MENTABENCH: BENCHMARKING MENTAL HEALTH SIGNAL DETECTION FROM TEXT WITH SMALL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-based mental-health signals could support just-in-time adaptive interventions (JITAIs). In settings such as smartphones, they must also use limited computing resources and protect sensitive user text. Although LLMs can detect these signals, their computational demands and reliance on remote servers can introduce delays and privacy risks, making them a poor fit for practical use. Small language models (SLMs) offer a promising alternative; however, their capability for mental health signal detection remains underexplored. We introduce MENTABENCH, a unified benchmark for evaluating SLMs on text-based mental health signal detection. We curated a large scale mental health corpus which comprises 21 public datasets spanning 40 tasks and diverse text sources, and evaluates 11 representative SLMs under zero-shot, few-shot, and fine-tuning regimes. We benchmarking results show that fine-tuned SLMs on mental health datasets can outperform traditional machine learning and encoder-based baselines as well as LLMs. Evaluating on both predictive performance and deployment-related metrics, we further find that Gemma4-E4B achieves the best performance, while LLaMA3.2-1B and LLaMA3.2-3B reach a more balanced trade-off for real-world deployment scenarios. MENTABENCH provides a standardized data source and benchmarking tools for assessing the generalization, adaptability, and deployability of efficient language models for mental health monitoring and future JITAI systems. Code is available at: [MentaBenchCode].

1 INTRODUCTION

Mental health conditions affect hundreds of millions of people worldwide (Piao et al., 2024; Weaver et al., 2025). Textual data, including social media posts (Hamilton et al., 2025), personal journals (Zhao et al., 2013), and self-reports, have increasingly been studied as rich sources of behavioral and psychological signals for mental health research (Hussain et al., 2025; Zhunis et al., 2022), such as identifying early warning signs of mental distress (Chancellor & De Choudhury, 2020; De Choudhury et al., 2013). Just-in-time adaptive interventions (JITAIs) offer a framework for continuous monitoring and timely support in everyday settings. Empirical evidence suggests that JITAIs can yield measurable improvements in mental health outcomes (Iorfino et al., 2021; Hickie et al., 2019). For example, studies show that successful JITAIs can reduce depression and anxiety symptoms, with benefits sustained for 3–6 months (Elmer et al., 2025). These findings highligh the urgent need for robust models capable of timely detection and monitoring of mental health-related signals from naturalistic text.

Recent advances in large language models (LLMs) show promise for mental health prediction and reasoning tasks, which offering opportunities for in-context understanding to better support patient assessment (Xu et al., 2024). The ability to process free-form text makes LLMs attractive for a variety of mental health applications, including components of digital phenotyping. However, deploying LLMs in JITAI settings remains challenging. Their high computational, memory, and energy requirements, together with their reliance on centralized and cloud infrastructure, limit accessibility and raise practical concerns in privacy-sensitive and resource-constrained environments (Jin et al., 2025). In contrast, small language models (SLMs), *typically with fewer than 7 billion parameters* (Van Nguyen et al., 2024), are designed for lower-resource and lower-latency inference.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Prior work shows that inference latency generally increases with model scale, whereas SLMs can provide substantially better efficiency in practical deployments (Gondara et al., 2025), including multi-fold speedups ($\sim 3\times$) (Gloeckle et al., 2024; Gao et al., 2025) and up to $20\times$ lower latency in real-world systems (Goel et al., 2026). SLMs allow on-device inference by avoiding cloud data transmission and improves accessibility by delivering mental health support directly to users, reducing reliance on network connectivity (Mandal et al., 2025) and traditional clinical barriers (Bunyi et al., 2021). These properties make SLMs a nature candidate for privacy-preserving, low-latency mental health monitoring systems, particularly when timely detection is necessary for adaptive intervention (Figure 13).

Unfortunately, evaluations of using language models (LMs), including LLMs and SLMs, for text-based mental health signal detection remain limited in several ways. First, existing datasets differ substantially in data sources, annotation protocols, target constructs, and label definitions (Eichstaedt et al., 2018; Althoff et al., 2016), which can cause results to reflect benchmark-specific conventions rather than truly mental-health-relevant learning. As a result, performance on a single dataset may not transfer reliably, and reported model performance may instead arise from dataset-specific artifacts, platform effects, or label leakage (Chancellor & De Choudhury, 2020). Moreover, prior work is often restricted to specific conditions or narrow domains, such as depression (Xu et al., 2024), suicidal thoughts (Marengo & Longobardi, 2025), and affective states (Zhang et al., 2024b), and commonly relies on social media data (Chancellor & De Choudhury, 2020; Harrigian et al., 2021), leaving open whether conclusions hold across different constructs, datasets, and data sources. Finally, prior studies often evaluate models in limited learning setting, such as zero-shot prompting or supervised fine-tuning, leaving unclear how much models benefit from task-specific supervision and whether those gains persist across datasets and mental health constructs.

In this work, we introduce MENTABENCH, a unified benchmark for evaluating SLMs on text-based mental health signal detection. To our knowledge, MENTABENCH is the *first benchmark to systematically examine SLMs across a broad collection of public mental health datasets*, multiple adaptation regimes, and deployment-aware efficiency constraints. Unlike prior evaluations limited to individual datasets, specific mental health conditions, or a single evaluation setting, MENTABENCH supports systematic comparisons of model performance across datasets, tasks, model sizes, and differen learning settings. We curated a large scale collection of 21 datasets and 40 tasks spanning diverse text sources and mental health-related constructs, and evaluates 11 SLMs under zero-shot, few-shot, and fine-tuning settings. This design enables direct analysis of cross-dataset robustness, task-level generalization, and the stability of model rankings across evaluation conditions. In addition, we incorporate deployment-aware metrics, including inference latency and resource efficiency, to assess whether models with strong predictive performance are also suitable for low-latency, resource-constrained, and potentially on-device mental health applications. Our contributions are threefold:

- **A large-scale data curation for mental health detection.** We introduce MENTABENCH, where we curated a large scale of mental health dataset consists of 21 datasets from diverse sources and 40 mental health-related tasks.
- **A systematic evaluation across different adaptation regimes.** We evaluate 11 SLMs under zero-shot, few-shot, and fine-tuning settings, allowing controlled comparison of how supervision, prompting, and model scale affect mental health signal detection. We observe that fine-tuned SLMs outperform traditional machine learning and encoder-based baselines, and outperform LLMs in both predictive performance and deployment efficiency while supporting preserving privacy inference.
- **A deployment-aware evaluation of SLMs.** We jointly evaluate SLMs' performance and efficiency-related metrics to provide practical insight into low-latency, resource-constrained, and on-device mental health monitoring. We also release an open-source on-device deployment package to the community to help researchers train and evaluate different models.

2 RELATED WORK

Language Models and Mental Health Prediction. Recent work has adapted LLMs for mental health prediction (Sarkar et al., 2022). MentalChat16K (Xu et al., 2025) introduces an LLM-generated conversational mental health dataset, while it primarily relies on LLM-augmented data and does not reflect on benchmarking diverse real-world mental health detection tasks. Mental-LLM (Xu et al.,

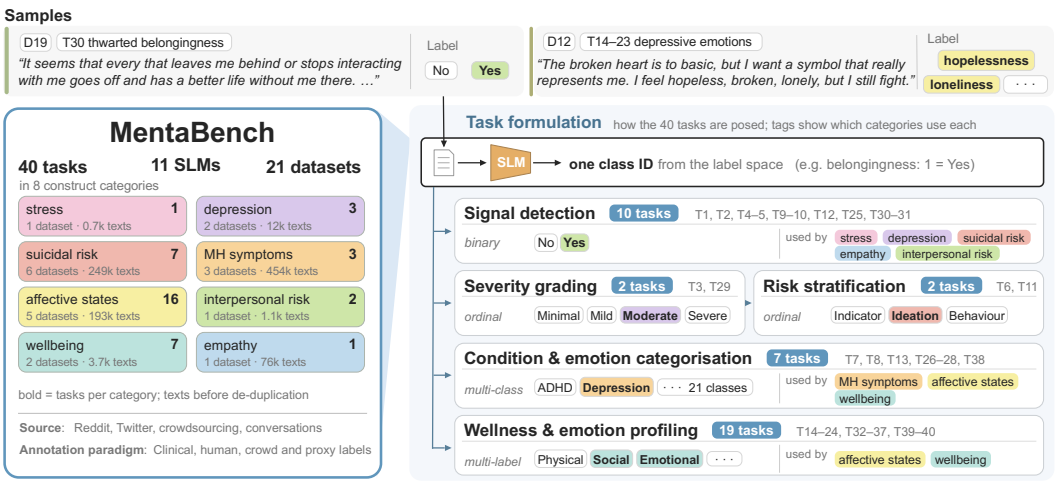


Figure 1: MentaBench workflow. MentaBench curated a large-scale cohort of datasets, including 40 mental health and well-being classification tasks from 21 datasets, and evaluates 11 SLMs using 11 metrics.

2024) shows that instruction-tuned LLMs could improve balanced accuracy and outperform task-specific baselines while improving cross-dataset generalization. Follow-up studies further develop domain-specialized models that achieve performance comparable to or exceeding larger LLMs on mental health tasks (Shi et al., 2025). Despite these advances, LLM-based approaches remain limited by privacy risks (Sarwar, 2025), deployment constraints (Maurya et al., 2025), high computational cost and inference time (Wang et al., 2023), highlighting the need for SLMs as a more practical alternative.

Evaluation and Benchmarking in Mental Health Prediction. Recent work has introduced benchmark-style evaluations for mental health detection, such as the CLPsych shared tasks (Coppersmith et al., 2015) and the CLEF eRisk series (Losada et al., 2018), but these work focus primarily on depression and suicide risk detection with limited evaluations.. Another study adapted LLMs for mental health prediction among limited mental health tasks from social media posts (Yang et al., 2024). Following, Mental-LLM (Xu et al., 2024) provided an evaluation of zero-shot and instruction-tuned LLMs across four mental health datasets. Moreover, a recent evaluation framework (Wang et al., 2025b) examined SLMs on health prediction tasks such as calori level predictions. However, none of past studies have explored extensive mental health evaluation and on-device deployment performance of SLMs. In comparison, our proposed MentaBench provides a large-scale, standardized benchmark for evaluating SLMs on text-based mental health prediction across a curated lare-scale collection of datasets.

3 MentaBench

3.1 Data Curation and Task

We first reviewed publicly available datasets across sources including Papers with Code, Kaggle, Hugging Face, and prior literature in mental health NLP. Guided by the goal of understanding how evaluation outcomes vary across datasets and SLMs, we organized the collected datasets according to key characteristics relevant to mental health signal detection, including data source (e.g., broad social media, crowdsourcing, curated corpora), label definition (e.g., proxy labels, human annotation, clinical/expert annotation), task formulation (e.g., binary classification, multi-class, multi-label), and target constructs (e.g., depression, stress, suicidality, emotional states, and interpersonal risk factors). Through reviewing dataset documentation and prior studies, we observed substantial variation in data generation processes and annotation assumptions across datasets.

Table 1: MentaBench: mental health detection consisting of 40 classification tasks among 21 datasets, in total 990,204 texts before pre-processing.

ID	Task (No. of Labels): Labels	Annotation Paradigm	Avg. Tok.	Size	Source	Category
D01 (Turcan & McKeown, 2019)	T1 (2): No; Yes	Crowdsourced annotation	101	715	Reddit	Stress
D02 (Naseem et al., 2022)	T2 (2): No; Yes; T3 (4): Minimal; Mild; Moderate; Severe	Clinical/Expert annotation	101	3,553	Reddit	Depression
D03 (Haque et al., 2021)	T4 (2): No; Yes	Proxy labeling	85	379	Reddit	Suicidal ideation
D04 (Gaur et al., 2019)	T5 (2): No; Yes; T6 (5): Supportive; Indicator; Ideation; Behaviour; Attempt	Clinical/Expert annotation	1,611	500	Reddit	Suicidal risks
D05 (kamruzzaman-asif, 2025)	T7 (21): COVID19_support; Addiction; ADHD; Alcoholism; Anxiety; Autism; Bipolar; BPD; Conspiracy; Depression; Divorce; Guns; Health anxiety; Jokes; Legal advice; Lonely; Meditation; Personal finance; Relationships; Schizophrenia; Teaching	Proxy-based	210	369,101	Social media	Symptoms
D06 (Ji et al., 2021)	T8 (5): Anxiety; SuicideWatch; Bipolar; Depression; OffMyChest	Human annotation	215	34,823	Reddit	Symptoms
D07 (Mafi & Alam, 2023)	T9 (2): Yes; No	Proxy labeling	173	14,242	Reddit	Suicidal risks
D08 (Syeda Amna, 2023)	T10 (2): Yes; No	Human annotation	28	1,785	Twitter	Suicidal risks
D09 (Infinite Dataset Hub, 2025)	T11 (3): High; Low; Moderate	Proxy labeling	17	100	Reddit	Suicidal risks
D10 (Lucidest, 2024)	T12 (2): Yes; No	Proxy labeling	168	232,073	Reddit	Suicidal risks
D11 (Ourafla, 2026)	T13 (4): Anxiety; Depression; Normal; Suicidal	Human annotation	92	49,612	Reddit + curated datasets	Symptoms
D12 (Abu Bakar Siddiqur Rahman, 2023)	T14–T23 (2): Absent; Present	Crowdsourced human annotation	129	100	Social media	Anger; Brain dysfunction; Emptiness; Hopelessness; Lack of energy; Loneliness; Sadness; Self-hate; Suicide intent; Worthless
D13 (Demszky et al., 2020)	T24, T39, T40 (2): Present; Absent	Crowdsourced annotation	17	28,438	Reddit	Neutral; Sadness; Disappointed
D14 (Rashkin et al., 2019)	T25 (2): Low; High	Crowdsourced annotation	17	76,492	Crowdsourcing	Empathy
D15 (Saravia et al., 2018)	T26 (6): Sadness; Joy; Love; Anger; Fear; Surprise	Human annotation	20	16,000	Twitter	Emotions
D16 (Li et al., 2017)	T27 (7): Anger; Disgust; Fear; Happiness; No emotion; Sadness; Surprise	Human annotation	14	102,979	Real-life conversations	Emotions
D17 (Barbieri et al., 2020)	T28 (3): Negative; Neutral; Positive	Human annotation	27	45,615	Twitter	Sentiment
D18 (Kayalvizhi et al., 2022)	T29 (3): No; Moderate; Severe	Human annotation	172	8,891	Social media	Depression
D19 (Garg et al., 2023)	T30 (2): No; Yes; T31 (2): No; Yes	Human annotation	138	1,057	Social media	T30: Belong; T31: Burden
D20 (Sathvik & Garg, 2023)	T32–T37 (2): Absent; Present	Human annotation	147	657	Reddit	Spiritual; Physical; Intellectual; Social; Vocational; Emotional
D21 (Liyanage et al., 2023)	T38 (4): Aspect 1; Aspect 2; Aspect 3; Aspect 4 (Specified in Appendix A)	Human annotation	26	3,092	Reddit	Wellness aspects

We curated a representative set of 21 datasets covering a broad spectrum of mental health–related tasks (ref. Table 1), including binary detection, ordinal severity or risk prediction, multiclass categorization, and one-versus-rest classification over potentially co-occurring states. Detailed task descriptions are provided in Appendix A. To ensure the benchmark is both informative and reproducible, we applied several selection criteria: (1) public accessibility and availability of usage information including dataset licenses, repository documentation, or platform-specific usage conditions where provided, (2) sufficient data quality and scale, and (3) diversity in task formulations and data sources. We further aimed to include datasets spanning a range of task formulations and textual domains, supporting broad evaluation coverage across mental health detection settings. More detailed information on dataset descriptions, benchmark construction criteria, manual audit procedures, and privacy-related data filtering are provided in Appendix Section A. The resulting benchmark encompasses heterogeneous text sources, enabling systematic analysis of cross-dataset robustness and the consistency of model performance across diverse mental health signal detection tasks.

3.2 Zero-shot, Few-shot and Finetune of SLMs

We evaluate SLMs in four settings: a general model used without task-specific adaptation, zero-shot prompting, few-shot prompting, and task-specific fine-tuning. These settings allow us to compare models used as-is with models given examples in a prompt or adapted through training.

Let $\mathcal{D} = \{D_1, \ldots, D_M\}$ denote a collection of datasets. Each dataset $D_m = \{(x_i^{(m)}, y_i^{(m)})\}_{i=1}^{n_m}$ consists of text inputs $x_i^{(m)} \in \mathcal{X}_m$ (e.g., social media posts, self-reports, or narratives) and labels $y_i^{(m)} \in \mathcal{Y}_m$, where the label space $\mathcal{Y}_m$ may vary across datasets. Among the 21 datasets, a total of 40 tasks were defined as shown in Table 1.

Given an input text x, a model f_θ produces a prediction $\hat{y} = f_\theta(x)$. We evaluate models under three adaptation regimes: (i) zero-shot, where predictions are made based solely on task descriptions without access to labeled examples; (ii) few-shot, where predictions are conditioned on a support set $\mathcal{S}_m = \{(x_j, y_j)\}_{j=1}^{k}$ drawn from D_m, with $k = 3$; and (iii) fine-tuning, where model parameters θ are updated using training data. In the zero-shot and few-shot settings, predictions are generated by conditioning on a task-specific prompt p_m that describes the label space $\mathcal{Y}_m$. In the few-shot regime, the prompt additionally includes the support set $\mathcal{S}_m$, such that: $\hat{y} = f_\theta(x \mid p_m, \mathcal{S}_m)$.

In the fine-tuning setting, for each task dataset D_m with size n_m, we construct training, validation, and test splits according to dataset size. If $n_m < 2000$, we partition the full dataset such that $|D_m^{\text{train}}| = 0.72n_m$, $|D_m^{\text{val}}| = 0.08n_m$, and $|D_m^{\text{test}}| = 0.20n_m$. For larger datasets ($n_m \geq 2000$), we subsample 2,000 instances using the same split ratio, as fine-tuning performance was generally strongest at around 1,000 training samples (ref. Table 11, 12, 13, 14, 15). The same partitioning protocol is applied consistently for fine-tuned SLMs, BERT baselines, and TF-IDF + Logistic Regression baselines. For all fine-tuning experiments, we perform three independent runs and report the average performance across runs.

4 Experimental Set-up

This section presents the experimental setup, including evaluation metrics, model configurations, and implementation details across zero-shot, few-shot, and fine-tuning settings.

4.1 Models

We evaluate 11 lightweight representative SLMs, with parameter sizes ranging from approximately 1B to 7B (Wang et al., 2024). Our model selection focuses on architectures suitable for resource-constrained and on-device settings. The evaluated models include LLaMA-based models (LLaMA 3.2-1B (Meta AI, 2024) and 3B (Grattafiori et al., 2024), TinyLLaMA-1.1B (Zhang et al., 2024a)), Gemma models (Gemma 4-E2B (Google, 2026a), and Gemma 4-E4B (Google, 2026b)), Qwen models (Qwen 2.5-3B (Team, 2024), Qwen 3-4B (Yang et al., 2025), and Qwen 3.5-4B (Qwen Team, 2026)), as well as other compact language models, including Falcon-1.3B (Almazrouei et al., 2023), StableLM-3B (Pinnaparaju et al., 2024), and Phi4Mini-3.8B (Abouelenin et al., 2025).

4.2 Evaluation

We evaluate model performance along two dimensions: predictive accuracy and deployment efficiency. For predictive performance, we report standard classification metrics, including accuracy (ACC), balanced accuracy (BACC), F1 score, area under the precision–recall curve (macro-AUPRC), precision, and recall. We primarily use BACC for cross-task comparisons due to class imbalance, with further justification and conceptual task groupings provided in Appendix Section B. For deployment efficiency, we measure on-device inference performance using latency and resource-related metrics and report time-to-first-token (TTFT), input tokens per second (ITPS), output encoding time (OET), output tokens per second (OTPS), total inference time, and peak memory usage (RAM). The predictive performance of the Gemma 4-E2B and Gemma 4-E4B models differs slightly from other models due to deployment constraints, as inference is conducted through an encapsulation-based pipeline. Since this pipeline may introduce variations in performance metrics, we evaluate these models under two settings: before and after deployment. For model comparison, we report results from the deployed setting to ensure consistency with practical usage scenarios.

4.3 Implementation

We convert textual inputs into embeddings and use BERT as well as TF-IDF with Logistic Regression (LR) to predict numeric labels. The number of test samples is kept consistent with the zero-shot evaluation setting. We also evaluate Gemini2.0-Flash-Lite and Gemini2.5-Flash-Lite as LLM baselines, using a temperature of zero and conducting three inference runs for each input text.

For SLMs, we used parameter-efficient fine-tuning via LoRA (rank $r = 16$, scaling factor $\alpha = 32$, dropout rate = 0.05). Model training is performed using the AdamW optimizer with a cosine learning rate scheduler. We set a warm-up ratio of 0.03 and apply gradient clipping with a maximum norm of 1.0. The training objective is standard cross-entropy loss. For all SLMs, the temperature is set to zero during inference. Experiments were conducted on NVIDIA A100 GPUs for perdiction performance evaluation. To assess practical on-device deployment performance, we additionally performed mobile inference experiments on an iPhone 15 Pro Max.

Across the datasets, we identified 31,507 duplicate instances (identical text and labels) among 990,204 samples, resulting in a final corpus of 958,697 unique texts. To ensure consistency across datasets, textual outputs are mapped to numerical label indices for evaluation. No additional text pre-processing

or normalization was applied, as preserving the original form of user-generated content is important for realistic evaluation.

5 RESULTS

In this section, we provide an overview of the model predictions based on our MentaBench benchmark among different settings and demonstrates the on-device performance of the SLMs.

5.1 PREDICTIVE PERFORMANCE ANALYSIS

Zero-shot and few-shot settings. We evaluate 11 SLMs under zero-shot and few-shot settings, with results averaged across 40 tasks reported in Table 2. Overall, we observe that pre-trained SLMs exhibit a basic capability for mental health signal detection from user-generated text, but leave substantial room for improvement.

Across both settings, most models achieve modest performance, with gemma4-E2B and gemma4-E4B consistently outperforming others, reaching around 0.65 BACC on average. The F1 scores closely track ACC and BACC (within approximately 5%), suggesting relatively balanced performance across classes. Transitioning from zero-shot to few-shot yields only marginal improvements, indicating that providing three in-context examples is insufficient for SLMs to acquire substantial task-specific knowledge. This raises the question of how much supervision is required for effective learning, which we further investigate with fine-tuned models. We also conducted prompt sensitivity analysis, more details can be found in Table 7.

At the task level, Gemma4-E4B and Gemma4-E2B achieve strong performance on tasks such as T1, T9, and T22 (BACC > 0.80), while performing poorly on more challenging tasks including T3 (depression), T6 (suicidal risks), T11 (suicidual risks), T26 (emotions), and T38 (wellness) (BACC < 0.50). Meanwhile, Qwen2.5-3B and phi4Mini-3.8B demonstrate strong performance on certain suicidal risk detection tasks T9 and T12 (BACC > 0.85), but exhibit less stable behavior across other related tasks. More detailed task-level BACC results are shown in Figures 9, 10, and 11.

We further group the 40 tasks into eight categories: stress, depression, suicidal risk, mental health symptoms, affective states, empathy, interpersonal risk factors (IRF), and wellbeing. The performance of SLMs under zero-shot and few-shot settings across these categories, along with the corresponding number of test samples, is shown in Figure 4. Although task difficulty may vary with the number of classes, we observe that SLMs generally perform well on stress and IRF tasks.

Fine-tuned Models. We evaluate the performance of fine-tuned SLMs on mental health detection tasks, with results reported in Table 2. Among the evaluated models, Gemma4-E4B achieves the best performance, reaching a BACC of 0.76, followed by LLaMA3.2-3B (BACC = 0.74) (ref. Figure 8). Models such as phi4Mini-3.8B, LLaMA3.2-1B, and Gemma4-E2B achieve comparable performance, with BACC of 0.7. Overall, fine-tuned SLMs consistently outperform their zero-shot and few-shot counterparts, yielding an average absolute improvement of approximately 10% across tasks. This indicates that fine-tuned SLMs are capable of achieving strong predictive performance across diverse mental health-related constructs. More details of few-shot inference and fine-tuning ablation can be found in Appendix Section C.1

Across fine-tuned models, Tasks 9 (suicidal risks) (BACC = 0.89), 10 (BACC = 0.84), and 12 (emotions) (BACC = 0.83) (binary suicidal risk prediction tasks) achieve the highest average performance across the 11 SLMs. At the category level, SLMs perform best on stress detection tasks, followed by IRF tasks, suicidal risk detection, affective states, wellbeing, mental health symptoms, and depression, with the lowest performance observed on empathy detection tasks. We further analyze task difficulty in relation to the number of classes. As shown in Figure 4, model performance generally decreases as the number of classes increases, indicating that multi-class settings pose greater challenges for SLMs. For tasks with more than five classes, performance becomes more variable across models.

Comparing precision and recall of the models, we found that most SLMs have errors balanced between false positives and false negatives. Fine-tuned models consistently achieve higher and more stable precision and recall, while zero-shot and few-shot settings exhibit substantially greater variability across tasks. As shown in Figure 2, LLaMA3.2-1B and Qwen3-4B exhibit variability in zero-shot and

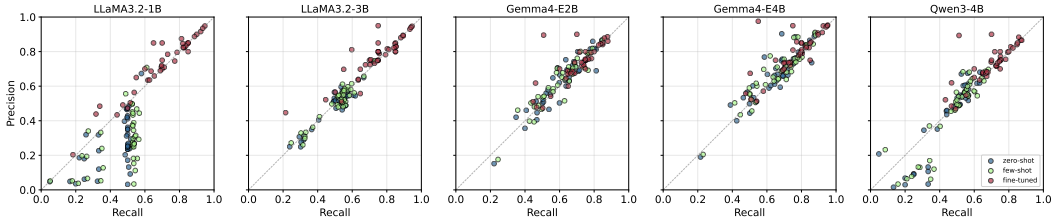


Figure 2: Task-level precision-recall distributions for representative SLMs under zero-shot, few-shot, and fine-tuned settings across MentaBench tasks. Each point represents the performance of a single task.

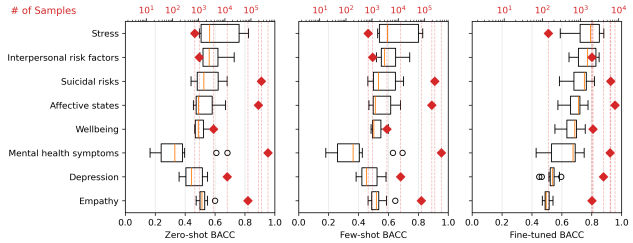


Figure 3: Comparison of SLM performance in terms of BACC across different task categories.

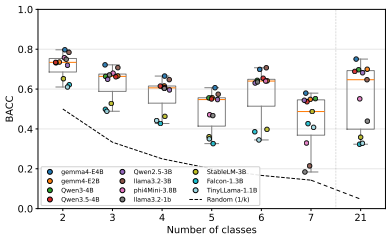


Figure 4: BACC distribution across 40 tasks for the fine-tuned SLMs, grouped by the number of classes.

few-shot setting when detecting mental health status. In contrast, Gemma-family models demonstrate stronger zero-shot and few-shot alignment. More information for dataset imbalance and annotation paradigm analysis can be found in Appendix Section C.1.

Our results show that all fine-tuned SLMs outperform both BERT and TF-IDF with Logistic Regression baselines, showing that SLMs has better capability in understanding mental health prediction tasks using text (Table 3). More baseline comparison against these traditional models can be found in Table 16, and Table 22 in Appendix Section C.1. Furthermore, the best-performing fine-tuned model, Gemma4-E4B, achieves comparable or better performance to the LLM baselines Gemini2.0-Flash-Lite and Gemini2.5-Flash-Lite as shown in Figure 5. These observations highlight the strong ability of well-trained SLMs to match or exceed the performance of larger LLMs while maintaining lower inference latency and supporting privacy-preserving local deployment, making them well-suited for real-world mental health monitoring and in-situ detection applications.

In qualitative analysis, we observe that fine-tuned SLMs can capture semantically meaningful mental health signals beyond simple keyword matching. As shown in Appendix C.2, the models differentiate protective future-oriented behaviour from suicidal intent, identify explicit hopelessness and self-harm expressions, and recognize concise indicators associated with lower suicide risk. The generated rationales also suggest a limited ability to provide partially interpretable explanations grounded in contextual cues and behavioural patterns, although their reliability and faithfulness require further investigation. Human oversight and domain expertise remain essential for the responsible interpretation and validation of such systems in real-world mental health settings.

5.2 Deployment Efficiency Analysis

Table 4 summarizes the on-device deployment performance of evaluated SLMs. Overall, the LLaMA family consistently achieves the most favorable deployment characteristics. Specifically, LLaMA3.2-1B attains the lowest TTFT across tasks, indicating superior responsiveness and generation latency, while consuming only 1.90 GB of RAM, which is the lowest among all evaluated models. TinyLLaMA-1.1B achieves the highest ITPS, suggesting faster decoding throughput once generation begins; however, its lower OTPS reflects longer overall execution times, limiting its practical advan-

Table 2: Model performance across zero-shot, few-shot, and after fine-tuning. Best results are in bold and second-best results are underlined.

Model	Zero-shot						Few-shot						After Fine-tuning					
	ACC	BACC	F1	AUPRC	Prec.	Rec.	ACC	BACC	F1	AUPRC	Prec.	Rec.	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-1B	0.48	0.43	0.30	0.43	0.27	0.43	0.51	0.47	0.31	0.46	0.28	0.46	0.77	0.69	0.69	0.63	0.72	0.69
LLaMA3.2-3B	0.52	0.50	0.48	0.51	0.49	0.50	0.52	0.51	0.49	0.52	0.51	0.51	0.80	0.74	0.74	0.67	0.77	0.74
TinyLLaMA-1.1B	0.49	0.47	0.45	0.48	0.47	0.48	0.51	0.48	0.47	0.49	0.50	0.49	0.59	0.56	0.54	0.55	0.57	0.57
Qwen2.5-3B	0.60	0.58	0.54	0.53	0.58	0.58	0.64	0.62	0.57	0.56	0.62	0.62	0.74	0.69	0.69	0.63	0.71	0.69
Qwen3-4B	0.49	0.46	0.40	0.45	0.44	0.46	0.50	0.48	0.42	0.47	0.46	0.48	0.75	0.70	0.70	0.64	0.73	0.70
Qwen3.5-4B	0.43	0.44	0.28	0.43	0.25	0.44	0.45	0.46	0.31	0.45	0.27	0.46	0.75	0.70	0.70	0.64	0.73	0.70
Falcon-1.3B	0.49	0.48	0.45	0.48	0.47	0.48	0.51	0.49	0.47	0.49	0.49	0.50	0.59	0.56	0.54	0.56	0.56	0.57
StableLM-3B	0.43	0.47	0.32	0.45	0.43	0.47	0.46	0.50	0.35	0.49	0.46	0.50	0.62	0.60	0.57	0.58	0.59	0.60
Phi4Mini-3.8B	0.56	0.56	0.50	0.51	0.57	0.56	0.60	0.60	0.54	0.54	0.60	0.60	0.77	0.71	0.70	0.64	0.74	0.71
Gemma4-E2B	0.66	0.61	0.58	0.55	0.62	0.61	0.66	0.62	0.59	0.56	0.63	0.61	0.76	0.70	0.71	0.65	0.73	0.71
Gemma4-E4B	**0.69**	**0.65**	**0.63**	**0.59**	**0.67**	**0.65**	**0.70**	**0.66**	**0.64**	**0.60**	**0.68**	**0.66**	**0.82**	**0.76**	**0.76**	**0.70**	**0.79**	**0.76**

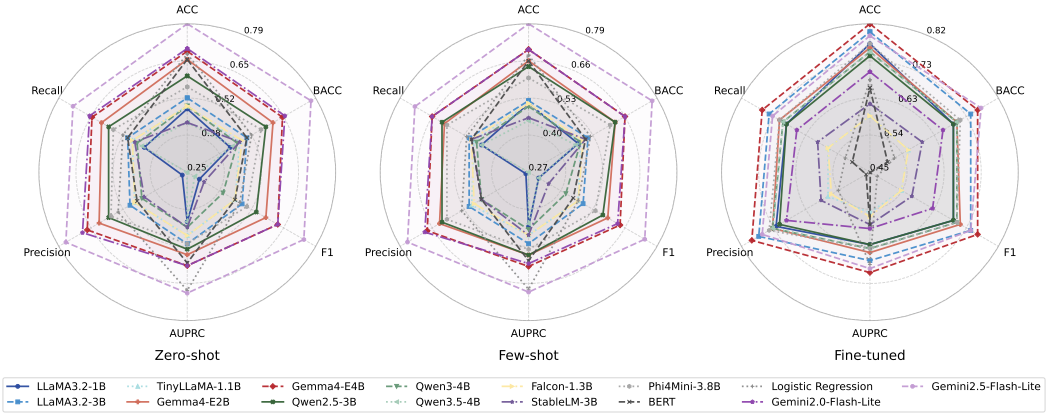


Figure 5: Radar-chart comparison of model performance across evaluation metrics under zero-shot, few-shot, and fine-tuned settings.

tage. We further observe a clear trade-off between deployment efficiency and predictive performance when comparing LLaMA3.2-1B and LLaMA3.2-3B: the larger model improves BACC by 0.05, at the cost of a 0.87,s increase in TTFT and a 2.78s increase in OET. Relative to other SLMs, LLaMA3.2-1B achieves approximately 2–3$\times$ higher ITPS than models such as Qwen2.5-3B while maintaining competitive predictive performance, and runs nearly 2$\times$ faster in total execution time than Gemma4-E4B, whose predictive accuracy is the highest, while requiring roughly one-third of the memory (Figure 6). Overall, LLaMA3.2-1B and LLaMA3.2-3B provide the strongest balance between deployment

Table 3: Performance comparison of BERT, Logistic Regression, and LLM baselines against the best-performing fine-tuned model Gemma4-E4B. Values are reported as mean $\pm$ standard deviation. All standard deviations of Gemini2.0-Flash-Lite, Gemini2.5-Flash-Lite and Gemma4-E4B are smaller than 0.01.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
BERT	0.66$\pm$0.16	0.50$\pm$0.19	0.45$\pm$0.21	0.58$\pm$0.19	0.46$\pm$0.22	0.50$\pm$0.19
Logistic Regression	0.68$\pm$0.14	0.52$\pm$0.17	0.47$\pm$0.19	0.68$\pm$0.16	0.54$\pm$0.22	0.52$\pm$0.17
Gemini2.0-Flash-Lite	0.70	0.66	0.63	0.59	0.69	0.66
Gemini2.5-Flash-Lite	0.79	0.77	0.74	0.69	0.76	0.73
Gemma4-E4B	0.82	0.76	0.76	0.70	0.79	0.76

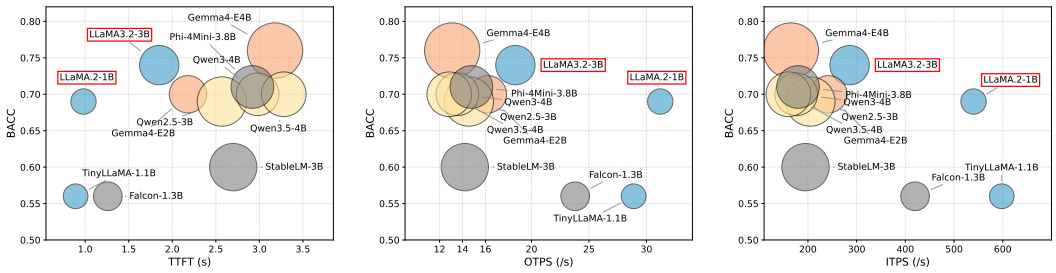


Figure 6: Tradeoff between deployment efficiency and predictive performance across SLMs. Bubble sizes represent RAM usage, while colors indicate different model families.

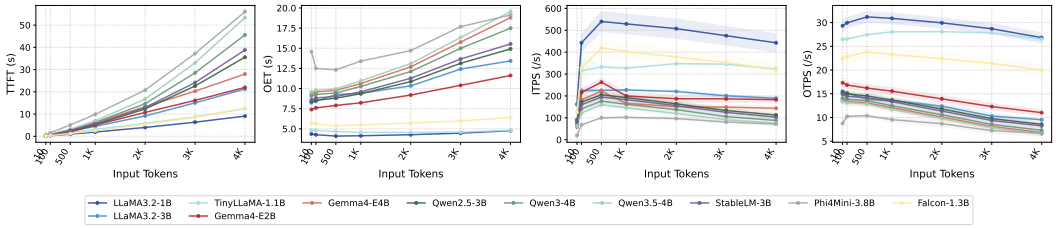


Figure 7: Scaling behavior of SLM deployment efficiency under increasing input token lengths.

efficiency and predictive performance for real-world mental health text detection, with the preferred model depending on latency requirements and available hardware resources.

We also explores impact of input token on the deployment efficiency. Figure 7 shows a consistent trade-off between input length and deployment efficiency across SLMs. As input tokens increase, latency metrics (TTFT and OET) rise substantially, while throughput metrics (ITPS and OTPS) generally decline. Among all models, LLaMA3.2-1B maintains the strongest efficiency profile, achieving the highest ITPS and OTPS with comparatively low latency even under long-context inputs. In contrast, larger models such as Qwen3-4B and gemma4-E4B exhibit steeper degradation in latency and throughput as context length increases. The results further indicate that smaller SLMs scale more robustly under long-input settings, highlighting their suitability for low-latency and resource-constrained on-device deployment scenarios.

We demonstrate mobile deployment of mental health text detection models (Figure 12). Using MentaBench, users can select task-specific models across datasets, while fine-tuned SLMs embedded in lightweight APIs enable low-latency, privacy-preserving on-device inference without external servers. This pipeline helps bridge psychological NLP research and real-world digital mental health applications.

6 Conclusion

In this work, we introduced MentaBench, a unified benchmark for evaluating SLMs on text-based mental health signal detection across diverse datasets, tasks, adaptation settings, and deployment constraints. Through systematic evaluation of 11 SLMs on 21 public datasets and 40 tasks under zero-shot, few-shot, and fine-tuning settings, our results demonstrate that fine-tuned SLMs substantially outperform zero-shot and few-shot configurations, while also exceeding traditional machine learning baselines and achieving competitive performance relative to LLMs. Beyond predictive performance, our deployment-aware analysis shows that several SLMs provide a strong balance between accuracy, latency, throughput, and memory efficiency, highlighting their practical suitability for low-latency and potentially on-device mental health monitoring applications. Prediction errors may have important implications in high-risk mental health settings; real-world intervention and crisis-response integration is beyond our scope and discussed in Appendix Section C. Overall, MentaBench establishes a standardized foundation for future research on scalable, privacy-preserving, and deployment-ready language models for mental health.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Abu Bakar Siddiqur Rahman. Depressionemo. `https://github.com/abuBakarSiddiqurRahman/DepressionEmo`, 2023. Accessed: 2026-05-07.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476, 2016.

Australian Institute of Health and Welfare. Suicide and self-harm. `https://www.aihw.gov.au/mental-health/topic-areas/health-wellbeing/suicide-and-self-harm`, 2024. Accessed: 2026-05-07.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 1644–1650, 2020.

Lily A Brown, Edwin D Boudreaux, Sarah A Arias, Ivan W Miller, Alexis M May, Carlos A Camargo Jr, Craig J Bryan, and Michael F Armey. C-ssrs performance in emergency department patients at high risk for suicide. *Suicide and Life-Threatening Behavior*, 50(6):1097–1104, 2020.

John Bunyi, Kathryn E Ringland, and Stephen M Schueller. Accessibility and digital mental health: considerations for more accessible and equitable mental health apps. *Frontiers in digital health*, 3: 742196, 2021.

H Cai, Y Gao, S Sun, N Li, F Tian, H Xiao, J Li, Z Yang, X Li, Q Zhao, et al. Modma dataset: A multimodal open dataset for mental-disorder analysis. arxiv 2020. *arXiv preprint arXiv:2002.09283*, 2 (3), 2020.

Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43, 2020.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 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, 2015.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pp. 128–137, 2013.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4040–4054, 2020.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

Timon Elmer, Markus Wolf, Evelien Snippe, and Urte Scholz. A social support just-in-time adaptive intervention for individuals with depressive symptoms: feasibility study with a microrandomized trial design. *JMIR mental health*, 12(1):e74103, 2025.

Xiangxiang Gao, Weisheng Xie, Yiwei Xiang, and Feng Ji. Falcon: Faster and parallel inference of large language models through enhanced semi-autoregressive drafting and custom-designed decoding tree. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23933–23941, 2025.

Muskan Garg, Amirmohammad Shahbandegan, Amrit Chadha, and Vijay Mago. An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11960–11969, 2023.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*, pp. 514–525, 2019.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024.

Vatsal Goel, Rishon Dsouza, Nikhil Ega, Amey Ramesh Rambatla, Rob Friel, Shuai Shao, and Yash Sheth. Luna-2: Scalable single-token evaluation with small language models. *arXiv preprint arXiv:2602.18583*, 2026.

Lovedeep Gondara, Jonathan Simkin, Graham Sayle, Shebnum Devji, Gregory Arbour, and Raymond Ng. Small or large? zero-shot or finetuned? guiding language model choice for specialized applications in healthcare. *Machine Learning and Knowledge Extraction*, 7(4):121, 2025.

Google. Gemma 4 e2b instruct. `https://huggingface.co/google/gemma-4-E2B-it`, 2026a. Accessed: 2026-05-07.

Google. Gemma 4 e4b instruct. `https://huggingface.co/google/gemma-4-E4B-it`, 2026b. Accessed: 2026-05-07.

Google DeepMind. Gemini 3.1 pro model card, February 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pp. 3123–3128. Reykjavik, 2014.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. 2025.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yifan Wu, YK Li, et al. Deepseek-coder: when the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

Jessica L Hamilton, Melissa J Dreier, Bianca Caproni, Jennifer Fedor, Krina C Durica, and Carissa A Low. Improving the science of adolescent social media and mental health: Challenges and opportunities of smartphone-based mobile sensing and digital phenotyping. *Journal of Technology in Behavioral Science*, 10(2):301–319, 2025.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. Deep learning for suicide and depression identification with unsupervised label correction. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*, pp. 436–447. Springer, 2021.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pp. 15–24, 2021.

Ian B Hickie, Elizabeth M Scott, Shane P Cross, Frank Iorfino, Tracey A Davenport, Adam J Guastella, Sharon L Naismith, Joanne S Carpenter, Cathrin Rohleder, Jacob J Crouse, et al. Right care, first time: a highly personalised and measurement-based care model to manage youth mental health. *Medical Journal of Australia*, 211:S3–S46, 2019.

Muhammad Hussain, Caikou Chen, Muzammil Hussain, Muhammad Anwar, Mohammed Abaker, Abdelzahir Abdelmaboud, and Iqra Yamin. Optimised knowledge distillation for efficient social media emotion recognition using distilbert and albert. *Scientific Reports*, 15(1):30104, 2025.

Infinite Dataset Hub. Suicideprediction dataset. https://huggingface.co/datasets/infinite-dataset-hub/SuicidePrediction, 2025. Accessed: 2026-05-07.

Frank Iorfino, Vanessa Wan Sze Cheng, Shane P Cross, Hannah F Yee, Tracey A Davenport, Elizabeth M Scott, and Ian B Hickie. Right care, first time: developing a theory-based automated protocol to help clinically stage young people based on severity and persistence of mental illness. *Frontiers in public health*, 9:621862, 2021.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, 2021.

Yu Jin, Jiayi Liu, Pan Li, Baosen Wang, Yangxinyu Yan, Huilin Zhang, Chenhao Ni, Jing Wang, Yi Li, Yajun Bu, et al. The applications of large language models in mental health: Scoping review. *Journal of Medical Internet Research*, 27:e69284, 2025.

kamruzzaman-asif. Reddit mental health classification dataset. https://huggingface.co/datasets/kamruzzaman-asif/reddit-mental-health-classification, 2025. Accessed: 2026-05-07.

S Kayalvizhi, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, et al. Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 331–338, 2022.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, 2017.

Chandreen Liyanage, Muskan Garg, Vijay Mago, and Sunghwan Sohn. Augmenting reddit posts to determine wellness dimensions impacting mental health. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pp. 306–312, 2023.

David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk: early risk prediction on the internet. In *International conference of the cross-language evaluation forum for european languages*, pp. 343–361. Springer, 2018.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024.

Lucidest. reddit-suicidal-classify-kaggle. https://huggingface.co/datasets/Lucidest/reddit-suicidal-classify-kaggle, 2024. Accessed: 2026-05-07.

Md Mafiul Hasan Matin Mafi and Md. Sabbir Alam. Suicidal ideation detection reddit dataset, 2023. URL https://data.mendeley.com/datasets/z8s6w86tr3/2.

Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. Towards privacy-aware mental health ai models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2502.00451*, 2025.

Davide Marengo and Claudio Longobardi. Detecting suicidal ideation in adolescence using self-reported emotional and behavioral patterns: Comparing machine learning and large language model predictions. *Assessment*, pp. 10731911251406405, 2025.

Ritesh Maurya, Nikhil Rajput, MG Diviit, Satyajit Mahapatra, and Manish Kumar Ojha. Exploring the potential of lightweight large language models for ai-based mental health counselling task: a novel comparative study. *Scientific Reports*, 15(1):22463, 2025.

Meta AI. Llama 3.2-1b-instruct. https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct, 2024. Accessed: 2026-05-07.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM web conference 2022*, pp. 2563–2572, 2022.

Ourafla. Mental health text classification dataset. https://huggingface.co/datasets/ourafla/Mental-Health_Text-Classification_Dataset, 2026. Accessed: 2026-05-07.

Xiangdan Piao, Jun Xie, and Shunsuke Managi. Continuous worsening of population emotional stress globally: universality and variations. *BMC Public Health*, 24(1):3576, 2024.

Nikhil Pinnaparaju, Reshinth Adithyan, Duy Phung, Jonathan Tow, James Baicoianu, Ashish Datta, Maksym Zhuravinskyi, Dakota Mahan, Marco Bellagente, Carlos Riquelme, et al. Stable code technical report. *arXiv preprint arXiv:2404.01226*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5370–5381, 2019.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3687–3697, 2018.

Shailik Sarkar, Abdulaziz Alhamadani, Lulwah Alkulaib, and Chang-Tien Lu. Predicting depression and anxiety on reddit: a multi-task learning approach. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 427–435. IEEE, 2022.

SM Sarwar. Fedmentalcare: towards privacy-preserving fine-tuned llms to analyze mental health status using federated learning framework. *arXiv preprint arXiv:2503.05786*, 2025.

MSVPJ Sathvik and Muskan Garg. Multiwd: Multiple wellness dimensions in social media posts. *Authorea Preprints*, 2023.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5263–5276, 2020.

Jiayu Shi, Zexiao Wang, Jiandong Zhou, Chengyu Liu, Poly ZH Sun, Erying Zhao, and Lei Lu. Mentalqlm: A lightweight large language model for mental healthcare based on instruction tuning and dual lora modules. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Debbie L Stoewen. Dimensions of wellness: Change your habits, change your life. *The Canadian veterinary journal*, 58(8):861, 2017.

Syeda Amna. Suicide ideation detection. https://github.com/syeda434am/Suicide-Ideation-Detection, 2023. Accessed: 2026-05-07.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Elsbeth Turcan and Kathleen McKeown. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*, 2019.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, et al. A survey of small language models. *arXiv preprint arXiv:2410.20011*, 2024.

Kimberly A Van Orden, Tracy K Witte, Kelly C Cukrowicz, Scott R Braithwaite, Edward A Selby, and Thomas E Joiner Jr. The interpersonal theory of suicide. *Psychological review*, 117(2):575, 2010.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*, 2024.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87, 2025a.

Xiaofei Wang, Hayley M Sanders, Yuchen Liu, Kennarey Seang, Bach Xuan Tran, Atanas G Atanasov, Yue Qiu, Shenglan Tang, Josip Car, Ya Xing Wang, et al. Chatgpt: promise and challenges for deployment in low-and middle-income countries. *The Lancet Regional Health–Western Pacific*, 41, 2023.

Xin Wang, Ting Dang, Xinyu Zhang, Vassilis Kostakos, Michael J Witbrock, and Hong Jia. Healthslm-bench: Benchmarking small language models for mobile and wearable healthcare monitoring. *arXiv preprint arXiv:2509.07260*, 2025b.

Nicole Davis Weaver, Gregory J Bertolacci, Emily Rosenblad, Sama Ghoba, Matthew Cunningham, Kevin S Ikuta, Madeline E Moberg, Vincent Mougin, Chieh Han, Eve E Wool, et al. Global, regional, and national burden of suicide, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet Public Health*, 10(3):e189–e202, 2025.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. Mentalchat16k: A benchmark dataset for conversational mental health assistance. *arXiv preprint arXiv:2503.13509*, 2025.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 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, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pp. 4489–4500, 2024.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024a.

Tianyi Zhang, Songyan Teng, Hong Jia, and Simon D'Alfonso. Leveraging llms to predict affective states via smartphone sensor features. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 709–716, 2024b.

Xuan Zhao, Niloufar Salehi, Sasha Naranjit, Sara Alwaalan, Stephen Voida, and Dan Cosley. The many faces of facebook: Experiencing social media as performance, exhibition, and personal archive. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1–10, 2013.

Assem Zhunis, Gabriel Lima, Hyeonho Song, Jiyoung Han, and Meeyoung Cha. Emotion bubbles: Emotional composition of online discourse before and after the covid-19 outbreak. In *Proceedings of the ACM Web Conference 2022*, pp. 2603–2613, 2022.

A Datasets and Benchmark-construction

A.1 Benchmark Construction

We define our benchmark-construction criteria as below:

- **Label Interpretability**: a dataset was judged interpretable if the original documentation provided (i) label names, (ii) clear label definitions, and (iii) a clearly documented relationship between text instances and labels that allows the classification task to be reproduced without undocumented assumptions. Datasets failing any of these elements were excluded.
- **Usable sample size**: after removing empty/invalid records and exact duplicates, a dataset was retained only if it contained ≥ 100 labelled examples, sufficient for our predefined training/validation/test splits.
- **Reproducibility**: a dataset was judged reproducible if it could be accessed from the documented source and converted into the same text–label task by following the published procedures (or our documented preprocessing steps) without undocumented ad hoc decisions.

Our manual audit consisted of a structured review of each candidate dataset against a fixed checklist covering text–label mappings, class distributions, duplicate records, label standardisation, and train/validation/test split consistency.

Our primary inclusion criteria were that a dataset: (i) publicly available; (ii) contained textual instances paired with labels relevant to mental health, mental wellbeing, emotions, affective states, stress, suicidal risk, or closely related constructs; (iii) could be formulated as a discrete text-classification task; (iv) provided sufficient documentation to identify the text source, label definitions, and annotation or proxy-labelling procedure; (v) contained enough valid labelled instances to support model training or evaluation, and (vi) has additional diversity in constructs, data sources, annotation paradigms, or task formulations. As such, we exclude many datasets that did not comply with these criteria such as DAIC-WOZ/Extended DAIC (Gratch et al., 2014) and MODMA (Cai et al., 2020).

A.2 Datasets

The datasets included in the MentaBench are introduced as follows:

D01. Dreaddit Turcan & McKeown (2019): Dreaddit is a Reddit-based stress detection dataset constructed from multiple online domains. Each text segment is annotated as either stress or non-stress, supporting binary stress detection from social media text.

D02. DepSeverity Naseem et al. (2022): This dataset is aimed at accurately identifying users' depression severity levels, annotated by CLEF eRisk organizers based on DSAS and clinical standards, categorizing the dataset into four severity levels: Minimal, Mild, Moderate, and Severe.

D03. SDCNL Haque et al. (2021): SDCNL is a Reddit-based dataset collected from communities related to depression and suicidal ideation. Posts from suicide-related communities are treated as suicidal ideation cases, while posts from depression-related communities are treated as non-suicidal ideation cases. The dataset is used for binary suicidal ideation detection.

D04. CSSRS-SuicideGaur et al. (2019): This dataset contains samples related to depressive suicidal ideation and behaviors. Five hundred users were randomly selected, and domain experts annotated the data using the Columbia-Suicide Severity Rating Scale (Brown et al., 2020). The resulting labels fall into five categories: supportive, indicator, ideation, behavior, and attempt. This dataset uses the CC BY 4.0 license.

D05. Reddit Mental Health Classification (kamruzzaman-asif, 2025): A Reddit-based mental health classification dataset distributed through Hugging Face. We use a filtered subset of the dataset and formulate a multi-class classification task over mental-health-related subreddit/category labels, including addiction, anxiety, depression, bipolar disorder, schizophrenia, loneliness, health anxiety, and other support-related categories.

D06. Sentiment Analysis for Mental Health (Ji et al., 2021): A mental health sentiment analysis dataset containing text samples annotated with multiple emotional and psychological state categories for evaluating mental health-related sentiment and emotion classification tasks.

D07. Suicidal Ideation Detection Reddit Dataset (Mafi & Alam, 2023): A Reddit-based suicidal ideation detection dataset in which posts from suicide-related and non-suicide-related subreddits were assigned proxy labels based on subreddit source membership for binary suicide risk classification. This dataset uses the CC BY 4.0 license.

D08. Suicidal Tweet Detection Dataset (Syeda Amna, 2023): A Twitter-based suicidal ideation detection dataset containing tweets labeled as suicidal or non-suicidal for binary suicide risk classification, curated for studying suicide-related language patterns in social media text.

D09. SuicidePrediction Dataset (Infinite Dataset Hub, 2025): A small synthetic suicide-risk prediction dataset containing short text descriptions annotated into three risk levels: High Risk, Moderate Risk, and Low Risk.

D10. (Lucidest, 2024): A Reddit-based suicidal post classification dataset containing posts labeled as suicidal or non-suicidal for binary suicide risk detection using social media text.

D11. (Ourafla, 2026): A mental health text classification dataset containing social media texts annotated across multiple psychological and emotional state categories for multi-class mental health classification tasks. This dataset uses the CC BY 4.0 license.

D12. (Abu Bakar Siddiqur Rahman, 2023): A depression-related emotion dataset containing social media texts annotated with multiple emotional and psychological state labels for fine-grained mental health emotion classification.

D13. (Demszky et al., 2020): A large-scale Reddit emotion dataset released by Google Research, containing comments annotated with 27 emotion categories and neutrality labels for fine-grained emotion understanding tasks. This dataset uses the Apache-2.0 license.

D14. (Rashkin et al., 2019): An empathy-focused conversational dataset released by Facebook Research, containing dialogue responses grounded in emotionally labeled personal situations for empathy and emotional understanding tasks. This dataset uses the CC BY-NC 4.0 license.

D15. (Saravia et al., 2018): An emotion classification dataset containing short texts annotated with six basic emotion categories for evaluating emotional state recognition from natural language.

D16. (Li et al., 2017): A multi-turn dialogue dataset containing daily conversational exchanges annotated with emotion and communication intention labels for dialogue emotion understanding tasks. This dataset uses the CC BY-NC-SA 4.0 license.

D17. (Barbieri et al., 2020): A benchmark suite of Twitter-based social media classification tasks containing manually annotated tweets for sentiment, emotion, irony, offensiveness, and related language understanding tasks.

D18. (Kayalvizhi et al., 2022): A social media depression detection dataset introduced in the LT-EDI 2022 shared task, containing texts annotated with multiple depression severity levels for mental health classification research.

D19. Interpersonal Risk Factors Dataset (Garg et al., 2023): A social media mental health dataset annotated for interpersonal risk factors related to mental disturbance. Following the original dataset, we formulate two binary tasks: thwarted belongingness and perceived burdensomeness.

D20. MultiWD (Sathvik & Garg, 2023): A Reddit-based multi-label wellness dimension dataset. We formulate six binary classification tasks corresponding to spiritual, physical, intellectual, social, vocational, and emotional wellness dimensions. This dataset uses the CC BY 4.0 license.

D21. Wellness Aspect Dataset (Liyanage et al., 2023): A Reddit-based four-class wellness aspect classification dataset for identifying wellness-related dimensions that may affect mental health. The original labels are provided as Aspect 1–4, so we retain these anonymous aspect identifiers rather than assigning unsupported semantic names.

All datasets were processed to retain only the text and task labels required for evaluation, and fields containing usernames, account identifiers, or other metadata were excluded to reduce the risk of sensitive personal information being exposed.

D02 (eRisk) and D04 (CSSRS-Suicide) explicitly report anonymization or pseudonymisation, while others omit direct account metadata or contain synthetic, crowdsourced, or purpose-written text. We did not apply additional formal anonymization and retained only the text and task labels required for evaluation, excluding available usernames and account metadata.

A.3 Task Definitions

The 40 classification tasks included in the MentaBench are defined as follows:

T1. Stress detection (D01, Dreaddit): Predicting each Reddit text segment to be stress or non-stress, indicating whether it reflects current psychological strain rather than merely mentioning stress as a topic.

T2. Depression detection (D02, DepSeverity): Determining whether a post represents depressive patterns. The four DepSeverity grades are collapsed into a binary label: Minimal as no depression, and Mild, Moderate, and Severe as depression.

T3. Depression severity (D02, DepSeverity): Predicting the corresponding depression severity level. Retaining the original four-level scale: Minimal, Mild, Moderate, and Severe.

T4. Suicidal ideation detection (D03, SDCNL): Determining whether a Reddit post expresses suicidal ideation. We use the dataset's released binary labels for this decision.

T5. Suicide-risk detection (D04, CSSRS-Suicide): Determining whether a user's posts indicate suicide risk. The five categories are collapsed into a binary label: Supportive remains Supportive, and Indicator, Ideation, Behavior, and Attempt are treated as risk.

T6. Suicide-risk severity (D04, CSSRS-Suicide): Predicting the suicide-risk category of a user's concatenated posts while retaining the original five labels: Supportive, Indicator, Ideation, Behavior, and Attempt.

T7. Mental-health category classification (D05): Assigning each post to one of 21 mental-health-related categories using the released category labels as a 21-way classification task.

T8. Community classification (D06): Predicting which of five Reddit communities a post comes from: Anxiety, SuicideWatch, Bipolar, Depression, or OffMyChest.

T9. Suicidal ideation detection (D07): Predicting whether a Reddit post indicates suicidal ideation using proxy labels derived from membership in suicide-related or non-suicide-related subreddits.

T10. Suicidal tweet detection (D08): Predicting whether a tweet contains suicidal content, using the released labels Potential Suicide post and Not Suicide post.

T11. Suicide-risk level (D09, SuicidePrediction): Assigning each short text to one of three suicide-risk levels: High Risk, Moderate Risk, or Low Risk.

T12. Suicidal post detection (D10): Predicting whether a Reddit post is suicidal, using the released labels suicidal and not suicidal.

T13. Mental-health category classification (D11): Assigning each text to one of four mental-health categories: Anxiety, Depression, Normal, or Suicidal.

T14. Anger (D12, DepressionEmo): Determining whether anger is present in the text. The Anger label is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T15. Brain dysfunction (D12, DepressionEmo): Determining whether brain dysfunction is present in the text. The released label "brain dysfunction (forget)" is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T16. Emptiness (D12, DepressionEmo): Determining whether emptiness is present in the text. The Emptiness label is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T17. Hopelessness (D12, DepressionEmo): Determining whether hopelessness is present in the text. The Hopelessness label is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T18. Lack of energy (D12, DepressionEmo): Determining whether lack of energy is present in the text. The Lack of energy label is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T19. Loneliness (D12, DepressionEmo): Determining whether loneliness is present in the text. The Loneliness label is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T20. Sadness (D12, DepressionEmo): Determining whether sadness is present in the text. The Sadness label is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T21. Self-hate (D12, DepressionEmo): Determining whether self-hate is present in the text. The Self-hate label is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T22. Suicide intent (D12, DepressionEmo): Determining whether suicide intent is present in the text. The Suicide intent label is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T23. Worthless (D12, DepressionEmo): Determining whether worthlessness is present in the text. The Worthlessness label is treated as a one-versus-rest classification task with Absent and Present as the two classes.

T24. Neutral (D13, GoEmotions): Determining whether a Reddit comment is neutral. The Neutral label is treated as a one-versus-rest classification task with Present and Absent as the two classes.

T25. Empathy (D14): Classifying each dialogue utterance as Low or High empathy, using the listener-empathy rating recorded for that conversation.

T26. Six-way emotion classification (D15): Assigning each short text to one of six emotions: Sadness, Joy, Love, Anger, Fear, or Surprise.

T27. Dialogue emotion classification (D16, DailyDialog): Assigning each dialogue turn to one of seven emotion labels: Anger, Disgust, Fear, Happiness, No emotion, Sadness, or Surprise.

T28. Tweet sentiment classification (D17, TweetEval): Assigning each tweet to one of three sentiment labels: Negative, Neutral, or Positive.

T29. Depression severity (D18): Assigning each social-media text to one of three depression severity levels: not depression, moderate, or severe.

T30. Thwarted belongingness (D19): Determining whether a post expresses thwarted belongingness. The task is treated as a binary classification problem with No and Yes as the two classes.

T31. Perceived burdensomeness (D19): Determining whether a post expresses perceived burdensomeness. The task is treated as a separate binary classification problem with No and Yes as the two classes.

T32. Spiritual wellness (D20, MultiWD): Determining whether the spiritual wellness dimension is present in a Reddit post. The task uses Absent and Present as the two classes.

T33. Physical wellness (D20, MultiWD): Determining whether the physical wellness dimension is present in a Reddit post. The task uses Absent and Present as the two classes.

T34. Intellectual wellness (D20, MultiWD): Determining whether the intellectual wellness dimension is present in a Reddit post. The task uses Absent and Present as the two classes.

T35. Social wellness (D20, MultiWD): Determining whether the social wellness dimension is present in a Reddit post. The task uses Absent and Present as the two classes.

T36. Vocational wellness (D20, MultiWD): Determining whether the vocational wellness dimension is present in a Reddit post. The task uses Absent and Present as the two classes.

T37. Emotional wellness (D20, MultiWD): Determining whether the emotional wellness dimension is present in a Reddit post. The task uses Absent and Present as the two classes.

T38. Wellness aspect classification (D21): Assigning each Reddit post to one of four labels: Aspect 1, Aspect 2, Aspect 3, or Aspect 4.

T39. Sadness (D13, GoEmotions): Determining whether sadness is present in a Reddit comment. The Sadness label is treated as a one-versus-rest classification task with Present and Absent as the two classes.

T40. Disappointed (D13, GoEmotions): Determining whether disappointment is present in a Reddit comment. The Disappointment label is treated as a one-versus-rest classification task with Present and Absent as the two classes.

B Detailed Evaluation Setup

While we report multiple evaluation metrics, we primarily focus on BACC due to substantial label imbalance across datasets and tasks. BACC provides a fairer assessment by averaging class-wise recall rather than being dominated by majority classes:

$$\text{BACC} = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}$$

where C denotes the number of classes, and TP_c and FN_c represent the true positives and false negatives for class c, respectively.

To provide higher-level analysis across heterogeneous datasets and label spaces, we group the 40 tasks into eight broader conceptual categories: *stress*, *interpersonal risk factors (IRF)*, *suicidal risks*, *affective states*, *wellbeing*, *mental health symptoms*, *depression*, and *empathy*. This abstraction is motivated by both psychological relevance and evaluation consistency. We found multiple datasets target closely related latent constructs despite using different label definitions or annotation procedures. For example, suicidal ideation, suicide severity, and suicide risk prediction tasks all aim to identify varying levels of self-harm or suicide-related risk, which we therefore analyse them jointly under *suicidal risks*. Similarly, datasets involving emotional reactions, sentiment, or affective expression are grouped as *affective states*, since they primarily evaluate emotion-related language understanding.

We categorize mental health symptoms into individual types including depression, stress and suicidal risks. Depression reflects a clinically oriented condition and symptom severity construct, whereas stress captures perceived or event-related strain. Suicidal risks are treated separately because suicidal ideation and behaviour represent high-risk outcomes that are related to, but not equivalent to, depression or general distress. Prior work also distinguishes suicide risk from broader mental illness (Australian Institute of Health and Welfare, 2024), noting that mental illness is a risk factor for suicide but does not fully determine suicidal behaviour.

We also separate *interpersonal risk factors (IRF)* from explicit suicidal-risk tasks because constructs such as perceived burdensomeness and thwarted belongingness are theorized as interpersonal vulnerability factors that contribute to suicidal desire, rather than direct expressions of suicidality (Van Orden et al., 2010). Similarly, *affective states* are separated from clinical symptom categories because they capture emotional expression, sentiment, and mood-like states rather than diagnostic or risk labels.

Additionally, *wellbeing* and *empathy* are treated as distinct categories because they reflect broader psychosocial functioning rather than mental illness symptoms. Wellness models commonly distinguish multiple dimensions such as physical, intellectual, emotional, social, spiritual, and vocational wellbeing, supporting our grouping of multidimensional wellness tasks under a separate wellbeing category (Stoewen, 2017). Empathy is also kept separate because it concerns interpersonal understanding and response quality, rather than the detection of distress, symptoms, or risk (Sharma et al., 2020).

Table 4: Deployment-related evaluation metrics across SLMs. Higher ITPS and OTPS indicate better throughput, while lower TTFT and OET indicate lower latency. Throughput is reported as mean, standard deviation (SD), and 95% confidence interval (CI), where available. Best results are in bold and second-best results are underlined.

Model	ITPS Mean ↑	ITPS SD	ITPS 95% CI	OTPS Mean ↑	OTPS SD	OTPS 95% CI	TTFT ↓ (s)	OET ↓ (s)	RAM (GB)	BACC
LLaMA3.2-1B	<u>540.00</u>	45.00	[428.21, 651.79]	**31.20**	1.10	[28.47, 33.93]	<u>0.98</u>	**4.10**	<u>1.90</u>	0.69
LLaMA3.2-3B	285.55	**9.32**	[262.40, 308.70]	18.60	**0.02**	[18.55, 18.65]	1.85	6.88	2.95	<u>0.74</u>
TinyLLaMA-1.1B	**597.96**	193.30	[117.78, 1078.14]	<u>28.89</u>	0.86	[26.75, 31.03]	**0.89**	<u>4.43</u>	**1.85**	0.56
Gemma4-E2B	242.00	18.00	[197.29, 286.71]	16.20	0.75	[14.34, 18.06]	2.18	7.90	2.80	0.70
Gemma4-E4B	165.00	13.00	[132.71, 197.29]	13.10	0.65	[11.49, 14.71]	3.18	9.77	4.15	**0.76**
Qwen2.5-3B	205.00	13.00	[172.71, 237.29]	14.55	0.60	[13.06, 16.04]	2.57	8.80	3.70	0.69
Qwen3-4B	176.00	13.00	[143.71, 208.29]	13.55	0.60	[12.06, 15.04]	2.98	9.45	3.20	0.70
Qwen3.5-4B	160.00	<u>12.50</u>	[128.95, 191.05]	12.85	<u>0.55</u>	[11.48, 14.22]	3.28	9.96	3.35	0.70
Falcon-1.3B	420.00	35.00	[333.06, 506.94]	23.80	1.20	[20.82, 26.78]	1.26	5.38	2.15	0.56
StableLM-3B	195.00	15.00	[157.74, 232.26]	14.20	0.70	[12.46, 15.94]	2.70	9.01	3.55	0.60
Phi4Mini-3.8B	179.73	17.90	[135.26, 224.20]	14.77	0.89	[12.56, 16.98]	2.92	8.67	3.20	0.71

C Additional Experimental Results

C.1 Empirical Results

Model training on task analysis

On-device experiments

We measure latency as $L = t_{\text{end}} - t_{\text{start}}$, latency speedup as $S_{\text{latency}} = L_{\text{baseline}}/L_{\text{model}}$, and throughput speedup as $S_{\text{throughput}} = R_{\text{model}}/R_{\text{baseline}}$, where L denotes latency and R denotes throughput. All on-device measurements are reported as the mean across three independent runs, with the corresponding standard deviations provided in Tables 4, 20, and 21. Representative inference-throughput results, including the mean, standard deviation, and 95% confidence interval for input tokens per second (ITPS) and output tokens per second (OTPS), are reported in Table 4.

Among the fine-tuned models, LLaMA3.2-3B shows great balance bewteen mental health predictions and on-device efficiency, and we present the training loss of the fine tuning process in Figure 8.

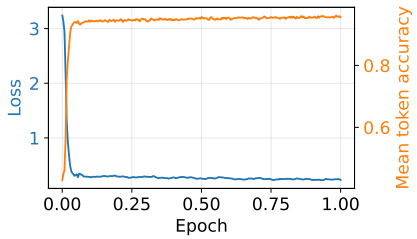


Figure 8: Training and validation performance curves of LLaMA3.2-3B during fine-tuning on MentaBench tasks.

The BACC results of SLMs in zero-shot, few-shot and fine-tuning setting are shown in Figure 9, Figure 10, and Figure 11. The results indicate that zero-shot performance depends strongly on task type. Tasks involving explicit suicidal-risk detection (e.g., Tasks 9, 10, and 12) generally achieve substantially higher BACC across most models. In contrast, several multi-class or affective-state tasks (e.g., Tasks 6, 7, 26–29, and 38) exhibit consistently lower performance across nearly all models, indicating greater difficulty in capturing nuanced emotional, interpersonal, or wellbeing-related constructs without supervised adaptation. Another notable pattern is the relatively stable performance of larger instruction-tuned models across heterogeneous tasks. Gemma-family models show fewer severe performance collapses and maintain moderate-to-high BACC even on challenging datasets, suggesting stronger robustness and prompt-following ability in zero-shot mental health prediction settings. Conversely, some smaller models exhibit highly unstable behavior, alternating between near-random and strong performance depending on the task.

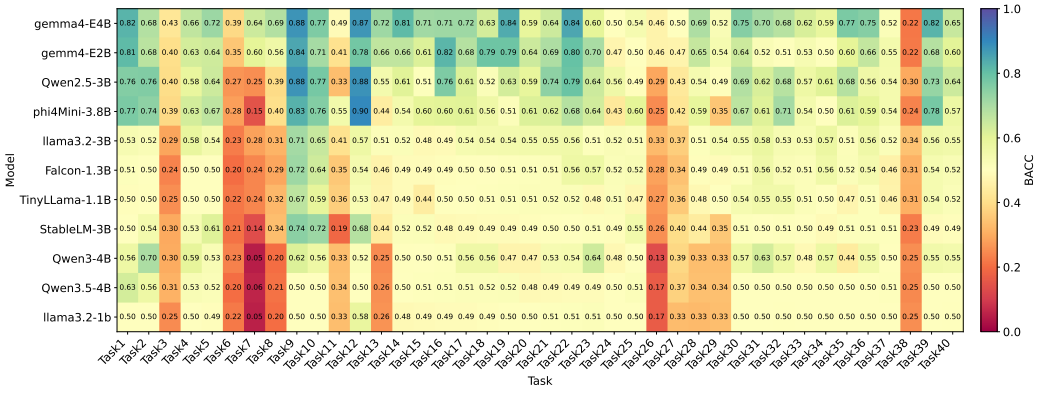


Figure 9: Distribution of BACC achieved by SLMs in the zero-shot setting across the 40 MentaBench tasks.

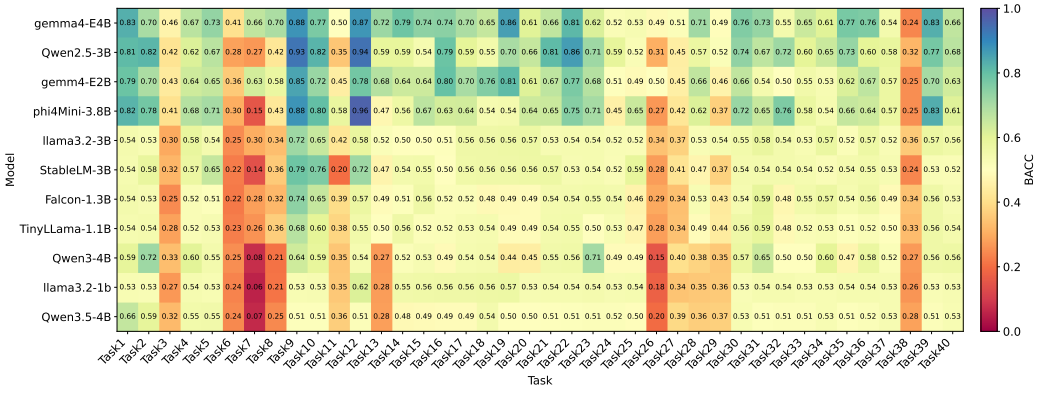


Figure 10: Distribution of BACC achieved by SLMs in the few-shot setting across the 40 MentaBench tasks.

In few-shot setting, the relative ranking among model families in BACC remains largely consistent with the zero-shot setting, that Gemma-family models exhibit the most stable cross-task behavior. Smaller models such as LLaMA3.2-1B and Qwen3.5-4B continue to remain close to chance-level performance on many tasks. The results show that the improvements from few-shot prompting are generally incremental rather than transformative.

Among all evaluated models, Gemma4-E4B achieves the strongest overall predictive performance, followed by LLaMA3.2-3B and phi4Mini-3.8B. The results also show clear variation in task difficulty across datasets and label spaces. In particular, several explicit suicidal-risk detection tasks, especially Tasks 9, 10, and 12, tend to obtain higher BACC values after supervised adaptation, suggesting that suicide-related linguistic signals may be more distinguishable in these benchmark settings. In contrast, several multi-class and affective-state tasks remain more challenging, including Tasks 3, 6, 26–29, and 38, which generally exhibit lower performance across models. These tasks may require more nuanced emotional reasoning, fine-grained semantic discrimination, or understanding of ambiguous psychosocial expressions.

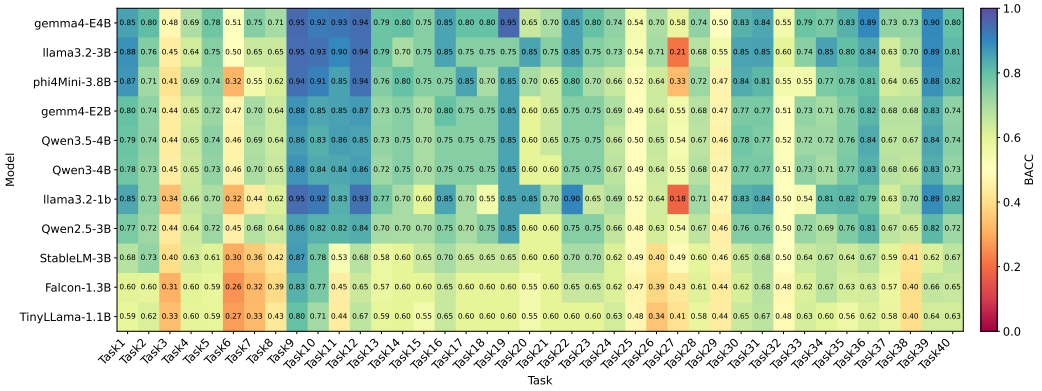


Figure 11: Distribution of BACC achieved by fine-tuned SLMs across the 40 MentaBench tasks.

Comparison among Large Language Models and Larger-Scale Small Language Models

Definitions of SLMs vary across the literature, with some studies using thresholds below 5B parameters (Lu et al., 2024) and others below 7B (Wang et al., 2025a). We therefore selected models in the 1–4B range, which are consistently regarded as small and offer a practical balance among deployment cost, on-device latency, and predictive performance. In our experiments, larger models did not substantially outperform Gemma4-E4B; given comparable accuracy, smaller models are preferable because of their greater efficiency. We additionally evaluated DeepSeek-Coder-6.7B (Guo et al., 2024) and LLaMA 3.1-8B (Grattafiori et al., 2025). As shown in Table 5, our best fine-tuned model consistently outperforms these larger models which require greater computational resources, making them less suitable for on-device deployment.

Table 5: Performance comparison across larger-scale small language models.

Model	ACC	BACC	F1	AUPRC
DeepSeek-Coder-6.7B	0.46	0.44	0.32	0.43
Llama-3.1-8B	0.59	0.58	0.51	0.52
Gemma4-E4B	**0.82**	**0.76**	**0.76**	**0.70**

To include strong fronteier LLM baselines, we further evaluated Gemini 3.1 Pro (Google DeepMind, 2026) across all 40 tasks, with the results shown in Table 6. Despite its greater model capacity, Gemini 3.1 Pro achieved performance broadly comparable to the Gemini Flash models and still underperformed our best fine-tuned SLM. These findings suggest that increasing LLM scale does not necessarily yield better performance for these mental-health classification tasks, whereas fine-tuned SLMs can provide a stronger balance between predictive accuracy and deployment efficiency. Moreover, API-based LLM inference requires sensitive text to be transmitted to external servers,

introducing potential privacy and data-exposure risks. In contrast, SLMs can support local or on-device inference, making them more suitable for privacy-sensitive mental-health applications.

Table 6: Performance comparison across stronger LLM baselines.

Model	ACC	BACC	F1	AUPRC
Gemini 3.1-Pro	0.70	0.67	0.63	0.60
Gemini 2.5-Flash-Lite	0.79	**0.77**	0.74	0.69
Gemma4-E4B	**0.82**	0.76	**0.76**	**0.70**

Prompt Strategies

We designed our prompt based on established mental health principles while considering on-device inference overhead. Compared with the Mental-LLM (Xu et al., 2024) prompting template, our prompt yielded similar results, demonstrating that the evaluated SLMs perform consistently across different prompt formulations. Results are shown in Table 7. Compared with the Mental-LLM prompting template, our prompt is more compact and thus more friendly for on-device settings.

Table 7: Performance comparison using our prompt template and the Mental-LLM prompt template.

Model	Prompt Template	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-3B	Ours	0.52	0.50	0.48	0.51	0.50	0.50
LLaMA3.2-3B	Mental-LLM	0.51	0.51	0.59	0.51	0.49	0.50
Qwen3.5-4B	Ours	0.43	0.44	0.28	0.43	0.25	0.44
Qwen3.5-4B	Mental-LLM	0.44	0.44	0.29	0.44	0.25	0.44
Phi4Mini-3.8B	Ours	0.56	0.56	0.50	0.51	0.57	0.56
Phi4Mini-3.8B	Mental-LLM	0.56	0.56	0.49	0.51	0.58	0.57
Gemma4-E4B	Ours	0.69	0.65	0.63	0.59	0.67	0.65
Gemma4-E4B	Mental-LLM	0.68	0.65	0.62	0.60	0.68	0.65

Few-shot inference and fine-tuning ablation

We conducted experiments with more than three shots and found no additional benefit from providing more examples (Table 8, Table 9, Table 10). Note that providing more examples to language models increases the prompt length and computational overhead, which may limit feasibility for on-device inference. The five-shot results are presented in the table below. Our results show only small differences across the few-shot settings, where the examples were randomly selected. Few-shot prompting provided no noticeable additional benefits over zero-shot prompting. In contrast, fine-tuning substantially improved performance.

Table 8: Performance comparison across models for zero-shot inference.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-3B	0.52	0.50	0.48	0.51	0.49	0.50
Qwen3.5-4B	0.43	0.44	0.28	0.43	0.25	0.44
Phi4Mini-3.8B	0.56	0.56	0.50	0.51	0.57	0.56
Gemma4-E4B	**0.69**	**0.65**	**0.63**	**0.59**	**0.67**	**0.65**

We conducted a training-size ablation using 500 (Table 11), 1000 (Table 12), 2000 (Table 13), and 5000 (Table 14) or more (Table 15) training instances. The results show that performance peaked at 1000 instances and declined when additional training data were used.

Comparison between TF-IDF and another embedding Linguistic Inquiry and Word Count (LIWC)

We conducted an experiment comparing LIWC- and TF-IDF-based solutions using the same logistic regression classifier and observed that LIWC outperform TF-IDF (Table 16). Therefore, TF-IDF has been compared as a baseline in our benchmark.

Dataset imbalance and annotation paradigm analysis

Observing the reasons why LR's relatively high AUPRC related to dataset imbalance, class imbalance likely contributes to the observed results but does not fully explain the case. Using the difference

Table 9: Performance comparison across models for three-shot inference.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-3B	0.52	0.51	0.49	0.52	0.51	0.51
Qwen3.5-4B	0.45	0.46	0.31	0.45	0.27	0.46
Phi4Mini-3.8B	0.60	0.60	0.54	0.54	0.60	0.60
Gemma4-E4B	**0.70**	**0.66**	**0.64**	**0.60**	**0.68**	**0.66**

Table 10: Performance comparison across models for five-shot inference.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-3B	0.51	0.50	0.49	0.52	0.50	0.51
Qwen3.5-4B	0.45	0.45	0.31	0.45	0.27	0.46
Phi4Mini-3.8B	0.59	0.59	0.53	0.53	0.60	0.59
Gemma4-E4B	**0.69**	**0.66**	**0.63**	**0.59**	**0.67**	**0.65**

AUPRC - min(Precision, Recall), we observed that mean gaps of 0.22 for binary classification tasks and 0.19 for multiclass tasks. We examined class imbalance and prediction patterns for T15 (binary) and T6 (five-class) for LR (Table 17 and Table 18). When predictions collapsed to a single class, AUPRC remained relatively high despite low precision and recall. This suggests that the models preserved some ranking ability even when the default decision rule produced imbalanced class assignments.

We also observe a similar pattern for models based on LR, BERT, and LIWC, with relatively high AUPRC despite low precision and recall. This indicates that class imbalance, probability calibration, and multiclass decision rules may contribute to this gap. Decision-threshold tuning may therefore be a useful direction for future analysis.

We investigate the relationship between model performance and annotation paradigms. We observe that proxy-labelled tasks achieved the highest average BACC, F1, precision, and recall, while crowdsourced annotations produced the highest average ACC and a comparable AUPRC (Table 19). Clinical/expert-annotated tasks showed the lowest average threshold-based performance. However, these differences may reflect task difficulty rather than annotation quality alone. In particular, the proxy-labelled group mainly consisted of suicidal-risk and symptom-classification tasks, making it difficult to disentangle the effects of annotation paradigm from task category. We therefore cannot conclude that model performance is directly associated with annotation quality, but this relationship represents an important direction for future investigation.

C.2 Case Study

This section presents representative examples illustrating the qualitative behaviour of fine-tuned SLMs LLaMA3.2-3B. The primary focus of this work is benchmark construction and systematic quantitative evaluation rather than detailed qualitative error analysis, which we leave for future investigation. Accordingly, the examples presented here are intended to provide illustrative insight into model behaviour and the ability of fine-tuned SLMs to generate interpretable and logically grounded reasoning. For each case, we include the original post, model prediction, confidence score, generated reasoning, and a brief qualitative analysis.

Case 1: Non-suicidal financial planning (D08).

Post: "I've been working the last 6 month without pay because of the nature of my job and DHS was being slow to approve my hours... I put 1k into savings. Paid off my credit card. Bought myself the

Table 11: Performance comparison across models with $N = 500$ training samples.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-3B	0.79	0.70	0.70	0.62	0.73	**0.73**
Qwen3.5-4B	0.73	0.69	0.68	0.60	0.68	0.65
Phi3Mini-3.8B	0.74	0.70	0.67	0.63	**0.77**	**0.73**
Gemma4-E4B	**0.80**	**0.75**	**0.72**	**0.69**	**0.77**	**0.73**

Table 12: Performance comparison across models with $N = 1{,}000$ training samples.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-3B	0.80	0.74	0.74	0.67	0.77	0.74
Qwen3.5-4B	0.75	0.70	0.71	0.64	0.73	0.70
Phi3Mini-3.8B	0.77	0.71	0.70	0.64	0.74	0.71
Gemma4-E4B	**0.82**	**0.76**	**0.76**	**0.70**	**0.79**	**0.76**

Table 13: Performance comparison across models with $N = 2{,}000$ training samples.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-3B	0.79	0.72	0.73	0.67	0.74	0.76
Qwen3.5-4B	0.75	0.70	0.71	0.64	0.73	0.69
Phi3Mini-3.8B	0.75	0.70	0.69	0.63	0.72	0.70
Gemma4-E4B	**0.81**	**0.74**	**0.75**	**0.69**	**0.78**	**0.78**

Table 14: Performance comparison across models with $N = 5{,}000$ training samples.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-3B	0.75	0.67	0.67	0.60	0.70	0.68
Qwen3.5-4B	0.69	0.65	0.65	0.58	0.66	0.63
Phi3Mini-3.8B	0.71	0.66	0.64	0.59	0.67	0.65
Gemma4-E4B	**0.76**	**0.71**	**0.69**	**0.65**	**0.73**	**0.70**

Table 15: Performance comparison across models with $N = 10{,}000$ training samples.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
LLaMA3.2-3B	0.72	0.65	0.64	0.47	0.67	**0.65**
Qwen3.5-4B	0.67	0.63	0.61	0.55	0.63	0.60
Phi3Mini-3.8B	0.68	0.62	0.60	0.58	0.62	0.62
Gemma4-E4B	**0.73**	**0.67**	**0.65**	**0.63**	**0.70**	**0.65**

Table 16: Performance comparison between traditional baselines and Gemma4-E4B. Results are reported as mean ± standard deviation.

Model	ACC	BACC	F1	AUPRC	Prec.	Rec.
LIWC+LR	0.63 ± 0.13	0.60 ± 0.13	0.58 ± 0.15	0.62 ± 0.16	0.58 ± 0.14	0.60 ± 0.13
TF-IDF+LR	0.68 ± 0.14	0.52 ± 0.17	0.47 ± 0.19	0.68 ± 0.16	0.54 ± 0.22	0.52 ± 0.17
Gemma4-E4B	$\mathbf{0.82 \pm 0.00}$	$\mathbf{0.76 \pm 0.00}$	$\mathbf{0.76 \pm 0.00}$	$\mathbf{0.70 \pm 0.00}$	$\mathbf{0.79 \pm 0.00}$	$\mathbf{0.76 \pm 0.00}$

Table 17: Class-level precision, recall, class proportion, and prediction proportion for Task 15 using linear regression.

Class	Prec.	Rec.	Class Proportion	Prediction Proportion
Absent	0.61	1.00	60.70%	100.00%
Present	0.00	0.00	39.30%	0.00%
All	0.30	0.50	–	–

Table 18: Class-level precision, recall, class proportion, and prediction proportion for Task 6 using linear regression, showing substantial prediction collapse toward the Ideation class. Three classes received no predictions, illustrating the model's poor handling of the multiclass distribution.

Class	Prec.	Rec.	Class Proportion	Prediction Proportion
Supportive	1.00	0.10	21.40%	2.10%
Indicator	0.00	0.00	20.00%	0.00%
Ideation	0.35	1.00	34.30%	97.90%
Behavior	0.00	0.00	15.00%	0.00%
Attempt	0.00	0.00	9.30%	0.00%
All	0.27	0.22	–	–

Table 19: Performance varied across annotation paradigms. Clinical/expert-annotated tasks showed comparatively lower BACC, F1, precision, and recall, although this category contained only four tasks.

Annotation Paradigm	No. of Tasks	ACC	BACC	F1	AUPRC	Precision	Recall
Proxy labeling	5	0.64	**0.59**	**0.57**	**0.65**	**0.59**	**0.59**
Clinical/Expert annotation	4	0.65	0.36	0.29	0.50	0.25	0.36
Crowdsourced annotation	15	**0.67**	0.56	0.49	0.64	0.47	0.56
Other human annotation	16	0.65	0.45	0.40	0.52	0.46	0.45

desktop I need to continue my education... I'm really proud of how I've handled this money so far, and I'm proud of the financial planning I've been doing."

Prediction: Non-Suicidal

SLM Reasoning: "The decision to classify this individual as Non-Suicidal is based on the fact that they have a stable financial situation, with a steady income, savings, and a plan to continue their education."

In Case 1, the model correctly identifies future-oriented planning, financial stability, and educational goals as protective indicators inconsistent with suicidal ideation. Rather than focusing on prior financial hardship, the model emphasizes adaptive coping and positive functioning signals. This suggests that the SLM's ability distinguish stressful life circumstances from genuine suicide-risk language.

Case 2: Explicit suicidal ideation (D08).

Post: "I am completely willing to end things if she truly no longer loves me... I will not live without her and I will absolutely choose this over life... death is the only thing that brings me peace that I can grab."

Prediction: Suicidal

SLM Reasoning: "The individual is willing to end their life if their ex-girlfriend no longer loves them, indicating a strong emotional attachment and desperation. They mention that death is the only thing that brings them peace, suggesting a sense of hopelessness and despair."

In Case 2, the model accurately captures explicit self-harm intent and severe hopelessness cues. Importantly, the reasoning references semantic meaning rather than isolated keywords, focusing on emotional desperation and life-ending preference statements. This indicates that the model's ability of detecting high-risk suicide signals in emotionally complex narratives.

Case 3: Low suicide risk through protective behavior (D10).

Post: "Participates in regular physical exercise and maintains a healthy diet."

Prediction: Low Risk

SLM Reasoning: "The individual's physical health is a strong indicator of their overall well-being, suggesting a low risk of suicidal thoughts. Regular exercise and a healthy diet contribute to a stable mental state, reducing the likelihood of suicidal ideation."

In Case 3, the model associates exercise and healthy lifestyle behaviors with psychological stability and lower suicide risk. Although the text is extremely short, the prediction remains confident, suggesting that the SLM can leverage concise protective indicators effectively. This demonstrates robustness even under sparse textual input conditions.

C.3 Limitation and Future Direction

Despite the strong performance of SLMs, prediction errors in mental health detection may still have serious consequences, particularly in high-risk settings. Our study does not address how such predictions should be integrated into real-world interventions, which requires careful human oversight and ethical system design.

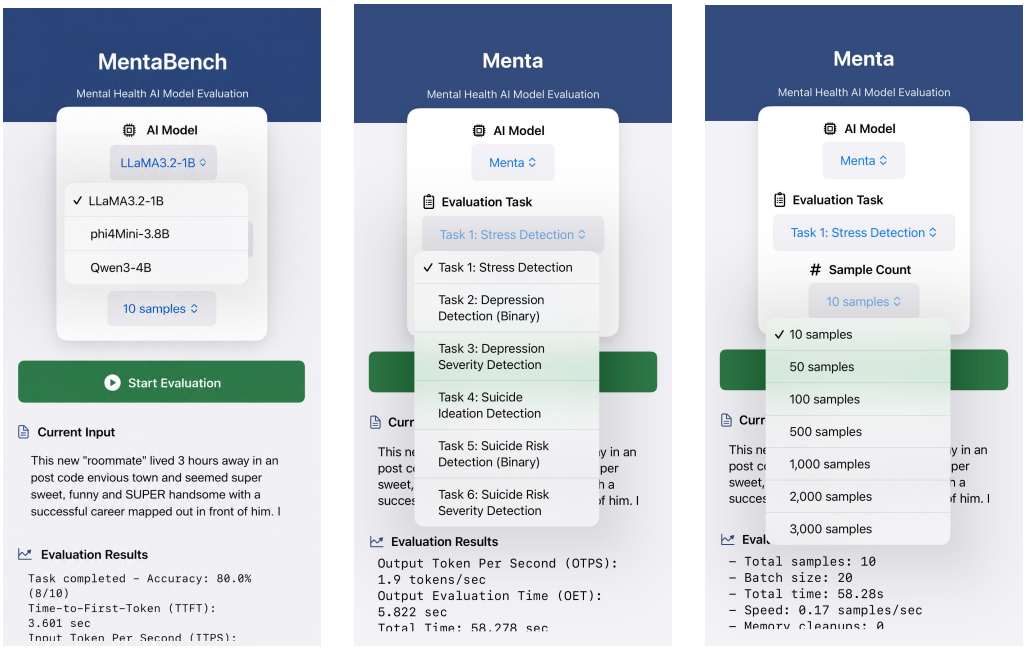


Figure 12: On-device deployment demonstration of the fine-tuned SLMs across 40 tasks. The interface shows input texts and model predictions for model selection (left), task selection (middle), and sample count selection (right).

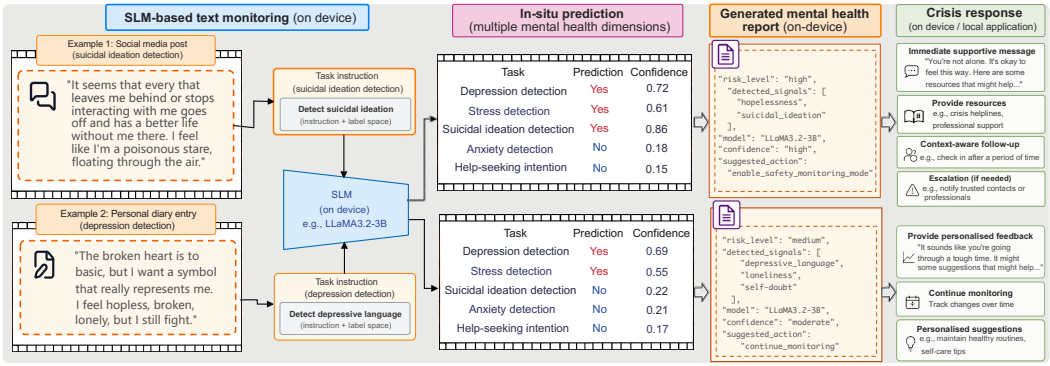


Figure 13: The workflow for on-device SLM-based mental health text monitoring. Incoming text is processed using task-specific instructions to generate predictions across multiple mental health dimensions and structured mental health reports, which may inform downstream monitoring, personalised feedback, or escalation. Downstream intervention and crisis-response components are illustrative and are not evaluated in MentaBench.

Table 20: Deployment metrics across input token lengths (I). ITPS and OTPS are reported as mean ± standard deviation. Higher ITPS and OTPS indicate better performance, whereas lower TTFT, OET, and RAM are preferred. Across models, increasing input length substantially increased TTFT and RAM usage, while output throughput generally declined at longer contexts. LLaMA3.2-1B maintained comparatively high throughput and low latency across input lengths, whereas larger models showed steeper latency increases as context length grew.

Model	Tokens	ITPS ↑	OTPS ↑	TTFT (s) ↓	OET (s) ↓	RAM (GB) ↓
llama3.2-1B	10	162.00 ± 33.75	29.33 ± 1.07	0.10	4.36	1.90
llama3.2-1B	100	442.80 ± 40.75	29.95 ± 1.08	0.26	4.27	1.91
llama3.2-1B	500	540.00 ± 45.00	31.20 ± 1.10	0.96	4.10	1.95
llama3.2-1B	1,000	529.20 ± 44.55	30.89 ± 1.09	1.92	4.14	2.00
llama3.2-1B	2,000	507.60 ± 43.63	29.95 ± 1.08	3.97	4.27	2.10
llama3.2-1B	3,000	475.20 ± 42.21	28.70 ± 1.06	6.35	4.46	2.20
llama3.2-1B	4,000	442.80 ± 40.75	26.83 ± 1.02	9.07	4.77	2.30
llama3.2-3B	10	161.92 ± 2.53	15.54 ± 0.06	0.13	8.24	2.95
llama3.2-3B	100	224.16 ± 12.65	15.21 ± 0.31	0.51	8.42	2.97
llama3.2-3B	500	225.43 ± 4.74	13.83 ± 0.02	2.29	9.26	3.03
llama3.2-3B	1,000	227.45 ± 1.62	13.60 ± 0.25	4.47	9.41	3.10
llama3.2-3B	2,000	220.54 ± 4.78	12.39 ± 0.56	9.15	10.33	3.25
llama3.2-3B	3,000	200.77 ± 6.63	10.31 ± 0.63	15.04	12.42	3.40
llama3.2-3B	4,000	189.10 ± 1.38	9.53 ± 0.08	21.26	13.43	3.55
StableLM-3B	10	81.90 ± 11.25	14.91 ± 0.72	0.19	8.58	3.55
StableLM-3B	100	159.90 ± 13.58	14.48 ± 0.71	0.69	8.84	3.57
StableLM-3B	500	195.00 ± 15.00	14.20 ± 0.70	2.63	9.01	3.63
StableLM-3B	1,000	183.30 ± 14.54	13.35 ± 0.68	5.53	9.59	3.70
StableLM-3B	2,000	156.00 ± 13.42	11.36 ± 0.63	12.91	11.27	3.85
StableLM-3B	3,000	124.80 ± 12.00	9.37 ± 0.57	24.15	13.66	4.00
StableLM-3B	4,000	103.35 ± 11.25	8.24 ± 0.53	38.82	15.53	4.15
gemm4-E2B	10	70.51 ± 13.50	17.33 ± 0.78	0.20	7.39	2.80
gemm4-E2B	100	215.96 ± 16.89	16.85 ± 0.76	0.52	7.60	2.82
gemm4-E2B	500	262.79 ± 18.00	16.20 ± 0.75	1.96	7.90	2.88
gemm4-E2B	1,000	199.94 ± 17.82	15.55 ± 0.73	5.07	8.23	2.95
gemm4-E2B	2,000	187.26 ± 16.89	13.93 ± 0.70	10.75	9.19	3.10
gemm4-E2B	3,000	186.86 ± 15.69	12.31 ± 0.65	16.14	10.40	3.25
gemm4-E2B	4,000	182.76 ± 14.84	11.02 ± 0.62	21.98	11.62	3.40
gemma4-E4B	10	58.40 ± 9.75	13.62 ± 0.66	0.24	9.40	4.15
gemma4-E4B	100	182.50 ± 11.48	13.36 ± 0.66	0.62	9.58	4.18
gemma4-E4B	500	221.70 ± 13.00	13.10 ± 0.65	2.33	9.77	4.25
gemma4-E4B	1,000	166.20 ± 12.47	12.05 ± 0.62	6.10	10.62	4.35
gemma4-E4B	2,000	151.40 ± 11.33	10.09 ± 0.57	13.31	12.69	4.55
gemma4-E4B	3,000	148.80 ± 9.90	8.12 ± 0.51	20.28	15.76	4.75
gemma4-E4B	4,000	143.60 ± 9.75	6.81 ± 0.49	28.00	18.80	4.95

Table 21: Deployment metrics across input token lengths (II). ITPS and OTPS are reported as mean ± standard deviation. Higher ITPS and OTPS indicate better performance, whereas lower TTFT, OET, and RAM are preferred. Similar scaling patterns were observed across the remaining models, with TTFT and RAM generally increasing and OTPS declining as input length grew. TinyLLaMA-1.1B and Falcon-1.3B were comparatively robust to longer inputs, maintaining relatively high throughput and low latency at 4,000 tokens.

Model	Tokens	ITPS ↑	OTPS ↑	TTFT (s) ↓	OET (s) ↓	RAM (GB) ↓
Qwen2.5-3B	10	90.20 ± 9.75	15.28 ± 0.61	0.18	8.38	3.70
Qwen2.5-3B	100	172.20 ± 11.91	14.99 ± 0.61	0.65	8.54	3.72
Qwen2.5-3B	500	205.00 ± 13.00	14.55 ± 0.60	2.51	8.80	3.78
Qwen2.5-3B	1,000	192.70 ± 12.60	13.68 ± 0.58	5.26	9.36	3.85
Qwen2.5-3B	2,000	164.00 ± 11.63	11.79 ± 0.54	12.28	10.86	4.00
Qwen2.5-3B	3,000	133.25 ± 10.48	9.75 ± 0.49	22.62	13.13	4.15
Qwen2.5-3B	4,000	112.75 ± 9.75	8.58 ± 0.46	35.59	14.92	4.30
Qwen3.5-4B	10	62.40 ± 9.38	13.36 ± 0.56	0.24	9.58	3.35
Qwen3.5-4B	100	121.60 ± 10.90	13.11 ± 0.56	0.90	9.76	3.38
Qwen3.5-4B	500	160.00 ± 12.50	12.85 ± 0.55	3.20	9.96	3.45
Qwen3.5-4B	1,000	145.60 ± 11.92	11.69 ± 0.52	6.95	10.95	3.55
Qwen3.5-4B	2,000	120.00 ± 10.83	9.77 ± 0.48	16.77	13.10	3.75
Qwen3.5-4B	3,000	91.20 ± 9.44	7.84 ± 0.43	33.02	16.33	3.95
Qwen3.5-4B	4,000	75.20 ± 9.38	6.55 ± 0.41	53.34	19.54	4.15

Continued on next page

Model	Tokens	ITPS ↑	OTPS ↑	TTFT (s) ↓	OET (s) ↓	RAM (GB) ↓
Qwen3-4B	10	70.40 ± 9.75	14.09 ± 0.61	0.21	9.08	3.20
Qwen3-4B	100	140.80 ± 11.63	13.82 ± 0.61	0.78	9.26	3.23
Qwen3-4B	500	176.00 ± 13.00	13.55 ± 0.60	2.91	9.45	3.30
Qwen3-4B	1,000	161.92 ± 12.47	12.47 ± 0.58	6.26	10.26	3.40
Qwen3-4B	2,000	137.28 ± 11.48	10.57 ± 0.53	14.66	12.11	3.60
Qwen3-4B	3,000	105.60 ± 10.07	8.54 ± 0.48	28.53	14.99	3.80
Qwen3-4B	4,000	88.00 ± 9.75	7.32 ± 0.45	45.59	17.49	4.00
phi4Mini-3.8B	10	19.27 ± 2.57	8.78 ± 0.67	0.63	14.58	3.20
phi4Mini-3.8B	100	67.77 ± 1.00	10.25 ± 0.18	1.57	12.49	3.23
phi4Mini-3.8B	500	100.13 ± 1.09	10.39 ± 0.10	5.09	12.32	3.30
phi4Mini-3.8B	1,000	102.35 ± 4.78	9.55 ± 0.54	9.88	13.40	3.40
phi4Mini-3.8B	2,000	96.78 ± 7.20	8.70 ± 0.70	20.78	14.71	3.60
phi4Mini-3.8B	3,000	80.92 ± 6.80	7.24 ± 0.65	37.21	17.68	3.80
phi4Mini-3.8B	4,000	71.55 ± 6.20	6.69 ± 0.60	56.05	19.13	4.00
TinyLLaMA-1.1B	10	90.71 ± 33.97	26.51 ± 0.27	0.15	4.83	1.85
TinyLLaMA-1.1B	100	315.17 ± 56.06	26.56 ± 0.04	0.35	4.82	1.86
TinyLLaMA-1.1B	500	332.77 ± 11.05	27.45 ± 0.09	1.54	4.66	1.90
TinyLLaMA-1.1B	1,000	326.96 ± 4.81	28.05 ± 0.09	3.09	4.56	1.95
TinyLLaMA-1.1B	2,000	347.43 ± 0.72	28.09 ± 0.06	5.79	4.56	2.05
TinyLLaMA-1.1B	3,000	345.37 ± 1.10	27.85 ± 0.10	8.72	4.60	2.15
TinyLLaMA-1.1B	4,000	322.92 ± 6.54	26.53 ± 0.78	12.42	4.82	2.25
Falcon-1.3B	10	126.00 ± 26.25	22.37 ± 1.16	0.12	5.72	2.15
Falcon-1.3B	100	327.60 ± 30.91	22.85 ± 1.18	0.35	5.60	2.16
Falcon-1.3B	500	420.00 ± 35.00	23.80 ± 1.20	1.23	5.38	2.20
Falcon-1.3B	1,000	403.20 ± 34.29	23.32 ± 1.19	2.52	5.49	2.25
Falcon-1.3B	2,000	378.00 ± 33.20	22.37 ± 1.16	5.34	5.72	2.35
Falcon-1.3B	3,000	352.80 ± 32.08	21.42 ± 1.14	8.55	5.98	2.45
Falcon-1.3B	4,000	319.20 ± 30.51	19.99 ± 1.10	12.58	6.40	2.55

Table 22: Performance of BERT-base across all MentaBench tasks. Differences between ACC and BACC on several tasks further highlight the impact of class imbalance.

Task	Dataset	ACC	BACC	F1	AUPRC	Precision	Recall
T1	D01	0.71	0.70	0.69	0.82	0.74	0.70
T2	D02	0.73	0.50	0.42	0.73	0.36	0.50
T3	D02	0.73	0.25	0.21	0.37	0.18	0.25
T4	D03	0.54	0.54	0.51	0.60	0.56	0.54
T5	D04	0.79	0.50	0.44	0.58	0.39	0.50
T6	D04	0.34	0.20	0.10	0.32	0.07	0.20
T7	D05	0.31	0.05	0.03	0.17	0.06	0.05
T8	D06	0.34	0.21	0.13	0.29	0.34	0.21
T9	D07	0.91	0.91	0.91	0.97	0.91	0.91
T10	D08	0.90	0.90	0.90	0.96	0.90	0.90
T11	D09	0.54	0.55	0.50	0.55	0.58	0.55
T12	D10	0.88	0.88	0.88	0.95	0.88	0.88
T13	D11	0.59	0.44	0.36	0.51	0.45	0.44
T14	D12	0.57	0.47	0.36	0.55	0.30	0.47
T15	D12	0.86	0.50	0.46	0.62	0.43	0.50
T16	D12	0.57	0.47	0.36	0.63	0.30	0.47
T17	D12	0.64	0.50	0.39	0.65	0.32	0.50
T18	D12	0.64	0.50	0.39	0.54	0.32	0.50
T19	D12	0.61	0.50	0.38	0.53	0.30	0.50
T20	D12	0.79	0.50	0.44	0.56	0.39	0.50
T21	D12	0.68	0.67	0.67	0.77	0.68	0.67
T22	D12	0.79	0.50	0.44	0.43	0.39	0.50
T23	D12	0.57	0.50	0.36	0.67	0.29	0.50
T24	D13	0.66	0.66	0.66	0.71	0.66	0.66
T25	D14	0.64	0.52	0.45	0.51	0.51	0.52
T26	D15	0.38	0.19	0.13	0.26	0.16	0.19
T27	D16	0.83	0.14	0.13	0.18	0.12	0.14
T28	D17	0.53	0.41	0.38	0.46	0.35	0.41
T29	D18	0.68	0.38	0.36	0.47	0.41	0.38
T30	D19	0.66	0.64	0.64	0.72	0.65	0.64
T31	D19	0.69	0.53	0.50	0.61	0.60	0.53
T32	D20	0.95	0.50	0.49	0.50	0.47	0.50
T33	D20	0.70	0.50	0.41	0.58	0.35	0.50
T34	D20	0.79	0.50	0.44	0.63	0.40	0.50
T35	D20	0.65	0.52	0.46	0.63	0.64	0.52
T36	D20	0.85	0.50	0.46	0.58	0.43	0.50
T37	D20	0.51	0.51	0.45	0.55	0.52	0.51
T38	D21	0.40	0.29	0.20	0.38	0.55	0.29
T39	D13	0.72	0.72	0.71	0.80	0.73	0.72
T40	D13	0.64	0.64	0.63	0.69	0.67	0.64