DOI: **10.37943/XSAJ1029**

Dastan Kapizov
PhD Student, School of Software Engineering
242781@astanait.edu.kz, orcid.org/0009-0005-3817-3614
Astana IT University, Kazakhstan
Praveen Kumar
PhD, Associate Professor, School of Software Engineering
Praveen.Kumar@astanait.edu.kz, orcid.org/0000-0002-9606-8960
Astana IT University, Kazakhstan

FINE-TUNING SMALL LANGUAGE MODELS FOR MENTAL HEALTH TEXT CLASSIFICATION: A PARAMETER-EFFICIENT APPROACH OUTPERFORMING ZERO-SHOT LARGE LANGUAGE MODELS

Abstract: Automated analysis of user-generated text is increasingly applied to scalable mental-health assessment, but large language models used for classification are expensive, latency-sensitive and raise data-governance concerns. We investigate whether parameter-efficient fine-tuning (PEFT) of a small model provides a better accuracy-efficiency trade-off. We fine-tune SmolLM2-1.7B with Low-Rank Adaptation on two Reddit-based benchmarks: SWMH five-class mental-health classification (34,823 training and 10,883 test posts) and Dreaddit binary stress detection (2,838 training and 715 test posts). Only 8.4 million parameters (about 0.5% of the model) are updated, and training completes on one NVIDIA A100 GPU. Evaluation is performed under two protocols: 1) primary, fully held-out test evaluation of the fine-tuned model, and 2) paired evaluation on shared 200-example subsets (to enable cross-model comparison with zero-shot GPT-4o-mini and GPT-4o). With length-normalized label log-likelihood scoring, the full held-out tests give 0.7056 accuracy, 0.7072 weighted F1 and 0.7207 macro F1 on SWMH and 0.8056 accuracy, 0.8048 weighted F1 and 0.8043 macro F1 on Dreaddit. Calibration analysis yielded an Expected Calibration Error (ECE) of 0.1826 and a multiclass Brier score of 0.4667 on SWMH, and 0.1757 and 0.1701 on Dreaddit. On paired comparisons, 0.6600 accuracy was seen for SmolLM2-1.7B on SWMH against 0.6250 and 0.6300 for GPT-4o-mini and GPT-4o, and 0.8000 on Dreaddit against 0.6500 and 0.6750, respectively. McNemar's exact test showed that the paired advantage was statistically significant on Dreaddit but not on SWMH. These results indicate that task-specific PEFT can make small models practically deployable and competitive with larger zero-shot models, with the strongest evidence of advantage on binary stress detection; the imperfect calibration means the outputs should support decisions rather than serve as clinical risk estimates.

Keywords: mental health classification, small language models, parameter-efficient fine-tuning, LoRA, text analysis, stress detection, depression detection, zero-shot learning

Introduction

Around the World more than 970 million people are affected by mental health disorders, with anxiety and depression accounting for a major share of the world disease burden, according to the World Health Organization [1]. But for timely support we need early detection but conventional screening is constrained by stigma, lack of specialist contact and late help seeking behaviour [2]. Online communities at the same time contain large quantities of self-reported distress signals, coping narratives and symptom language that are useful for scalable monitoring of mental-health risk [3].

Newer developments in NLP and large language models like GPT-4 allow better text understanding in zero-shot/few-shot situations [4]. In practice though, large proprietary models

Received: 26.02.2026 Accepted: 17.07.2026 Published: 30.06.2026

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

are not easily deployed for mental-health screening. Main barriers are high inference cost/latency sensitivity in real time workflows and governance issues when user text is sent to external services [5]. All of these constraints make operational feasibility unviable even at high raw model capability.

Small language models (SLMs) generally under 3B parameters, are a more favourable profile of deployment for specialized classification pipelines [6]. With parameter-efficient fine-tuning (PEFT) they are adapted to domain tasks without updating all model weights [7]. In particular, Low-Rank Adaptation (LoRA) updates a small subset of parameters which requires little compute and memory while maintaining good task performance [8].

Controlled evidence is scarce in two areas of increasing interest – mental-health NLP and efficient adaptation. First of all, few studies directly compare tuned SLMs with zero-shot larger models under matched evaluation conditions [9]. Second, much of the prior work was binary detection oriented whereas multiclass mental-health categorization is clinically informative but methodologically harder [10].

Herein we fill those gaps by evaluating PEFT of SmolLM2-1.7B on two benchmarks: SWMH (five class mental-health classification) and Dreaddit (binary stress detection). We have two very explicit protocols: 1) a full held-out test evaluation as the main outcome for the fine-tuned model and 2) paired evaluation on identical subsets for fair cross-model comparison with zero-shot GPT baselines. This separates deployment-relevant model quality from protocol mixing in comparative claims.

Literature review and problem statement

Now researchers look at social media platforms like Reddit, Twitter (X), and Facebook because people post thoughts about moods, inner struggles, or requests for help [11]. First analysts used word lists alongside standard classifiers to find speech traits, check emotional tone, and look for recurring subjects. As better ways of encoding meaning appeared, stronger network designs such as recurrent neural networks / convolutional neural networks / systems built on transformers were developed [12]. Of these, tools based on BERT and similar versions stand out for detecting depression signs [13], detecting self-harm tendencies [14], and separating tension/pressure [15].

Instruction-tuned large language models are new tools for analysis of mental health texts. But even LLMs like GPT-3.5 and GPT-4 give us nice replies in therapy-like exchanges but they give false information and cause safety issues [16]. Classification with zero-shot/few-shot prompts requires no training data [17], but tests are weak. Results are wildly different depending on how prompts are written [18]. Performance differs between different mental health diagnoses, where meanings are blurred [19]. Closely related clinical labels often trip these systems up. Large models also use lots of computing power - so responses are sluggish. That delays us when we need rapid widespread use.

In order to ease computational demands we have parameter-efficient tuning techniques to tune models to domains which do not require high resource use [20]. Rather than updating all weights, Low-Rank Adaptation places compact matrix pairs in transformer layers and ignores the original network. This saves both memory use and training time [8]. More recent versions, for example QLoRA, use compressed, quantized parameters together with low-rank updates [21]. Studies actually show that even smaller models tuned with LoRA outperform much larger models when applied to very specific tasks [22].

Various benchmark collections are available for testing methods, but often there is no structure for comparison. Signs of stress from Reddit communities are labeled in one such collection called "Dreaddit" [15]; Another, called SWMH, groups entries about psychological health into anxiety, mood changes, sadness, broad emotional strain, and thoughts about self-harm [23]. Even with all these resources, studies seldom group results neatly in binary and

DOI: 10.37943/XSAJ1029

multi-class tasks or pair finely tuned small models with zero-shot large models in a single framework.

And yet, one big question remains unanswered: whether parameter-efficient fine-tuning of small language models is as good as large ones without training examples. This is important when sorting mental health texts into different categories. Here, success would also bring speed, consistency, and real-world use.

The aim of the study

The aim of this study is to determine whether parameter-efficient fine-tuning of a small language model can provide accurate and computationally efficient mental-health text classification, and whether such a model can outperform zero-shot larger language models under matched evaluation conditions.

The objectives of the study

Herein, to meet this end the study fulfils the following objectives. (1) To adapt SmolLM2-1.7B for mental health text classification through Low-Rank Adaptation on multiclass and binary benchmarks. (2) To compare this adapted model to zero-shot GPT baselines on held-out test sets and paired subsets. (3) To characterise the model outside aggregate accuracy via class-wise performance, calibration and uncertainty-aware decision support. (4) To evaluate efficiency and deployment feasibility for resource-constrained deployment.

Methods and Materials

Datasets

Two publicly available Reddit-based datasets were used to test mental-health text classification in complementary settings. As a five-class task, the SWMH dataset uses posts from mental-health related subreddits r/Anxiety, r/Bipolar, r/Depression, r/OffMyChest, and r/SuicideWatch with labels Anxiety, Bipolar, Depression, OffMyChest, and SuicideWatch. In this work, SWMH was applied to predefined train, validation, and test partitions of 34,823, 8,706, and 10,883 samples. In the Dreaddit dataset, we have a binary stress classification task (Stressed versus Non-Stressed) with 2,838 training and 715 test samples. The Dreaddit training split was 2,554 training/284 validation samples for model selection - the official 715-sample test split was reserved for final evaluation. As a consequence, all reported full-test results are computed only on held-out test data and no test set is used during tuning. The two datasets also differ in structure: SWMH is class-imbalanced and reflects broader mental health discourse categories, whereas Dreaddit is more balanced around stress detection. Fig. 1 summarizes the exploratory analysis of the two corpora: the panels show the class distribution of each dataset together with the distribution of post lengths, which is visibly shorter for Dreaddit (median text length: 421 characters) than for SWMH.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

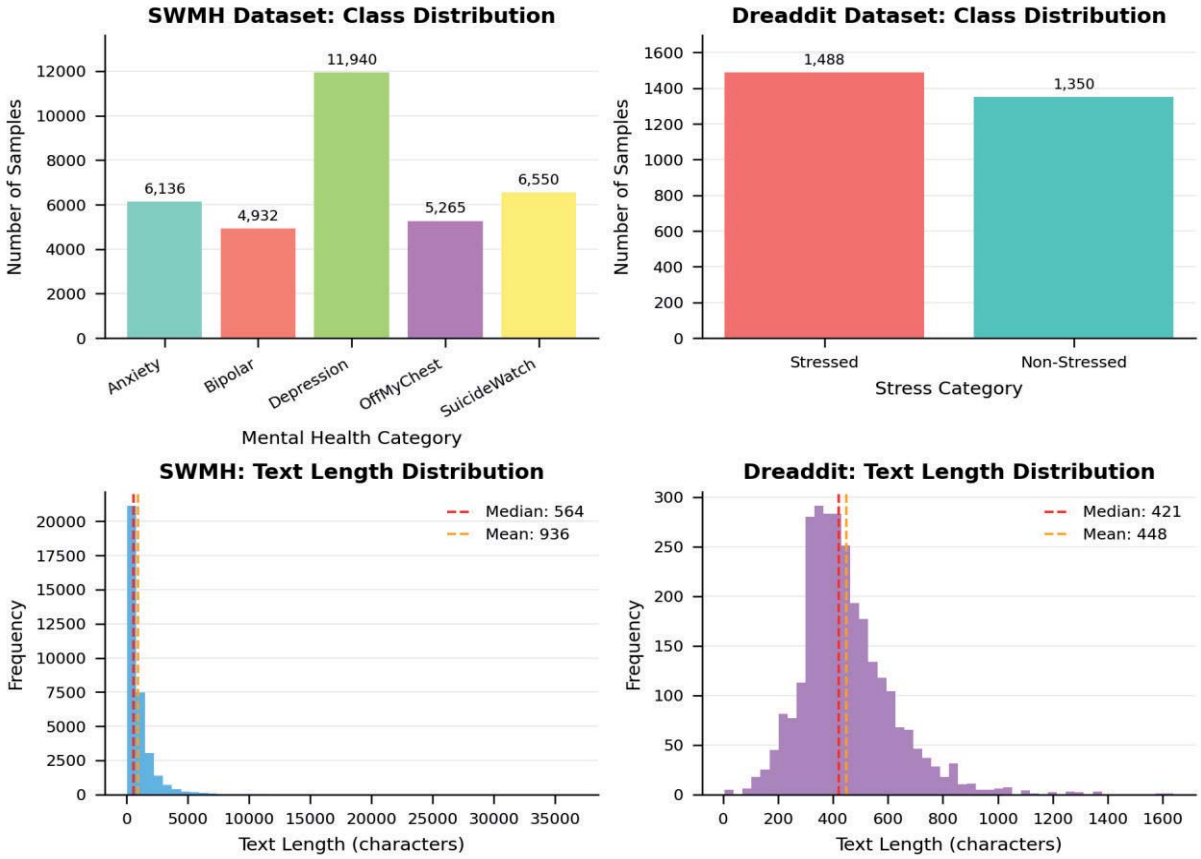

Figure 1. Exploratory data analysis of SWMH and Dreaddit datasets

Dataset labels are treated as benchmark annotations for NLP classification and not as clinical diagnoses.

Model architecture and configuration

We used HuggingFaceTB/SmolLM2-1.7B-Instruct as the base model. SmolLM2 is a decoder-only transformer with about 1.7B params, 32 transformer layers, hidden size 2048, 32 attention heads. To standardise conditions and put computational limits across experiments, all input were tokenized at max sequence length of 512 tokens (truncation where needed). Adaptation was carried out with LoRA rather than full-parameter fine-tuning: modules were added to the query and value attention projections within every layer, base-model weights were fixed. LoRA config was rank r=16, scaling a=32, dropout=0.05, which left us with roughly 8.4M trainable parameters (about 0.5% of param space) - sufficient to maintain adaptation, but substantially reduces the memory/compute burden as compared to full-parameter fine-This kept adaptation capacity with significantly lower memory/compute footprint than full-model fine-tuning.

Mathematical modeling of the classification framework

Let x denote an input text sample. After preprocessing and truncation, the text is inserted into a task-specific prompt $P(x)$. The tokenizer converts this prompt into a sequence of tokens:

$$z = (z_1, z_2, \dots, z_T). \quad (1)$$

Each token is mapped to an embedding vector, and the decoder-only transformer processes the sequence through L layers:

DOI: 10.37943/XSAJ1029

$$H^{(0)} = E(z) + U, \tag{2}$$

$$H^{(\ell)} = Transformer_{\ell}\left(H^{(\ell-1)}\right), \quad \ell = 1, \ldots, L. \tag{3}$$

Here, $E(z)$ denotes token embeddings and U denotes positional embeddings. In this study, the base model is SmolLM2-1.7B with 32 transformer layers, hidden size 2048, and 32 attention heads. The overall data flow is as follows: the raw Reddit post is cleaned and truncated, inserted into the prompt, tokenized, transformed into embeddings, passed through the decoder-only transformer, and then scored against candidate label tokens to obtain the final predicted class. To adapt the model efficiently, Low-Rank Adaptation is applied while the original pretrained weights remain frozen. For a pretrained weight matrix W_0, the adapted matrix is defined as

$$W = W_0 + \Delta W, \tag{4}$$

$$\Delta W = \frac{\alpha}{r} BA, \tag{5}$$

where A and B are trainable low-rank matrices, r is the adaptation rank, and α is the scaling factor. In this study, $r = 16$, $\alpha = 32$, and dropout $= 0.05$, which results in approximately 8.4 million trainable parameters, or about 0.5% of the full model parameters.

For classification, the model does not use a separate discriminative classifier head. Instead, each candidate label c is scored by its conditional log-likelihood given the prompt. If the label c consists of the token sequence $\left(\ell_{c,1}, \ell_{c,2}, \ldots, \ell_{c,M_c}\right)$, then its score is

$$s(c \mid x) = \sum_{j=1}^{M_c} \quad logp\left(\ell_{c,j} \mid P(x), \ell_{c,<j}\right). \tag{6}$$

Because labels may have different token lengths, length normalization is applied:

$$s_{norm}(c \mid x) = \frac{1}{M_c} \sum_{j=1}^{M_c} \quad logp\left(\ell_{c,j} \mid P(x), \ell_{c,<j}\right). \tag{7}$$

The normalized scores are converted into class probabilities using the softmax function:

$$p(c \mid x) = \frac{exp\left(s_{norm}(c|x)\right)}{\sum_{c' \in C} \quad exp\left(s_{norm}(c'|x)\right)}. \tag{8}$$

The final predicted class is then

$$\hat{y} = argmax_{c \in C} p(c \mid x). \tag{9}$$

This formulation makes the decision rule explicit and shows how the generative language model is transformed into a classifier through candidate-label scoring and normalized probabilistic selection.

During fine-tuning, the trainable parameters are optimized by minimizing the cross-entropy loss over the gold labels:

$$L = -\sum_{i=1}^{N} \quad logp(y_i \mid x_i). \tag{10}$$

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

Thus, the model is trained to assign the highest probability to the correct class among the predefined candidate labels.

Training protocol

Minimal preprocessing was done to maintain social-media language characteristics. For SWMH, subreddit-style label prefixes were removed, and input posts were truncated to 1500 chars before tokenization, with no other normalization such as capitalization, punctuation, or slang-word. For each task, task-specific ChatML prompt templates were used to maintain an instruction-following interface as with SmolLM2 and restrict outputs to correct label sets for both tasks (5 classes: SWMH; 2 classes: Dreaddit). Fine-tuning was performed in TRL SFTrainer on a single NVIDIA A100 (80 GB) using bfloat16 when available. Training used a maximum sequence length of 512 tokens, per-device batch size 16, gradient accumulation 2 (effective batch size 32), learning rate 2×10^{-4}, warmup ratio 0.1, cosine learning-rate scheduling, and epoch-wise evaluation with best_checkpoint, and epoch size of 3 (SWMH) and 5 (Dreaddit). The reported logged training runs are for SWMH: 301.2s (≈5.0 minutes), Dreaddit: 81.8 minutes. Training dynamics are shown in Figure 4 and 5.

Zero-shot baseline protocol

For external baselines, GPT-4o-mini and GPT-4o were evaluated using the OpenAI API with task-specific label-constrained prompts and deterministic decoding (temperature = 0). Because it was prohibitive to evaluate APIs on full test sets, baseline scoring was done on fixed random subsets of 200 test instances per dataset. Sampled indices with a fixed seed were generated once for each compared model in the paired protocol and were reused for all compared models. To preserve comparability, label mapping and decision rules must be identical across models. That design allows controlled relative comparison but does not substitute for full test evaluation.

Evaluation protocol

There were two evaluation protocols. Primary protocol: Full-held-out test evaluation of fine-tuned SmolLM2-1.7B on SWMH and Dreaddit. Secondary protocol: Pair-matched sample evaluation (N = 200 per dataset) of SmolLM2-1.7B, GPT-4o-mini, and GPT-4o on identically sampled random subsets. All results are reported as accuracy, weighted F1, and macro F1. All superiority claims over zero-shot GPT baselines are solely based on the paired protocol. Full test results appear as standalone performance evidence.

Evaluation metrics and statistical analysis

Model performance was measured in terms of accuracy, weighted precision, weighted recall, and weighted F1-score. For class-level analysis, precision/recall, F1-score, support, and confusion matrices were also analyzed. Expected Calibration Error and Brier score were also used to measure the probabilistic quality. 95% bootstrap confidence intervals were computed for reported metrics based on resampled hold-out test examples to quantify uncertainty in the reported metrics. The exact test was used for paired comparisons of SmolLM2-1.7B with zero-shot GPT baselines on identical subsets if the difference in paired classification correctness was statistically significant. Since repeated fine-tuning runs with different random seeds are not yet included in the current revision, reported uncertainty intervals are sampling variability over test instances and not variance across training seeds.

The main metrics were computed as follows. Accuracy is defined as

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \quad 1(\hat{y}_i = y_i). \tag{11}$$

For a class k, precision and recall are defined as

$$Precision_k = \frac{TP_k}{TP_k + FP_k}, \tag{12}$$

DOI: 10.37943/XSAJ1029

$$Recall_k = \frac{TP_k}{TP_k + FN_k}. \quad (13)$$

The F1-score for class k is

$$F1_k = \frac{2 \cdot Precision_k \cdot Recall_k}{Precision_k + Recall_k}. \quad (14)$$

Macro F1 is the unweighted average across all classes:

$$F1_{macro} = \frac{1}{K} \sum_{k=1}^{K} F1_k. \quad (15)$$

Weighted F1 is computed as

$$F1_{weighted} = \sum_{k=1}^{K} \frac{n_k}{N} F1_k, \quad (16)$$

where n_k is the number of examples in class k. Expected Calibration Error is defined as

$$ECE = \sum_{b=1}^{B} \frac{|B_b|}{N} |acc(B_b) - conf(B_b)|. \quad (17)$$

Here, B_b is the set of predictions in confidence bin b, $acc(B_b)$ is the empirical accuracy in that bin, and $conf(B_b)$ is the mean predicted confidence in that bin.

For the binary task, the Brier score is

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2. \quad (18)$$

For the multiclass task, the Brier score is

$$BS = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{K} \left(p_{i,c} - y_{i,c}\right)^2, \quad (19)$$

where $y_{i,c}$ is the one-hot encoded target value for class c.

For paired significance testing, McNemar's exact test is based on discordant prediction pairs:

$$b = \#\{SmolLM\ correct, GPT\ wrong\}, \quad (20)$$

$$c = \#\{SmolLM\ wrong, GPT\ correct\}. \quad (21)$$

Under the null hypothesis of equal paired accuracy, the discordant outcomes follow a binomial distribution:

$$b \sim Binomial(b + c, 0.5). \quad (22)$$

A two-sided exact p-value is then computed to determine whether the difference between the two classifiers is statistically significant.

Results

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

In Table 1, we report the full held-out test performance for the finely tuned SmolLM2-1.7B model on both datasets under length-normalized label log-likelihood scoring. For SWMH, the model was 0.7056, the weighted F1 was 0.7072, and the macro F1 was 0.7207. On the model, the accuracy was 0.8056 with a weighted F1 score of 0.8048 and a macro F1 score of 0.8043. Figure 2 visually represents these full aggregate metrics. Zero-shot GPT baselines are not compared with these full-test values as they were assessed under a separate paired protocol on matched subsets. Those results are reported later.

Table 1. Overall performance metrics of fine-tuned SmolLM2-1.7B

Model	Dataset	Accuracy	F1 (macro)	F1 (weighted)
SmolLM2-1.7B (Fine-tuned)	SWMH	0.7056	0.7207	0.7072
SmolLM2-1.7B (Fine-tuned)	Dreaddit	0.8056	0.8043	0.8048

Class-wise full-test behavior is summarized in Fig. 3 and detailed in Tables 2 and 3. On SWMH (n = 10,883), the strongest class F1 values are Anxiety (precision = 0.85, recall = 0.79, F1 = 0.82; n = 1,911) and Bipolar (precision = 0.81, recall = 0.79, F1 = 0.80; n = 1,493). Depression shows relatively high precision but limited recall (precision = 0.76, recall = 0.57, F1 = 0.65; n = 3,774), indicating under-detection of this class. OffMyChest shows the opposite pattern, with low precision and high recall (precision = 0.52, recall = 0.81, F1 = 0.63; n = 1,687), indicating over-assignment to this category. SuicideWatch is more balanced than in the earlier analysis (precision = 0.66, recall = 0.73, F1 = 0.70; n = 2,018). On Dreaddit (n = 715), class performance is also relatively balanced. Non-Stressed reaches precision = 0.83, recall = 0.75, and F1 = 0.79 (n = 346), while Stressed reaches precision = 0.78, recall = 0.86, and F1 = 0.82 (n = 369). This pattern suggests stable binary discrimination with recall favoring the Stressed class.

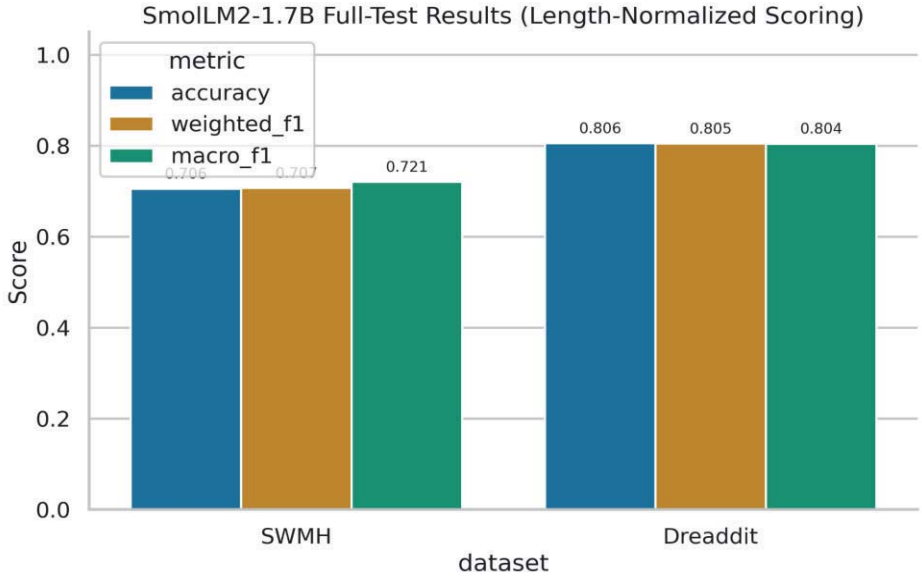

Figure 2. SmolLM2-1.7B full-test performance on SWMH and Dreaddit (accuracy, weighted F1, macro F1)

DOI: 10.37943/XSAJ1029

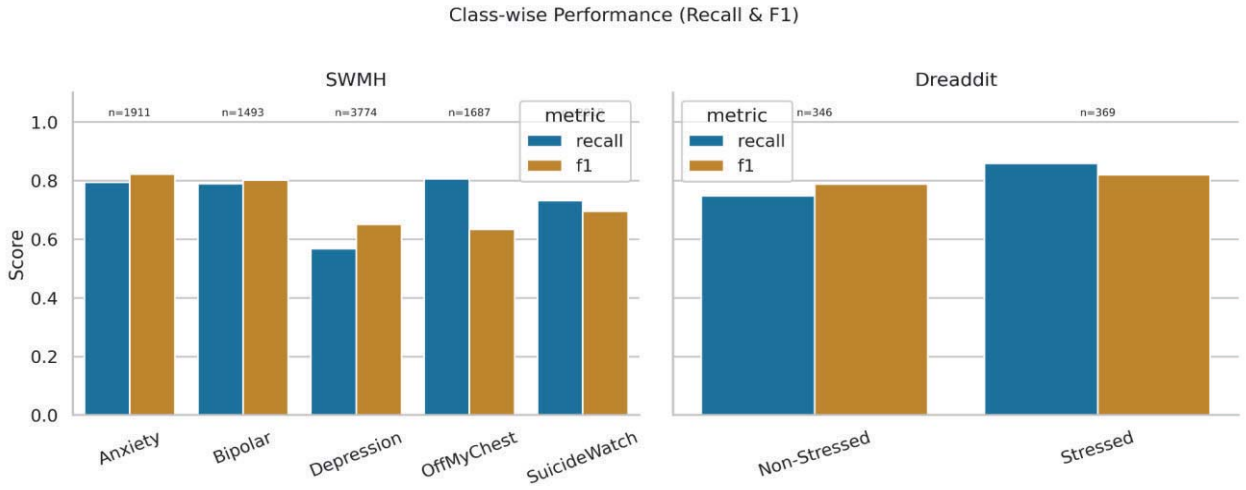

Figure 3. Class-wise full-test performance (Recall and F1) for SWMH and Dreaddit.

Table 2. SWMH full-test class-wise metrics

Category	Precision	Recall	F1-Score	Support
Anxiety	0.85	0.79	0.82	1911
Bipolar	0.81	0.79	0.80	1493
Depression	0.76	0.57	0.65	3774
OffMyChest	0.52	0.81	0.63	1687
SuicideWatch	0.66	0.73	0.70	2018

Table 3. Dreaddit full-test class-wise metrics

Category	Precision	Recall	F1-Score	Support
Non-Stressed	**0.83**	**0.75**	**0.79**	**346**
Stressed	0.78	0.86	0.82	369

Computational efficiency and convergence behavior are reported in Table 4 and Fig. 4-5. Fine-tuning required approximately 81.8 minutes for SWMH (3 epochs) and 301.2 seconds (approximately 5.0 minutes) for Dreaddit (5 epochs) on one NVIDIA A100 (80 GB), with approximately 8.4M trainable parameters (~0.5% of model parameters). SWMH reached its best validation loss at epoch 2 (1.8776), and Dreaddit reached its best validation loss at epoch 4 (2.1086), after which improvements plateaued, supporting best-checkpoint selection.

Table 4. Training time and computational requirements

Dataset	Training Samples	Epochs	Training Time	GPU Memory	Trainable Parameters
SWMH	34, 823	3	81.8 min	~45 GB	8.4M (0.5%)
Dreaddit	2, 838	5	5 min	~35 GB	8.4M (0.5%)

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

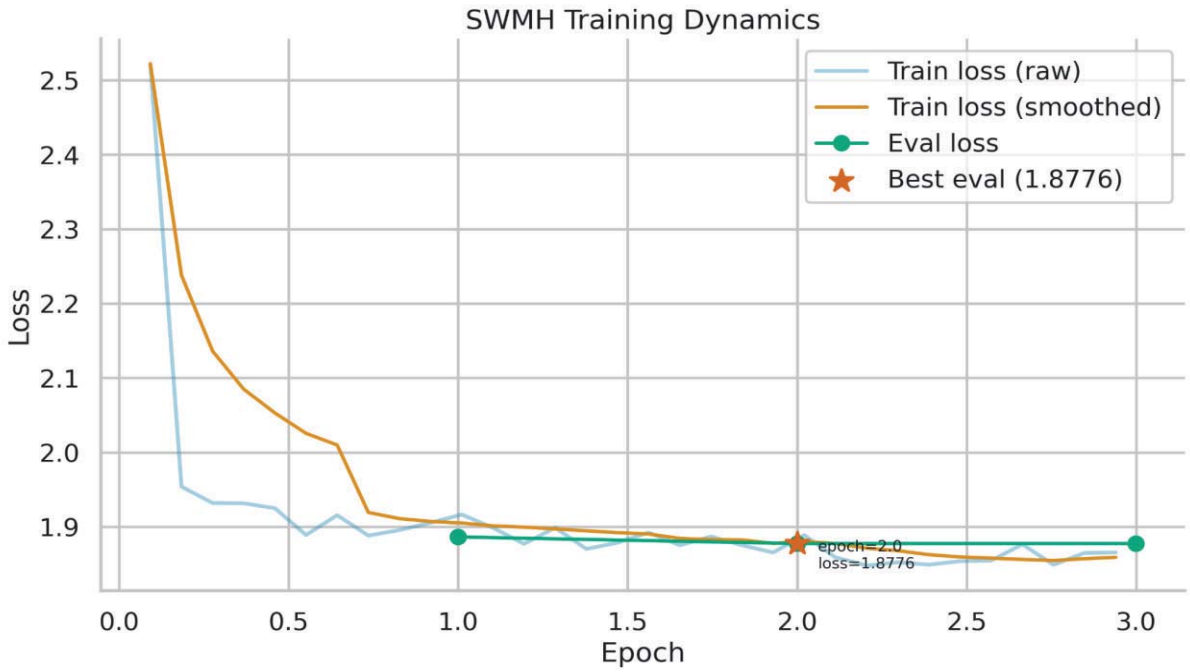

Figure 4. SWMH training dynamics (train/eval loss; best epoch=2).

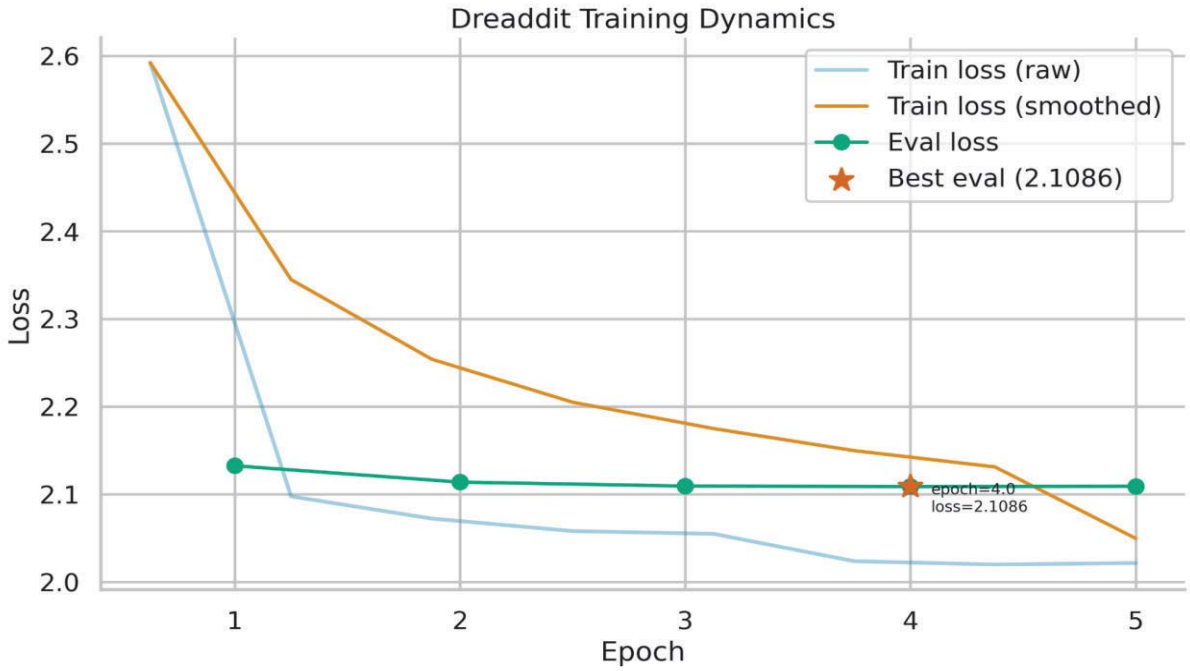

Figure 5. Dreaddit training dynamics (train/eval loss; best epoch=4).

A fair comparison of models was conducted on identical random subsets (N = 200 per dataset), with all models evaluated on the same test indices. On SWMH, SmolLM2-1.7B achieved paired accuracy = 0.6600, compared with 0.6250 for GPT-4o-mini and 0.6300 for GPT-4o. McNemar's exact test showed that these SWMH differences were not statistically significant ($p = 0.401062$ against GPT-4o-mini; $p = 0.429591$ against GPT-4o). On Dreaddit, SmolLM2-1.7B achieved a paired accuracy of 0.8000, compared with 0.6500 for GPT-4o-mini and 0.6750 for GPT-4o. These Dreaddit differences were statistically significant according to McNemar's exact test ($p = 0.000073$ against GPT-4o-mini; $p = 0.000621$ against GPT-4o).

DOI: 10.37943/XSAJ1029

Thus, the paired protocol supports a statistically reliable advantage for the fine-tuned small model on Dreaddit and only a numerical, non-significant advantage on SWMH. Fig. 6 presents this paired comparison: for each dataset the grouped bars give the accuracy of SmolLM2-1.7B, GPT-4o-mini and GPT-4o on exactly the same 200 test indices, so that the size of the gap between the fine-tuned small model and the two zero-shot baselines can be read directly for the multiclass and for the binary task.

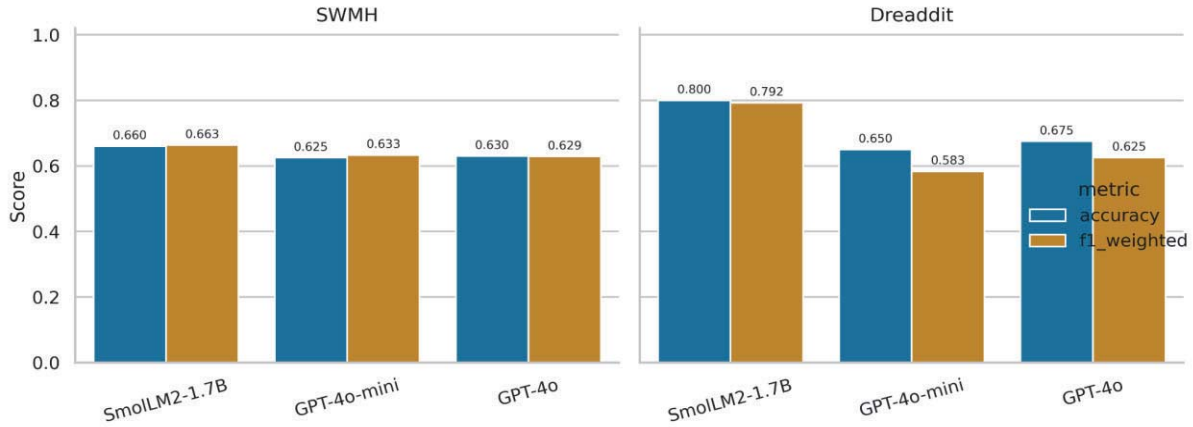

Figure 6. Fair paired comparison on identical test indices (N=200): SmolLM2-1.7B vs GPT-4o-mini vs GPT-4o.

To assess uncertainty in the reported full-test metrics, 95% bootstrap confidence intervals were computed over held-out test examples. For SWMH, the intervals were 0.6973-0.7144 for accuracy, 0.6989-0.7159 for weighted F1, and 0.7129-0.7289 for macro F1. For Dreaddit, the corresponding intervals were 0.7762-0.8336 for accuracy, 0.7747-0.8326 for weighted F1, and 0.7742-0.8320 for macro F1. These intervals provide an estimate of sampling uncertainty around the obtained performance metrics.

Calibration analysis indicated that the probabilistic outputs were informative but not perfectly calibrated. On SWMH, the Expected Calibration Error was 0.1826 and the multiclass Brier score was 0.4667; on Dreaddit, the Expected Calibration Error was 0.1757 and the binary Brier score was 0.1701. To show how this miscalibration is distributed over the confidence range, Fig. 7 reports reliability diagrams for both datasets, in which the mean predicted confidence of each bin is plotted against the empirical accuracy of the same bin, so that bins lying below the diagonal correspond to overconfident predictions. Several confidence bins were indeed overconfident relative to empirical accuracy. Accordingly, such confidence estimates should be interpreted in mental-health text analysis as decision-support signals and not as clinically useful risk probabilities.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

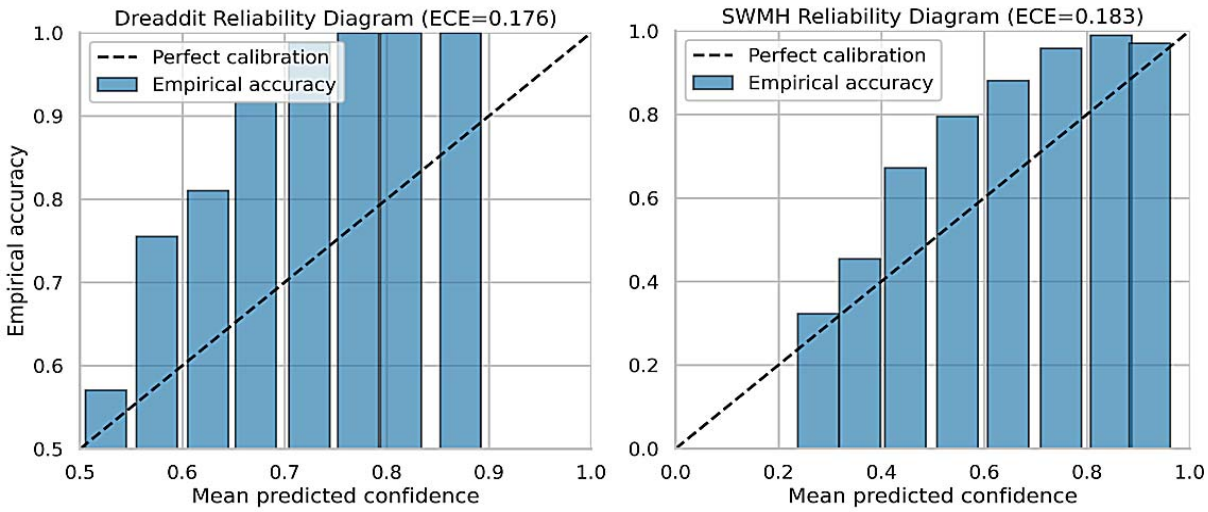

Figure 7. Reliability diagrams for Dreaddit and SWMH

Paired comparison analysis clarified the interpretation of cross-model differences. With respect to SWMH, SmolLM2-1.7B had paired accuracy of 0.6600 to 0.6250 for GPT-4o-mini and 0.6300 for GPT-4o, but not statistically significant with McNemar's exact test. Using McNemar's exact test on Dreaddit, SmolLM2-1.7B had paired accuracy of 0.8000 versus 0.6500 for GPT-4o-mini and 0.6750 for GPT-4o. In other words, the evidence points to a statistically reliable advantage in the fine-tuned small model and a numerical but non-significant advantage on the SWMH.

Discussion

Results show that parameter-efficient fine-tuning of small language models is a good strategy for mental-health text classification, but the comparative claim has to be precise. Full-held-out test evaluation with length-normalized label log-likelihood scoring - SmolLM2-1.7B: 0.7056, 0.7072 weighted F1 and 0.7207 macro F1 on SWMH and 0.8056, 0.8048, weighted F1 and 0.8043 macro F1 on Dreaddit. Compared with zero-shot GPT-4o-mini and GPT-4o on both datasets, the fine-tuned model was numerically stronger than the zero-shot GPT-4o-mini and GPT-4o but statistically significant only for Dreaddit. All of this points to a more robust conclusion regarding binary stress detection, as well as a more cautious claim of competitiveness for multiclass SWMH classification. More calibrated analysis also shows that while the model yields usable confidence estimates, these are imperfectly calibrated and thus should be interpreted as decision-support signals rather than clinically reliable risk probabilities.

These findings challenge the belief that larger language models are better for classification, at least for domain-specialist classification. In this respect, adaptation quality is likely to be of greater importance than parameter count on domain-specific tasks such as mental health signal detection, as here interpreted. That explanation is supported by earlier observations that domain-specific models are superior to domain-agnostic zero-shot systems in the corresponding targeted tasks [24], as here verified. It is also compatible with clinical NLP accounts where fine-tuned BERT-family systems were observed to outperform larger untuned alternatives on extraction and classification tasks [25], in such circumstances as represented here. In that latter context, the findings here are not dissimilar to those found for Reddit-based mental health classification, and equivalent to or better than current GPT baselines. Consistent with earlier SWMH-style benchmarks, full test performance is within or beyond the generally reported limits of BERT-era systems [26], and Dreaddit performance equivalent to or beyond classical BERT-base baselines [15].

DOI: 10.37943/XSAJ1029

Class-wise results show which classes the model continues to struggle with. In SWMH, Anxiety and Bipolar have higher recall than precision and hence are higher than the default, which suggests they are over-assigned; Depression has lower recall than precision which suggests that it is under-assigned. OffMyChest is polarized in the opposite direction, which suggests it is over-assigned to predictions. SuicideWatch is balanced as this pair is, yet is still safety-relevant. On Dreaddit, recall favours the Stressed class in a balanced fashion overall. These trends are consistent with a classifier whose distinctions between clearer affective signals it handles more reliably than classes that have broad or overlapping semantics. The confusion between Depression and SuicideWatch in particular is plausible given the lexical overlap of depression and suicidality discourse and the inherent ambiguity of user-generated labels in social media corpora [27].

In a deployment context, PEFT achieves a highly efficient profile: only ~0.5% of parameters were trainable, convergence was fast, and training completed in one A100 GPU with a moderate memory budget. From a practitioner's lens, compressed tuned models naturally come with decreased dependency on external API calls and enable better data-governance control over sensitive text streams [28-29]. Locally (or on a contained deployment) one could improve the privacy posture with respect to off-site pipelines [30-31].

Some limitations still apply. Notably, external baseline comparison is only to zero-shot GPT-4o mini and GPT-4o, not stronger few-shot/fine-tuned large-model baselines [32]. First, there are only 200 matched examples per dataset for paired cross-model comparison, reducing statistical power in a multiclass setting. Calibration was only explicitly assessed with Expected Calibration Error/Brier score/reliability diagrams; the resulting probabilities were not post hoc temperature-scaled. Fourth, uncertainty is bootstrapped over held-out test examples, not variance across multiple fine-tuning seeds. Last, the dataset labels are benchmark annotations from online communities and not clinical diagnoses; real-world clinical generalisation requires external validation and expert oversight [33-34].

In an ethical sense these models should be considered decision support tools and not diagnostic systems. False negatives in safety-critical categories and false positives that may cause stigma or unnecessary escalation both have explicit safeguards. Practical uses will need conservative thresholds, human oversight and clear escalation protocols. Work on multi-label/hierarchical formulations, calibrated confidence estimation, temporal risk modeling and external-domain validation is needed [35].

Conclusion

It is shown that parameter-efficient fine tuning of small language models can achieve good and practically relevant performance for mental-health text classification with only a fraction of model parameters updated. Full-held out test evaluation with length-normalized label log-likelihood scoring - SmolLM2-1.7B: 0.7056, 0.7072 weighted F1 and 0.7207 macro F1 on SWMH and 0.8056, 0.8048 weighted F1 and 0.8043 macro F1 on Dreaddit. Compared with identical subsets, the fine-tuned model was numerically better than the zero-shot GPT-4o-mini and GPT-4o on both tasks, but it was not statistically significant compared to Dreaddit. Such results support the conclusion that task-specific PEFT makes a compact model competitive with larger zero-shot systems in multiclass mental health classification, and demonstrably stronger in binary stress detection under matched evaluation. But at the same time such systems should be considered decision support tools and not diagnostic instruments. Future work should include larger model baselines, post hoc calibration, and external domain validation before high-stakes deployment is considered.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

References

[1] World Health Organization. (2022). World mental health report: Transforming mental health for all. Geneva: WHO. ISBN 978-92-4-004933-8. https://www.who.int/publications/i/item/9789240049338

[2] Gulliver, A., Griffiths, K. M., & Christensen, H. (2010). Perceived barriers and facilitators to mental health help-seeking in young people: A systematic review. BMC Psychiatry, 10(1), 113. https://doi.org/10.1186/1471-244X-10-113

[3] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. NPJ Digital Medicine, 3(1), 43. https://doi.org/10.1038/s41746-020-0233-7

[4] Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys, 56(2), Article 30, 1-40. https://doi.org/10.1145/3605943

[5] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623. https://doi.org/10.1145/3442188.3445922

[6] Schick, T., & Schütze, H. (2021). It's not just size that matters: Small language models are also few-shot learners. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, 2339-2352. https://doi.org/10.18653/v1/2021.naacl-main.185

[7] Ben Allal, L., Lozhkov, A., Bakouch, E., Penedo, G., von Werra, L., & Wolf, T. (2024). SmolLM2-1.7B-Instruct [Model card]. Hugging Face. https://huggingface.co/HuggingFaceTB/SmolLM2-1.7B-Instruct

[8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR 2022). https://openreview.net/forum?id=nZeVKeeFYf9

[9] Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., & Ananiadou, S. (2023). Towards interpretable mental health analysis with large language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 6056-6077. https://doi.org/10.18653/v1/2023.emnlp-main.370

[10] Losada, D. E., Crestani, F., & Parapar, J. (2020). Overview of eRisk 2020: Early risk prediction on the internet. Experimental IR Meets Multilinguality, Multimodality, and Interaction, 12260, 272-287. https://doi.org/10.1007/978-3-030-58219-7_20

[11] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. Proceedings of the International AAAI Conference on Web and Social Media, 7(1), 128-137. https://doi.org/10.1609/icwsm.v7i1.14432

[12] Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., & Leung, H. (2020). SenseMood: Depression detection on social media. Proceedings of the 2020 International Conference on Multimedia Retrieval, 407-411. https://doi.org/10.1145/3372278.3391932

[13] Ji, S., Yu, C. P., Fung, S. F., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. Complexity, 2018, Article 6157249. https://doi.org/10.1155/2018/6157249

[14] Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., … & Sheth, A. (2019). Knowledge-aware assessment of severity of suicide risk for early intervention. The World Wide Web Conference, 514-525. https://doi.org/10.1145/3308558.3313698

[15] Turcan, E., & McKeown, K. (2019). Dreaddit: A Reddit dataset for stress analysis in social media. Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), 97-107. https://doi.org/10.18653/v1/D19-6213

[16] Lawrence, H. R., Schneider, R. A., Rubin, S. B., Mataric, M. J., McDuff, D. J., & Jones Bell, M. (2024). The opportunities and risks of large language models in mental health. JMIR Mental Health, 11, e59479. https://doi.org/10.2196/59479

[17] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., … & Le, Q. V. (2022). Finetuned language models are zero-shot learners. International Conference on Learning Representations. https://openreview.net/forum?id=gEZrGCozdqR

[18] Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In International Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=RIu5lyNXjT

[19] Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. IEEE Intelligent Systems, 38(2), 15-23. https://doi.org/10.1109/MIS.2023.3254179

[20] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., … & Liu, Z. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 5(3), 220-235. https://doi.org/10.1038/s42256-023-00626-4

[21] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36, 10088-10115. https://doi.org/10.5555/3666122.3666563

[22] Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2024). Mental-LLM: Leveraging large language models for mental health prediction via online text data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(1), 1-32. https://doi.org/10.1145/3643540

[23] Ji, S., Li, X., Huang, Z., & Cambria, E. (2022). Suicidal ideation and mental disorder detection with attentive relation networks. Neural Computing and Applications, 34(13), 10309-10319. https://doi.org/10.1007/s00521-021-06208-y

[24] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., … & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1), 1-23. https://doi.org/10.1145/3458754

[25] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730

[26] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: Publicly available pretrained language models for mental healthcare. Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022), 7184-7190. https://aclanthology.org/2022.lrec-1.778/

[27] Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., … & Schwartz, H. A. (2019). Suicide risk assessment with multi-level dual-context language and BERT. Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 39-44. https://doi.org/10.18653/v1/W19-3005

[28] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., … & Leskovec, J. (2023). Holistic evaluation of language models. Transactions on Machine Learning Research. https://openreview.net/forum?id=iO4LZibEqW

[29] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645-3650. https://doi.org/10.18653/v1/P19-1355

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

[30] Patterson, D., Gonzalez, J., Holzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. Computer, 55(7), 18-28. https://doi.org/10.1109/MC.2022.3148714

[31] Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. Proceedings of the Conference on Fairness, Accountability, and Transparency, 79-88. https://doi.org/10.1145/3287560.3287587

[32] Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). Large language models are human-level prompt engineers. International Conference on Learning Representations. https://openreview.net/forum?id=92gvk82DE-

[33] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, 2980-2988. https://doi.org/10.1109/ICCV.2017.324

[34] Manikonda, L., & De Choudhury, M. (2017). Modeling and understanding visual attributes of mental health disclosures in social media. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 170-181. https://doi.org/10.1145/3025453.3025932

[35] Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8), 1819-1837. https://doi.org/10.1109/TKDE.2013.39