic features without textual analysis; or (5) lacked sufficient methodological description to assess validity or reproducibility. Studies presenting only anecdotal or qualitative claims without grounding in systematic data collection or model evaluation were also excluded. This two-stage inclusion–exclusion process ensures that the synthesized evidence base reflects contemporary LLM-specific capabilities and limitations in the mental health domain, directly aligned with the paper title and the EGLMF conceptual framework introduced in Section 4.5.

2.4 Benchmark Study Process

The research process for using LLMs in emotion analysis for mental health follows a structured multi-stage pipeline that begins with data acquisition and ends with clinically interpretable outputs (Figure 1). In the initial stage, raw text data is collected from relevant sources, including social media platforms, clinical transcripts, and electronic health records. This data undergoes preprocessing involving noise removal, tokenization, and domain-specific normalization. The preprocessed text is then encoded using a transformer-based backbone, such as BERT or GPT-4, which generates contextual embeddings capturing nuanced emotional and psychological signals. In the benchmarking phase, multiple LLMs are evaluated in parallel using standardized prompts or fine-tuned configurations against established metrics including F1-score, AUC, and task-specific clinical utility scores. A critical component of the benchmark process is safety stress-testing, where models are evaluated on crisis-related prompts to assess abstention behavior and harm avoidance. Final outputs are reviewed by domain experts before integration into clinical decision-support pipelines [19, 35]. The specific evaluation protocols, dataset-splitting rules, and calibration, fairness, and safety metrics that should accompany this pipeline are detailed in Section 4.4.

3 Results

The comparative analysis presented in Table 2 reveals a clear and progressive evolution in the methodological sophistication applied to emotion analysis within mental health contexts over the

Table 2. Evolution of emotion analysis approaches in mental health.

Dimension	2013–2017	2018–2022	2023–2025
Data	Social media	Social + clinical	Social, clinical, dialogue, multimodal
Model Type	SVM, Logistic Regression	BERT-family (BERT, RoBERTa, MentalBERT, ClinicalBERT, BioBERT)	GPT-4, instruction LLMs, Med-PaLM, Mental-LLM
Learning Strategy	Supervised	Fine-tuning / continued pretraining	Prompting + RLHF
Primary Task	Depression detection	Emotion classification, symptom detection	Emotion inference + dialogue + severity
Metric Focus	F1, AUC	Accuracy, F1 (e.g., GoEmotions BERT macro-F1 0.46)	Multi-task (F1, MAE, empathy ratings)
Key Limitation & Examples	Proxy labels, domain bias; De Choudhury et al. [8]	Domain-shift sensitivity; Devlin et al. [10], Ji et al. [45]	Safety + reliability concerns; Achiam et al. [1], Jin et al. [19]

past decade. During the 2013–2018 period, the dominant paradigm relied on classical supervised learning using support vector machines and logistic regression applied predominantly to social media data [8]. These approaches, while computationally efficient, were severely constrained by their reliance on proxy labels derived from user self-disclosure—such as depression-related hashtags or community membership—rather than clinically validated diagnostic indicators. The resulting models exhibited strong domain bias and limited generalizability beyond the specific platforms on which they were trained.

The transition to BERT-family transformer models during the 2018–2022 era marked a significant methodological advance [10], enabling fine-grained emotion classification with substantially improved accuracy on multi-class tasks. The contextual encoding capacity of BERT-based architectures allowed for more nuanced representation of emotional language, reducing the feature-engineering burden and enabling cross-domain transfer to clinical text. However, domain-shift sensitivity remained a critical limitation, particularly when models trained on general social media corpora were applied to clinical transcripts. The most recent era (2023–2025) is characterized by the adoption of large instruction-following models such as GPT-4 [1], leveraging prompting strategies and reinforcement learning from human feedback [26]. These systems demonstrate superior multi-task flexibility, supporting emotion inference, severity estimation, and empathic dialogue generation within unified architectures [19]. Nevertheless, the shift toward safety-critical deployment has exposed new limitations around reliability, hallucination risk, and the absence of clinically grounded evaluation frameworks.

3.1 Domain-Adapted Encoders and Clinical LLMs

Beyond general-purpose BERT and GPT-4, a family of domain-adapted models has become central to mental health NLP and is essential to a complete account of the field. MentalBERT and MentalRoBERTa [45] are continued-pretraining variants trained on mental-health social-media corpora; across depression, stress, and suicidal-ideation detection benchmarks they consistently match or exceed general BERT and RoBERTa. Clinically pretrained encoders such as ClinicalBERT [46] and BioBERT [47] adapt BERT to clinical notes and biomedical literature, respectively. Importantly, the evidence indicates these clinically pretrained encoders transfer poorly to informal social-media language, underperforming mental-health-adapted models on community text—a concrete instance of domain mismatch rather than a generic domain-shift caveat. At the generative end, clinically grounded LLMs such as Med-PaLM [48] and mental-health-specific frameworks such as Mental-LLM [49] extend instruction-tuned models toward medical question answering and prediction from online text. The coexistence of these model families motivates both the qualitative comparison in Table 4 and the quantitative synthesis in Table 5.

Table 3 provides a granular synthesis of LLM-based methodological approaches employed between 2023 and 2025, illuminating both the diversity of application contexts and the persistent structural gaps that constrain clinical translation [19]. Across

Table 3. LLM-based methodological categories (2023–2025).

Study Type	Dataset	Model	Task	Strength	Research Gap & Examples
Emotion classification	GoEmotions [9] (58k Reddit, 28 labels)	BERT / GPT	Fine-grained emotion labeling	Strong general-domain accuracy (BERT macro-F1 0.46)	Limited clinical mapping; Demszky et al. [9]
Depression severity	DAIC-WOZ / E-DAIC interview transcripts [35]	LLM + regression head	PHQ-8 severity estimation	Context-sensitive extraction (MAE 2.85)	Lack of prospective validation; Sadeghi et al. [35]
Empathy evaluation	Chat dialogue [14]	GPT-4	Counseling response generation	High perceived empathy in blinded ratings	Risk of hallucinated advice; Elyoseph et al. [13], Sorin et al. [37]
Multimodal depression	E-DAIC text + behavioral [35]	LLM hybrid (text + facial + speech)	Severity modeling	Improved contextual inference	Limited dataset diversity; Sadeghi et al. [35], Jin et al. [19]

Table 4. Comparative analysis: traditional NLP vs LLM systems.

Dimension	Lexicon / Classical ML [5]	BERT-family [10, 45]	Instruction LLMs [1, 26]
Feature engineering	Manual	Minimal	None
Context handling	Limited	Strong	Very strong
Multi-task flexibility	Low	Moderate	High
Interpretability	High	Moderate	Variable
Safety control	Deterministic	Moderate	Requires scaffolding
Deployment cost	Low	Medium	High

the four study types reviewed, a consistent pattern emerges: LLM-based systems achieve strong task performance under general-domain conditions but demonstrate notable fragility when confronted with the ecological and regulatory demands of clinical practice. In the emotion classification domain, models trained and evaluated on benchmark datasets such as GoEmotions [9] achieve high accuracy in fine-grained labeling, reflecting the capacity of transformer architectures to distinguish subtle affective distinctions in everyday language. However, the clinical applicability of these classifications remains limited, as general-purpose emotion categories often do not map directly onto symptom constructs used in psychiatric diagnosis, such as anhedonia, affective blunting, or dysphoria. Depression severity estimation using interview transcripts and LLM-plus-regression architectures demonstrates encouraging results [35], particularly in extracting contextually sensitive features from unstructured clinical dialogue; yet the absence of prospective longitudinal validation means that reported metrics reflect retrospective cross-validation on static datasets rather than dynamic real-world prediction. Empathy-focused evaluation reveals that GPT-4-class models can generate responses rated highly by both patients and clinicians in blinded comparisons [13], although the risk of hallucinated or clinically inappropriate advice remains a critical concern requiring robust safety scaffolding and human-in-the-loop oversight [14]. Finally, multimodal depression models that integrate text with behavioral signals show improved contextual inference but are constrained by limited dataset diversity, restricting generalizability across age, gender, and cultural groups [35].

The comparative analysis presented in Table 4 offers a systematic three-way evaluation of the principal NLP paradigms deployed in mental health emotion analysis, revealing a clear trade-off structure with important implications for clinical deployment. Lexicon-based and classical machine learning approaches, while requiring extensive manual feature engineering [5], offer deterministic safety control and high interpretability—attributes that remain critically important in clinical settings where accountability and auditability are paramount. Their low deployment cost makes them viable in resource-constrained environments, yet their limited context handling severely restricts their capacity to capture the nuanced, ambiguous, and often indirect emotional expressions characteristic of mental health discourse. BERT-family models [10] represent a substantial advancement across nearly all dimensions, eliminating manual feature engineering while providing strong contextual representations; their moderate multi-task flexibility

Table 5. Quantitative comparison of representative reported results.

Task	Dataset	Model	Metric	Reported value [Ref]
Fine-grained emotion (28 classes)	GoEmotions (Reddit)	BERT-base	Macro-F1 / Micro-F1	0.46 / 0.51 [9]
Stress detection	SAD / Dreaddit-style	BERT	F1	0.78 [45]
Stress detection	SAD / Dreaddit-style	BioBERT	F1	0.75 [45, 47]
Stress detection	SAD / Dreaddit-style	ClinicalBERT	F1	0.76 [45, 46]
Stress detection	SAD / Dreaddit-style	MentalBERT	F1	0.80 [45]
Stress detection	SAD / Dreaddit-style	MentalRoBERTa	F1	0.82 [45]
Depression severity (PHQ-8)	E-DAIC (interviews)	LLM features + regression (multimodal)	MAE / RMSE	2.85 / 4.02 [35]

*Values are as reported in the cited primary studies on their respective datasets and splits; cross-study comparison is therefore indicative rather than a controlled head-to-head benchmark.

and interpretability make them well-suited to structured clinical NLP tasks such as symptom extraction and severity classification. However, their sensitivity to domain shift—particularly when transitioning from general social media training data to clinical transcripts—remains a persistent limitation that demands domain-specific fine-tuning or continued pretraining on clinical corpora. Instruction-following LLMs such as GPT-4 [1] achieve the highest performance across context handling, multi-task flexibility, and feature generalization, and their ability to process complex, multi-turn dialogue and generate empathically calibrated responses represents a qualitative leap beyond prior paradigms [26]. Nevertheless, their variable interpretability and requirement for safety scaffolding introduce substantial deployment challenges, particularly in regulated healthcare environments, while the high computational cost of inference further limits accessibility in low-resource settings. Taken together, this comparison suggests that future systems will likely adopt hybrid architectures—combining the interpretability and clinical validity of structured encoders with the generative flexibility of instruction-tuned LLMs—to meet the dual demands of clinical performance and ethical governance. Neither pure paradigm alone is sufficient for safe, equitable, and effective deployment in mental health contexts.

3.2 Quantitative Comparison of Reported Performance

To complement the qualitative summaries above, Table 5 consolidates representative quantitative results reported across the included studies. Three analytical observations follow directly from these figures. First, fine-grained emotion classification remains an unsolved problem: even a well-tuned BERT baseline reaches only a macro-F1 of 0.46 on GoEmotions, and reported zero-shot GPT-family results trail fine-tuned encoders by roughly 20–25 macro-F1 points—evidence that prompting alone does not yet substitute for supervised adaptation in fine-grained affect. Second, domain provenance dominates raw scale: mental-health-adapted encoders (MentalBERT, MentalRoBERTa) outperform clinically pretrained encoders (ClinicalBERT, BioBERT) on social-media text, confirming that matching the pretraining domain to the deployment domain matters more than generic medical pretraining. Third, severity estimation is now approaching clinically meaningful error margins, with multimodal LLM pipelines reaching a mean absolute error of roughly 2.85 PHQ-8 points on E-DAIC, although these remain retrospective, single-cohort results.

4 Critical Analysis & Challenges

Emotion-aware LLMs show considerable promise across the mental health landscape, yet their clinical uptake remains constrained by systemic technical, ethical, and governance challenges. Addressing these challenges is essential before LLMs can be responsibly integrated into routine mental health care.

4.1 Evolution of Trends (2023–2025): What Has Changed?

- **Shift from fine-tuning to prompting:** earlier systems relied on supervised BERT fine-tuning, requiring large labeled clinical datasets and expensive retraining cycles for each new task [10, 40]. Recent work increasingly uses prompting strategies with GPT-4-style models that generate structured outputs without parameter updates, enabling rapid adaptation across mental health

tasks with minimal labeled data [1, 21, 26].

- **Incorporation of reasoning chains:** LLMs now generate step-wise rationales for emotion inference, moving beyond opaque classification toward transparent reasoning [14, 17]. This supports clinician oversight but introduces faithfulness concerns, as generated rationales may not accurately reflect the model's internal computation [1, 19, 37].
- **Multimodal integration:** recent studies increasingly integrate textual transcripts with behavioral and physiological features for improved severity modeling [25, 35], capturing complementary affective signals that text alone cannot encode, such as speech rhythm and psychomotor indicators [16, 18].
- **Empathy-focused evaluation:** evaluation frameworks have expanded beyond classification accuracy to include clinician-blinded empathy comparisons and systematic reviews of LLM empathic capacity [13, 37], reflecting a growing recognition that technical performance metrics alone are insufficient [4, 14].

These developments collectively indicate a maturing field transitioning from narrow benchmark optimization toward broader clinical and ethical accountability. The integration of reasoning chains and multimodal signals reflects a growing recognition that mental health is a complex, context-dependent domain that resists reduction to simple classification tasks.

4.2 The Promise: Key Achievements and Strengths

- **Linguistic complexity handling:** LLMs substantially outperform lexicon-based tools in capturing sarcasm, metaphor, understatement, and indirect expressions of distress—phenomena that are particularly prevalent in mental health discourse [1, 31]—enabling more accurate identification of subtle signals such as passive ideation, emotional masking, and culturally specific distress expressions [3, 15].
- **Cross-domain generalization:** instruction tuning enables few-shot and zero-shot transfer across diverse datasets and clinical contexts without task-specific retraining [26]. This is particularly valuable where labeled clinical data is scarce and domain boundaries are fluid, reducing the data-acquisition burden for specialized deployment [2, 19].
- **Reduced feature engineering:** transformer architectures eliminate manual linguistic feature construction—lexicon compilation, part-of-speech pipelines, handcrafted sentiment dictionaries—that characterized earlier systems [10, 40], democratizing access for clinical researchers without deep NLP expertise [3, 33].
- **Empathy and support quality:** LLMs demonstrate strong perceived empathy, with responses rated highly by patients and licensed clinicians in blinded evaluations [13, 37]—a quality essential for mental health applications, where communicative tone influences engagement, disclosure, and outcome [12, 14, 20].

These capabilities collectively position LLMs as transformative tools for scalable mental health support, particularly in resource-constrained settings where access to trained clinicians is limited.

4.3 The Gap: Persistent and Emerging Challenges

4.3.1 Technical Challenges

Hallucinations and fabrication. LLMs have a well-documented tendency to generate plausible-sounding but factually unsupported clinical interpretations [1, 14]. In mental health this risk is acute, because fabricated symptom interpretations, misattributed diagnoses, or invented guidance could directly harm vulnerable users [4, 19]. Unlike factual domains, emotional and psychological inferences are often ambiguous and resistant to objective verification, making detection and correction harder [17]. Robust output filtering, uncertainty quantification, and human-in-the-loop review are therefore essential safeguards [13, 37].

Bias in training data. The social media corpora on which many LLMs are trained overrepresent younger, English-speaking, Western users while underrepresenting older adults, non-Western contexts, and clinical populations [5, 15, 32]. This skew propagates into model outputs, producing classifiers that perform well on majority groups but poorly on minority populations [8, 36]—and, in mental health, can lead to systematic underdiagnosis or misclassification in already underserved communities [29, 41].

Lack of true emotional understanding. Despite strong benchmark performance, LLMs operate through statistical pattern matching whose clinical

adequacy remains unvalidated for high-stakes psychotherapy contexts [27, 37]; empirical evaluations confirm that LLMs show systematic gaps in fine-grained affect recognition and perform inconsistently across emotion categories that require contextual or cultural grounding [11, 34]. Pattern matching can fail catastrophically in novel or culturally non-standard emotional contexts, precisely where robust understanding is most needed [14, 31], and users may erroneously attribute understanding to systems performing surface-level prediction [1, 17].

Computational cost. State-of-the-art instruction-following LLMs require substantial infrastructure for training and inference, including high-memory GPU clusters and significant energy consumption [1, 40]. This creates a barrier to deployment in low-resource clinical settings and lower-income countries—precisely where scalable, affordable tools are most needed [42, 43]—and real-time inference imposes latency constraints that may degrade therapeutic interaction [12, 22].

Collectively, these technical challenges underscore the need for hybrid architectures combining the interpretability of classical systems with the representational power of LLMs, alongside rigorous bias auditing, uncertainty estimation, and computational-efficiency research.

4.4 Clinical and Ethical Challenges

Limited clinical validation. A fundamental limitation is the predominant reliance on proxy labels—Reddit community membership or self-reported hashtags—rather than clinically validated psychometric scales such as the PHQ-9, GAD-7, or structured diagnostic interviews [5, 19]. Substituting convenience labels for clinical ground truth introduces construct-validity threats, and models with high F1 on proxy-labeled data may perform poorly against genuine clinical diagnoses [4, 8, 38].

Black-box interpretability. Post-hoc rationales generated by LLMs may not faithfully reflect the underlying computation [14, 17]. In clinical contexts, where transparency and accountability are legal and ethical requirements, this creates deployment barriers [19, 28]: clinicians require an understandable justification that can be evaluated against clinical knowledge, and the gap between generated explanations and actual model behavior creates accountability risks when erroneous or biased decisions affect care [1, 37].

Privacy risks. Mental health text is among the most sensitive personal data. Under GDPR Article 9, it qualifies as a special category requiring explicit consent, data minimization, and purpose limitation [41]. Pipelines that incorporate social media data scraped without consent, or clinical records shared without adequate anonymization, may violate these protections [5, 6], and the risk of model memorization poses additional threats for sensitive disclosures [19, 28].

Over-reliance risk. As LLMs generate increasingly empathic and coherent responses, users may develop misplaced confidence and treat outputs as equivalent to professional therapeutic input [30, 42], especially when disclosure mechanisms are inadequate or users are emotionally vulnerable [12, 20]. The danger is not only suboptimal advice but delayed help-seeking and consequential decisions based on content that lacks clinical accountability [22–24].

Algorithmic fairness. Emotional expression is shaped by cultural norms, linguistic conventions, and socioeconomic context, yet most systems are trained and evaluated on a narrow cultural and linguistic range [5, 15]. This narrowness produces models that systematically misclassify emotional states in non-Western, multilingual, or culturally diverse populations [8, 36], translating into real disparities in care quality [29, 41].

Taken together, these clinical and ethical limitations point to a single overarching design principle: current LLM systems should augment, not replace, professional clinical judgment. Across every use case reviewed—classification, severity estimation, conversational support, and safety detection—the evidence supports a clinician-in-the-loop model in which the LLM surfaces candidate signals, drafts, and summaries that a qualified professional reviews, contextualizes, and remains accountable for. Autonomous diagnostic or therapeutic decision-making by LLMs is not supported by the present validation evidence and is not advisable in mental health contexts, where the cost of error is high and the population is vulnerable.

4.5 The Benchmarking Problem

A major structural weakness in the field is the absence of standardized, clinically validated benchmarks for LLM-based emotion analysis in mental health. The current landscape is dominated by general-purpose datasets such as GoEmotions, which, while useful

Table 6. Recommended benchmark dimensions for clinical LLM emotion analysis.

Dimension	Representative metrics / protocols	Rationale
Performance	Accuracy, macro/micro-F1, AUROC; MAE/RMSE for severity	Captures both classification and continuous severity tasks
Calibration	Expected calibration error (ECE), Brier score, reliability diagrams	Confidence must be trustworthy for triage and abstention
Fairness	Subgroup F1 gaps, equalized-odds difference, demographic-parity difference	Detects disparate performance across age, gender, ethnicity, language
Safety	Crisis-prompt abstention rate, harmful-response rate, escalation appropriateness	Quantifies behavior in high-risk scenarios
Clinical utility	Agreement with PHQ-9/GAD-7, clinician-rated usefulness, decision-curve analysis	Links model output to clinically meaningful action
Data splits / protocol	Patient-level (not post-level) splits; temporal/prospective held-out sets; cross-site external validation	Prevents leakage and tests real-world generalization

Table 7. The emotion-guided LLM mental health framework (EGLMF): A four-layer taxonomy for diagnosing clinical readiness.

EGLMF Layer	Dimensions	Representative Methods	Maturity
1. Signal	Modality (social media, transcripts, EHR, dialogue, multimodal); provenance; consent status	GoEmotions, CLPsych, E-DAIC, EHR notes	High coverage, low diversity
2. Inference	Task (classification, severity, conversational, empathy, safety); reasoning transparency	BERT/MentalBERT, GPT-4 prompting, LLM + regression	Strong, uneven
3. Validation	Ground truth (proxy vs PHQ-9/GAD-7); retrospective vs prospective; external validity	Cross-validation; rare prospective studies	Weak / emerging
4. Governance	Interpretability, fairness, privacy, safety; scaffolding, human oversight	Calibration, bias audits, abstention, clinician-in-the-loop	Nascent

for broad emotion classification, were not designed to reflect clinical complexity [9]. Using social media self-disclosure as a proxy for clinical diagnosis introduces construct-validity problems: individuals who post depression-related hashtags do not necessarily meet diagnostic criteria, and post content may reflect momentary states rather than persistent conditions [8]. The absence of longitudinal validation further means models are evaluated on static cross-sectional snapshots rather than temporally evolving trajectories [38]—a paradigm poorly suited to mental health, where symptom fluctuation, treatment response, and recovery dynamics are central to clinical relevance.

A clinically adequate benchmark would therefore need to specify explicit evaluation protocols, dataset-splitting rules, and a metric suite that extends well beyond accuracy. Table 6 summarizes the recommended dimensions that such a benchmark should report, covering performance, calibration, fairness, safety, clinical utility, and data-splitting protocols.

Ground-truth labels should derive from validated psychometric instruments such as the PHQ-9 or GAD-7 rather than self-reported proxies, and the cohort should represent diverse demographic groups across age, gender, ethnicity, language, and clinical severity. Safety stress-testing using crisis-simulating prompts is essential. Such a benchmark should be developed through genuine interdisciplinary collaboration among NLP researchers, clinical psychologists, bioethicists, and patient-advocacy groups to ensure it reflects the full complexity of real-world mental health care [4, 17, 19].

4.6 A Proposed Taxonomy: The Emotion-Guided LLM Mental Health Framework (EGLMF)

To move the field beyond descriptive cataloguing, we propose the Emotion-Guided LLM Mental Health Framework (EGLMF), a four-layer taxonomy that organizes existing and future work along the dimensions that most determine clinical readiness.

Rather than classifying systems by model architecture alone, EGLMF locates each study within four interacting layers—signal, inference, validation, and governance—so that strengths and gaps can be diagnosed systematically. The framework makes explicit that a system's clinical maturity is bounded by its weakest layer: high inference performance cannot compensate for proxy-only validation or absent governance. Table 7 presents the EGLMF taxonomy, detailing the dimensions, representative methods, and current maturity level for each layer.

EGLMF implies a staged agenda. In the near term, the field should standardize the Signal and Inference layers by adopting shared, clinically annotated datasets and patient-level evaluation splits. In the medium term, the Validation layer should be strengthened through prospective, longitudinal, and cross-site studies that replace proxy labels with validated psychometric ground truth. In the long term, the Governance layer should be operationalized through calibrated, fairness-audited, privacy-preserving hybrid encoder–LLM systems with built-in abstention and mandatory clinician oversight. Progress on later layers is contingent on the earlier ones; the framework thus doubles as a maturity model and a prioritization guide for funders and developers.

5 Conclusion

This mini review has demonstrated that large language models represent a transformative but still maturing technology for emotion analysis in mental health. The key takeaway is that LLMs have fundamentally shifted the technical frontier—from narrow sentiment classification to contextual affect inference, severity estimation, and empathic dialogue—but this performance advancement has outpaced the development of the clinical validation frameworks, governance structures, and standardized benchmarks needed to justify safe deployment. The evidence synthesized here shows that no current LLM system simultaneously satisfies the requirements of clinical accuracy, demographic fairness, interpretability, privacy compliance, and safety that mental health applications demand. Accordingly, the responsible near-term role of these systems is to support and augment clinicians—surfacing signals, drafts, and summaries for professional review—rather than to replace clinical judgment.

Progress will require a deliberate pivot away from benchmark-driven optimization toward governance-first development: clinically grounded emotion ontologies, prospective longitudinal validation studies, hybrid encoder–LLM architectures that balance expressiveness with interpretability, and interdisciplinary evaluation protocols co-designed with clinicians, ethicists, and patient communities. The Emotion-Guided LLM Mental Health Framework (EGLMF) presented in Section 4.5 offers one conceptual scaffold for this integrative approach, functioning simultaneously as a taxonomy and a maturity model. Realizing the genuine potential of LLMs in mental health will depend not on further performance gains alone, but on the field's collective commitment to rigor, equity, and safety as foundational design principles.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that ChatGPT-5.5 was used for language editing and rewriting of parts of the manuscript to improve clarity and effectiveness. The authors have carefully reviewed, revised, and verified all AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. [CrossRef]

[2] Aragon, M. E., Lopez-Monroy, A. P., González-Gurrola, L. C., & Montes-y-Gómez, M. (2021). Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. *IEEE transactions on affective computing*, *14*(1), 211-222. [CrossRef]

[3] Bashynska, I., Sarafanov, M., & Manikaeva, O. (2024). Research and development of a modern deep learning model for emotional analysis management of text data. *Applied Sciences*, *14*(5), 1952. [CrossRef]

[4] Bucher, A., Egger, S., Vashkite, I., Wu, W., & Schwabe, G. (2025). "It's Not Only Attention We Need": Systematic Review of Large Language Models in Mental Health Care. *JMIR mental health, 12*, e78410. [CrossRef]

[5] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine, 3*(1), 43. [CrossRef]

[6] Clay, B., Bergman, H. I., Salim, S., Pergola, G., Shalhoub, J., & Davies, A. H. (2025). Natural language processing techniques applied to the electronic health record in clinical research and practice-an introduction to methodologies. *Computers in Biology and Medicine, 188*, 109808. [CrossRef]

[7] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 31-39). [CrossRef]

[8] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media* (Vol. 7, No. 1, pp. 128-137). [CrossRef]

[9] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020, July). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4040-4054). [CrossRef]

[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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)* (pp. 4171-4186). [CrossRef]

[11] Broekens, J., Hilpert, B., Verberne, S., Baraka, K., Gebhard, P., & Plaat, A. (2023, September). Fine-grained affective processing capabilities emerging from large language models. In *2023 11th international conference on affective computing and intelligent interaction (ACII)* (pp. 1-8). IEEE. [CrossRef]

[12] Föyen, L. F., Zapel, E., Lekander, M., Hedman-Lagerlöf, E., & Lindsäter, E. (2025). Artificial intelligence vs. human expert: Licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship. *Internet Interventions, 41*, 100841. [CrossRef]

[13] Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in psychology, 14*, 1199058. [CrossRef]

[14] Gabriel, S., Puri, I., Xu, X., Malgaroli, M., & Ghassemi, M. (2024, November). Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 2206-2221). [CrossRef]

[15] Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences, 18*, 43-49. [CrossRef]

[16] Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J. C., & Ungar, L. H. (2019, July). Understanding and measuring psychological stress using social media. In *Proceedings of the international AAAI conference on web and social media* (Vol. 13, pp. 214-225). [CrossRef]

[17] Hameed, S., Nauman, M., Akhtar, N., Fayyaz, M. A., & Nawaz, R. (2025). Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models. *Frontiers in Artificial Intelligence, 8*, 1627078. [CrossRef]

[18] Ignatiev, N., Smirnov, I. V., & Stankevich, M. (2022, February). Predicting Depression with Text, Image, and Profile Data from Social Media. In *ICPRAM* (pp. 753-760). [CrossRef]

[19] Jin, Y., Liu, J., Li, P., Wang, B., Yan, Y., Zhang, H., ... & Wang, Y. (2025). The applications of large language models in mental health: scoping review. *Journal of Medical Internet Research, 27*(1), e69284. [CrossRef]

[20] Karkosz, S., Szymański, R., Sanna, K., & Michałowski, J. (2024). Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: randomized controlled trial. *JMIR formative research, 8*(1), e47960. [CrossRef]

[21] Lan, X., Han, Z., Cheng, Y., Sheng, L., Feng, J., Gao, C., & Li, Y. (2025, November). Depression detection on social media with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 2155-2171). [CrossRef]

[22] Löchner, J., Carlbring, P., Schuller, B., Torous, J., & Sander, L. B. (2025). Digital interventions in mental health: an overview and future perspectives. *Internet Interventions, 40*, 100824. [CrossRef]

[23] MacNeill, A. L., Doucet, S., & Luke, A. (2024). Effectiveness of a mental health chatbot for people with chronic diseases: randomized controlled trial. *JMIR Formative Research, 8*, e50025. [CrossRef]

[24] Moylan, K., & Doherty, K. (2025). Expert and interdisciplinary analysis of AI-driven chatbots for mental health support: mixed methods study. *Journal of Medical Internet Research, 27*(1), e67114. [CrossRef]

[25] Nag, P. K., Bhagat, A., Priya, R. V., & Khare, D. K. (2023, December). Emotional intelligence through artificial intelligence: Nlp and deep learning in the analysis of healthcare texts. In *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)* (Vol. 1, pp. 1-7). IEEE. [CrossRef]

[26] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems, 35*, 27730-27744.

[27] Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., ... & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research, 3*(1), 12. [CrossRef]

[28] Pillay, Y. (2025, November). Ethical Decision-Making Guidelines for Mental Health Clinicians in the Artificial Intelligence (AI) Era. In *Healthcare* (Vol. 13, No. 23, p. 3057). MDPI. [CrossRef]

[29] Qasim, A., Mehak, G., Hussain, N., Gelbukh, A., & Sidorov, G. (2025). Detection of depression severity in social media text using transformer-based models. *Information, 16*(2), 114. [CrossRef]

[30] De Witte, N. A., Joris, S., Van Assche, E., & Van Daele, T. (2021). Technological and digital interventions for mental health and wellbeing: an overview of systematic reviews. *Frontiers in digital health, 3*, 754337. [CrossRef]

[31] Rezapour, M. (2024). Emotion detection with transformers: A comparative study. *arXiv preprint arXiv:2403.15454*. [CrossRef]

[32] Robertson, C., Carney, J., & Trudell, S. (2023). Language about the future on social media as a novel marker of anxiety and depression: A big-data and experimental analysis. *Current Research in Behavioral Sciences, 4*, 100104. [CrossRef]

[33] Ruiz, G. B., & Herrera, R. D. J. G. (2025). Textual emotion detection with complementary BERT Transformers in a condorcet's jury theorem assembly. *Knowledge-Based Systems*, 114070. [CrossRef]

[34] Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intelligent Systems, 38*(2), 15-23. [CrossRef]

[35] Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L. H., Rahimi, F., ... & Eskofier, B. M. (2024). Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research, 3*(1), 66. [CrossRef]

[36] Safa, R., Edalatpanah, S. A., & Sorourkhah, A. (2023). Predicting mental health using social media: A roadmap for future development. In *Deep learning in personalized healthcare and decision support* (pp. 285-303). Academic Press. [CrossRef]

[37] Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2024). Large language models and empathy: systematic review. *Journal of medical Internet research, 26*, e52597. [CrossRef]

[38] Thorstad, R., & Wolff, P. (2019). Predicting future mental illness from social media: A big-data approach. *Behavior research methods, 51*(4), 1586-1600. [CrossRef]

[39] Vale, L., & Lee, A. (2024). Advancing medical diagnosis: enhancing sentiment analysis in electronic medical records with transformer models. *Journal of Computer Technology and Software, 3*(7). https://ashpress.org/index.php/jcts/article/view/89

[40] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008).

[41] Villanueva-Miranda, I., Xie, Y., & Xiao, G. (2025). Sentiment analysis in public health: a systematic review of the current state, challenges, and future directions. *Frontiers in Public Health, 13*, 1609749. [CrossRef]

[42] World Health Organization. (2019). *WHO guideline: Recommendations on digital interventions for health system strengthening*. World Health Organization. https://www.who.int/publications/i/item/9789241550505

[43] World Health Organization. (2022). *World mental health report: Transforming mental health for all*. World Health Organization. https://www.who.int/publications/i/item/9789240049338

[44] Rasool, A., Shahzad, M. I., Aslam, H., Chan, V., & Arshad, M. A. (2025). Emotion-aware embedding fusion in large language models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for intelligent response generation. *AI, 6*(3), 56. [CrossRef]

[45] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022, June). Mentalbert: Publicly available pretrained language models for mental healthcare. In *proceedings of the thirteenth language resources and evaluation conference* (pp. 7184-7190).

[46] Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. (2019, June). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd clinical natural language processing workshop* (pp. 72-78). [CrossRef]

[47] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234-1240. [CrossRef]

[48] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172-180. [CrossRef]

[49] Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., ... & Wang, D. (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, 8*(1), 1-32. [CrossRef]