TinyMentalLLMs Enable Depression Detection in Chinese Social Media Texts

Jinyuan Xu[1,†], **Tian Lan**[2,†], **Mathieu Valette**[1], **Pierre Magistry**[1], **Lei Li**[3,4,*]
[1]ERTIM, Institut National des Langues et Civilisations Orientales, Paris, France
[2]Milkuya Studio, Hangzhou, China
[3]VitaSight, Shanghai, China
[4]University of Washington, Seattle, USA
{jinyuan.xu, mvalette, pierre.magistry}@inalco.fr
tianlan@milkuya.com, lenny.lilei.cs@gmail.com

Abstract

Depression remains a major global mental health concern, bringing a higher risk of suicide and growing social costs tied to mental disorders. Leveraging social media as a valuable source of emotional signals, we identify two limitations in current NLP-based depression detection frameworks: (1) prediction systems often lack clear, user-friendly explanations for predictions in Depression Detection, and (2) the computational and confidentiality demands of LLMs are misaligned with the need for dependable, privacy-focused small-scale deployments. To address these challenges, we introduce TinyMentalLLMs (TMLs), a compact framework that offers two **key contributions**: (a) the construction of a small yet representative dataset through psychology-based textometry, and (b) an efficient fine-tuning strategy centered on multiple aspects of depression. This design improves both accuracy and F1 scores in generative models with 0.5B and 1.5B parameters, consistently yielding over 20% performance gains across datasets. TMLs achieve results on par with, and deliver better text quality than, much larger state-of-the-art models.

1 Introduction

Depression is one of the most common mental disorders and a major contributor to suicide worldwide. According to the WHO's March 2023 report[1], about 280 million people worldwide are affected. China alone accounts for approximately 2.8 million cases (Wang et al., 2024a), but many receive no effective treatment.

The rise of social media, such as Weibo, has formed an ecosystem where users routinely share their moods, with many posting emotion-related content each week. These dense, time-stamped traces enable large-scale, cost-effective early warning systems for depression.

Early depression-detection systems relied on handcrafted lexical and psycholinguistic features combined with traditional machine learning methods (Islam et al., 2018; Guntuku et al., 2017). Although these models are interpretable, they lack deep semantic understanding. Transformer-based classifiers such as BERT offer stronger contextual representations (Malviya et al., 2021), but they still struggle with long texts (Gao et al., 2021) and are not easily interpretable by humans.

Consequently, researchers are increasingly adopting generative large language models (LLMs) for depression detection (Lan et al., 2024; Hu et al., 2024; Xin Yan, 2023; Lai et al., 2023), which generate both a label and a summary-style explanation (Wang et al., 2024c; Yang et al., 2024b; Xu et al., 2024; Yang et al., 2023), thereby contextualizing the output for improved transparency and interpretability. While large-parameter models typically achieve superior performance, they require substantial computational resources and raise concerns regarding privacy, security, and cost (Yao et al., 2024). In contrast, smaller-parameter models demand fewer resources but often suffer from quality issues such as hallucination. In the specific context of depression detection on Chinese social media, although a number of datasets have been released[2] (Harrigian et al., 2021), most are designed for binary classification. To date, there is still no dataset suitable for post-label summary generation and interpretability analysis.

To address these challenges, in this paper, we introduce *TinyMentalLLMs* (TMLs), a family of compact generative models (0.5B and 1.5B parameters) tailored to Chinese social-media texts. We employ a multi-turn fine-tuning strategy and a new heuristic

[1]https://www.who.int/news-room/fact-sheets/detail/depression

[2]https://github.com/bucuram/depression-datasets-nlp

Proceedings of Recent Advances in Natural Language Processing,pages 1352–1363
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-157

survey dataset to achieve performance comparable to larger models. This approach aims to enhance the quality of the generated text and provide more structured analytical explanations, making it more suitable for scenarios with limited computational resources. Our research aims to detect potential depressive risks and extreme depressive moods as an auxiliary tool for mental health support, rather than directly diagnosing clinical depression or other psychological disorders.

To achieve these goals, we make the following contributions:

- **Construction of a small depression survey dataset:** Currently, there is no suitable survey-based dataset exists for Chinese social media users' depressive states. To address this, we employed Lafon-based textometry (Lafon, 1980), a proven effective method for mining semantic dimensions in textual data, facilitating both quantitative and qualitative analysis (Pincemin, 2022; Eensoo and Valette, 2012), and a heuristic semantic framework (Anonymous, 2024) for multidimensional feature selection. Experts then conducted semi-manual labeling on a specially representative dataset, producing a small high-quality depression survey dataset.

- **Proposed Strategies for Domain-Specific Fine-tuning and Model Implementation in Chinese Depression Detection task: Tiny Mental LLMs:** To enhance the applicability of small language model-based detection systems and improve interpretability, we implemented an efficient multi-turn fine-tuning process. Accompanied by a textometry-semantic strategy, this process uses a newly created dataset and existing datasets. The approach aims to mitigate formatting errors that arise during the generation of analytical explanations by small models.

- **Evaluation of comparative results between different methods:** We conducted experimental evaluations on several mainstream Chinese general-purpose large language models with more parameters and some supervised methods. We compared their performance with our two TML models, demonstrating that, despite their smaller size, our models achieve comparable or superior results and offer better text quality from a semantic information perspective.

2 Related Works

Early research on depression detection and mental health analysis relied on classical machine learning (Cortes, 1995; Breiman, 2001; McCallum et al., 1998; Dreiseitl and Ohno-Machado, 2002) and neural networks (Schmidhuber, 2015), often employing diverse text feature extraction techniques. Linguistic features (Arora and Arora, 2019; Eensoo and Valette, 2012; Yang et al., 2020), lexeme- or n-gram-based statistical extraction (Ramos et al., 2003), and textometry approaches focusing on semantic chunks (Brown et al., 1992) were used across different corpora, alongside psychological dictionaries (Tausczik and Pennebaker, 2010).

Word embedding techniques (Mikolov et al., 2013; Pennington et al., 2014; Joulin et al., 2016) were introduced to better capture semantic information. When integrated with deep learning frameworks (Zhang and Wallace, 2015; Zeiler and Fergus, 2014; Graves, 2013; Wang et al., 2016), they improved semantic feature extraction. Recently, Transformer architectures, especially BERT, have dominated classification tasks, leveraging attention mechanisms and advanced pre-training/fine-tuning to advance the state of the art in mental health detection (Zhai et al., 2024; Ji et al., 2022).

With the rise of ChatGPT[3], generative large language models such as LLaMA (Touvron et al., 2023), GLM (Du et al., 2022), Qwen (Bai et al., 2023), DeepSeek (DeepSeek-AI et al., 2025), and Claude[4] are used across disciplines, including mental health detection. Researchers employ these models with prompt strategies like Chain-of-Thought (Wei et al., 2023; Li, 2024) to evaluate different applications (Shi et al., 2024; Li et al., 2024; Cai et al., 2024), e.g. mental health, and some fine-tune them via multi-turn dialogues for real-world scenarios (Hu et al., 2024; Xu et al., 2024).

3 Dataset

To align with our multi-stage training framework, we constructed three types of training sets and three corresponding test sets (see Table 1). The raw data originated from the SWDD) (Cai et al., 2023)

[3] https://chat.openai.com/
[4] https://claude.ai/

and WU3D (Wang et al., 2020) datasets, both collected from Sina Weibo (China's largest microblogging platform). Each dataset is organized by user, with each user contributing dozens to hundreds of posts labeled as either depression (positive) or non-depression (negative). To focus on more active users, we restricted our dataset to those with over 60 posts and total text lengths of 3,000–5,000 characters.

We constructed three sub-datasets for training. The SWDD dataset includes self-reported depression markers, forming a positive sub-dataset from confirmed depressed users and a matched amount of negative data (meeting length/post-count criteria). Since the labels are limited to depression or non-depression, we name this sub-dataset **SWDD self-reported-label** (SR Label).

The second sub-dataset is a small yet representative survey of depressed users. We applied textometry-based semantic modeling to positive users from the SR Label dataset to identify the most representative cases, which then underwent a two-step AI-human collaborative annotation flow:

- **AI-assisted pre-screening**: GPT-o1 API generated preliminary labels via structured prompts incorporating DSM-5 [5] criteria.
- **Expert verification**: Two experts with advanced degrees (Master's or higher) in psychology conducted a dual-blind review, achieving a 92.3% inter-annotator agreement rate and a Fleiss' kappa (McHugh, 2012) of 0.89. Discrepancies were resolved through iterative consensus meetings, ensuring rigorous alignment in structured analyses.

This step aims to select some "golden", representative data in the production of such depression detection datasets. We name this dataset the **SWDD self-reported-survey** (SR Survey) dataset, with detailed construction steps outlined in section 4.1.

The third sub-dataset involves splitting the second sub-dataset based on semantic dimensions derived from textometry. This process enhances the model's analytical and interpretative capabilities during training by aligning it with different semantic dimensions identified in the survey. We name this subset the **SWDD self-reported multi-dims** (SR Multi-Dims) dataset.

[5] https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/dsm-5-fact-sheets

In our testing phase, we utilized three test datasets. The first originates from the **WU3D** dataset, the second comprises users from the SWDD dataset who have self-reported (**SWDD SR Test**), and the third includes SWDD users without self-reported depression but labeled as positive by expert reviewers (**SWDD Test**). For these test datasets, approximately 200 users were randomly selected for the positive data within the original dataset range, and the negative data were also randomly chosen.

We specifically isolated the SWDD self-reported users from the SWDD dataset because labeling depression in internet data is exceptionally challenging. Even if users mention numerous symptoms of depression or expressions of depressive mood in their posts, these cannot definitively classify them as depressed. Data from self-reported markers are considered to have fewer annotation errors and to be more representative, making them more suitable for use in low-resource and constrained settings during training or fine-tuning. Furthermore, although some sub-datasets are derived from the same original dataset, there is no overlap between the test datasets, and the datasets used for training are also independent of each other.

Dataset Type	Dataset Name	Positive Users	Negative Users
train	SR Label	733	733
	SR Multi-Dims	540	270
	SR Survey	90	45
test	SWDD SR Test	206	197
	SWDD Test	200	200
	WU3D	200	200

Table 1: Datasets Overview.

4 Methodology

In this section, we discuss the construction of efficient, small models for depression detection under resource constraint. (See Figure 1 and Figure 2)

4.1 Survey Dataset construction

This new survey dataset is derived from the self-reported section of the SWDD dataset. We employed the Lafon Specificity-based method (Lafon, 1980), an established approach in textometry, for extracting semantic dimensions from textual data. This method is useful for both quantitative and

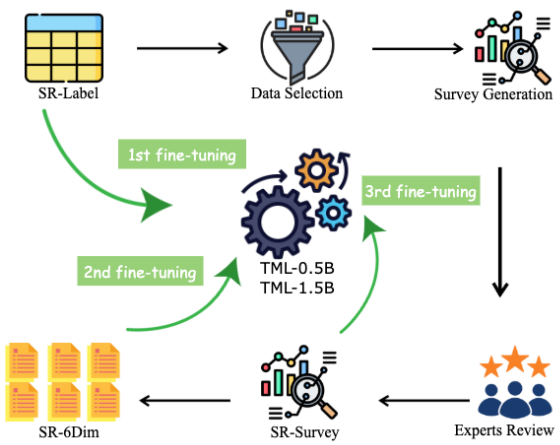

Figure 1: Dataset Construction and Fine-Tuning Process.

qualitative analyses and is commonly used in corpus linguistics and digital humanities. It facilitated the filtering and multi-dimensional modeling of data from representative target user groups. Subsequently, we used the GPT-o1 preview model[6] for dimension-specific user survey generation. Finally, two psychology experts cross-verified and refined each generated survey to validate accuracy and relevance.

The Lafon Specificity (Lafon, 1980) allows us to assess a corpus containing two sub-corpora, positive and negative, by analyzing each repeated semantic segment (whether unigrams or n-grams with important meaning) and their relevance to one of the sub-corpora under study. It assigns a score to each repeated segment (Salem, 1986), identifying those with the highest semantic relevance to the target sub-corpus. This method also helps reduce the impact of high-frequency words common to both sub-corpora.

Using this method, we extracted thousands of semantically relevant and highly scored repeated segments from positive user texts. we divided these segments into six semantic categories based on existing psychological questionnaires, criteria (Kroenke et al., 2001; American Psychiatric Association et al., 2000; Beck et al., 1996) and the characteristics of internet texts (Mothe et al., 2022) and Chinese semantics. These six dimensions include two levels: a primary level including **Negative Emotions**, **Depressive Psychological States**, and **Clinical Symptoms**; and a secondary level

[6] https://openai.com/index/introducing-openai-o1-preview/

Input data
{user id:posts[text_1,text_2,...,text_n]}
SR Label
instruction: 请仔细阅读以下推文，判断这名用户是否表现出抑郁情绪。如果存在抑郁情绪，回答“是”；如果没有，回答“否”。 *Please carefully read the following tweets and determine whether this user exhibits depressive emotions. If depressive emotions are present, answer "Yes"; otherwise, answer "No".*
output: 是/否 *Yes/No*
SR Survey
instruction: 请仔细阅读下列推文，并作出判断：用户是否有抑郁情绪？如果有，回答“是”；如果没有，回答“否”。然后分步骤说明你的分析。 *Please carefully read the following tweets and determine: Does the user exhibit depressive emotions? If yes, answer "Yes"; if no, answer "No." Then, explain your analysis step-by-step.*
output: 答案：“否”; **总结和分析：** **抑郁心理状态（主要判断标准）：原因;** **抑郁相关的医疗表达（主要判断标准）：原因;** **抑郁相关的临床症状（主要判断标准）：原因;** **负面情绪（次要判断标准）：原因;** **造成抑郁的潜在外因（次要判断标准）：原因;** **抑郁相关的语言表达模式（次要判断标准）：原因;** *Answer: "No": Summary and Analysis:* *Depressive Psychological State (Primary Criterion): reasons;* *Depression-Related Medical Expressions (Primary Criterion):reasons;* *Depression-Related Clinical Symptoms (Primary Criterion):reasons;* *Negative Emotions (Secondary Criterion):reasons;* *Potential External Causes of Depression (Secondary Criterion):reasons;* *Depression-Related Language Patterns (Secondary Criterion):reasons;*
SR Multi-Dim
instruction: 请仔细阅读以下文本，分析用户的DIM1/DIM2/.../DIM6，并提供你的推理理由。 *Please carefully read the following texts, analyze the user's display of* **DIM1/DIM2/.../DIM6**, *and provide your reasoning.*
output: **DIM1:[text1,text2,...textn], ... ,DIM6:[text1,text2,...textm]**

Figure 2: Data Samples for SR-Label, SR Survey, and SR-Multi-Dims.

including **Potential External Factors**, **Medically Related Aspects**, and **Special Language Expressions**. Experts then manually assigned each repeated segment to its respective category. .

We assigned and summed values for all repeated segments appearing in user posts, selecting the top 40 users in each dimension based on their scores, totaling 240 users (ensuring that scores in other dimensions did not exceed the score of the dimension under study). Subsequently, we randomly selected 20 users from the three main semantic categories and 10 users from the three secondary categories, totaling 90 positive user tweet data sets. Finally, we randomly selected another 45 users from the negative sub-corpus. This selection of highly representative users was used to complete the production of the survey dataset.

4.2 Training Strategies

We chose QWEN2.5-0.5B-Instruct and QWEN2.5-1.5B-Instruct (QwenTeam, 2024; Bai et al., 2023) as our base models because they are currently the only ultra-small models among the most popu-

lar Chinese LLMs. Due to the small model size and limited dataset, overfitting was a primary concern. To address this, we implemented a structured, step-by-step approach that progressively introduced more complex tasks, while applying techniques such as text shuffling, prompt variation, and sentence replacement. These strategies were crucial in helping the model generalize well to new data while maintaining its ability to handle both binary classification and detailed analysis of depressive text across multiple dimensions.

In the first stage of our fine-tuning process, we trained the model for one epoch on the **SWDD self-reported-label** dataset. The prompt instructed the model to decide whether the user in the provided text showed signs of depression. The prompt template used was: *"Please determine whether the user has depressive emotions based on the following text. If there are depressive emotions, answer 'yes'; if not, answer 'no'."*

This step focused purely on binary judgments, whether the user is depressed or not, with responses limited to "yes" or "no." By starting with this simple task, we ensured that the model could reliably classify depressive states before moving on to more complex tasks. This also helped the model become familiar with the typical language used in social media texts.

In the second stage, we moved to a more detailed analysis by dividing the task into six specific semantic dimensions of the depressive text. We used the **SWDD self-reported multi-dims** as the training set for this stage, the model was fine-tuned for two epochs on data corresponding to each dimension using relevant prompts. The goal was to ensure that the model could analyze user texts using these dimensions, preparing it for more nuanced and comprehensive analysis in the later stages.

In the final stage, the model's task was to generate both a binary decision and an explanation of the reasoning process, regarding whether the user's text showed depressive emotions and to explain its reasoning. To do this, we used the following prompt:

"Please determine whether the user has depressive emotions based on the following text. If there are depressive emotions, answer 'yes'; if not, answer 'no'. After providing the answer, please explain your reasoning step by step."

We fine-tuned the model for one epoch on the **SWDD self-reported Survey** dataset to improve its accuracy and provide clearer explanations for its decisions.

Given the small size of both our model and training dataset, it is evident that overfitting was a major concern. To mitigate overfitting during the fine-tuning process, we implemented the following strategies:

- **Shuffling User Text**: For each training epoch, we shuffled the order of user texts in the training set to introduce variability.
- **Prompt Randomization**: For the prompts used during training, we randomly selected a prompt from a list of alternatives. Each prompt had a similar meaning but varied in wording, ensuring that the model was not conditioned on the specific wording of a single prompt.
- **Sentence and Word Replacement**: For responses in the **SWDD self-reported multi-dims** and the **SWDD self-reported Survey** training set, we randomly replaced frequently occurring sentences and words with similar alternatives, using sentences or words with equivalent meanings. This further reduced the risk of the model overfitting to specific phrases or terms.

These strategies aimed to introduce diversity into the fine-tuning process and ensure that the model could generalize better to new data.

5 Experimental Settings

5.1 Baselines

We employed various supervised models, including BERT-based methods like **BERT-base-chinese** (Devlin et al., 2019), **Chinese-RoBERTa-wwm-ext** (Cui et al., 2019), and **StructBERT-mental** (Wang et al., 2019).

We also considered generative models, including **Llama3-8B-Chinese-Chat** (Wang et al., 2024b), **GLM-4-9B-Chat** (GLMTeam et al., 2024), and **Qwen2.5 series** (QwenTeam, 2024; Yang et al., 2024a).

Finally, to ensure a comprehensive evaluation, we also incorporated **GPT-4o**, **GPT-4o Mini**[7] and **DeepSeek-R1-671B**[8] (DeepSeek-AI et al., 2025) models into our baseline comparisons.

[7] https://chat.openai.com/
[8] https://www.deepseek.com/

Model	SWDD self-reported				SWDD				WU3D			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Baseline Supervised Models												
RoBERTa Chinese	0.76	0.68	0.99	0.80	0.85	0.78	0.97	0.86	0.74	0.66	0.96	0.78
BERT based Chinese	0.77	0.69	0.98	0.81	0.83	0.76	0.96	0.85	0.75	0.68	0.95	0.79
StructBERT-mental	0.77	0.70	0.97	0.81	0.86	0.79	0.97	0.87	0.77	0.69	0.96	0.80
Baseline Generative Models												
Llama3-8B-Chinese-Chat	0.83	0.84	0.82	0.83	0.86	0.83	0.90	0.86	0.79	0.81	0.76	0.79
GLM-4-9B-Chat	0.75	0.68	0.97	0.80	0.76	0.68	1.00	0.81	0.77	0.69	0.98	0.81
Qwen2.5-7B-Instruct	0.86	0.81	0.95	0.88	0.88	0.82	0.98	0.89	0.84	0.78	0.94	0.85
GPT-4o Mini	0.66	0.60	1.00	0.75	0.65	0.58	1.00	0.74	0.65	0.59	1.00	0.74
GPT-4o	**0.93**	0.88	**1.00**	**0.93**	**0.92**	0.86	**1.00**	**0.92**	**0.91**	0.85	**1.00**	**0.92**
DeepSeek-R1-671B	0.88	0.82	0.99	0.89	0.86	0.78	0.99	0.87	0.89	0.82	0.99	0.90
Qwen2.5-0.5B-Instruct	0.52	0.52	0.67	0.59	0.51	0.50	0.68	0.58	0.52	0.52	0.62	0.56
Qwen2.5-1.5B-Instruct	0.56	0.55	0.83	0.66	0.63	0.59	0.87	0.70	0.58	0.56	0.78	0.65
Fine-tuned Models												
TML-0.5B	0.81	0.83	0.80	0.82	0.84	0.89	0.77	0.82	0.78	0.82	0.71	0.76
TML-1.5B	0.87	**0.94**	0.80	0.87	0.84	**0.90**	0.76	0.82	0.81	**0.95**	0.66	0.78

Table 2: Performance of Baseline and Fine-tuned Models on SWDD self-reported, SWDD, and WU3D datasets. The best scores are in bold.

6 Results and Discussions

6.1 Experiment Setup

During the fine-tuning phase, we used an Nvidia A800 GPU with 80GB of memory. Fine-tuning was performed using LlamaFactory[9] (Zheng et al., 2024) with a learning rate of 2e-05, a batch size of 4, and a linear learning rate scheduler, ensuring full-parameter fine-tuning throughout the process.

For testing, the smaller models (TML-0.5B and TML-1.5B) were evaluated on a consumer-grade Nvidia RTX 4090 GPU, demonstrating their flexibility and accessibility. In contrast, the baseline generative models were tested on an Nvidia V100 32G GPU. We conducted two experiments. Even compared with ultra-large models like DeepSeek-R1-671B, GPT-4o, and GPT-4o Mini, our results remain competitive. All tests used the same generation parameters: temperature of 0.7, top p of 0.8, and top k of 20, based on the official configuration of the Qwen model. This setup allowed us to maintain consistency and compare the models' performance under the same conditions

6.2 Depression Classification

The results shown in Table 2 provide a clear comparison between the baseline models and the fine-tuned models across multiple datasets, highlighting the performance improvements achieved through fine-tuning.

Our fine-tuned models outperform the baseline models in many aspects:

- The TML-0.5B model delivers performance comparable to traditional supervised models like RoBERTa-Chinese. On the SWDD self-reported dataset, it achieves an accuracy of 0.81 and an F1 score of 0.82, closely matching the results of traditional models.

  A key advantage of TML-0.5B is its resource efficiency. Despite being a generative model, it remains competitive in computational and memory requirements, making it suitable for resource-limited scenarios. Moreover, unlike traditional supervised approaches, it provides explanations for its predictions, enhancing interpretability and transparency critical in tasks like depression detection.

- TML-1.5B demonstrates notable advantages over larger generative models (Llama3-8B-Chinese, GLM4-9B-Chat, and Qwen2.5-7B-Instruct). In terms of performance, it achieves an F1 score of 0.87 on the SWDD dataset, outperforming Llama3-8B-Chinese (0.83) and GLM4-9B-Chat (0.80). This shows that, even with fewer parameters, the fine-tuned TML-1.5B performs better on specific tasks, particularly in classification tasks like depression detection.

  Additionally, TML-1.5B offers an efficiency advantage over models like Qwen2.5-7B-Instruct. Despite having fewer parameters, it achieves higher accuracy and F1 scores across most datasets while consuming fewer computational resources.

- Compared to other baseline methods, our fine-tuned models show higher precision but lower recall, reflecting more precise yet cautious judgments. For example, the TML-1.5B

[9] https://github.com/hiyouga/LLaMA-Factory

model achieves a precision of 0.94 on the SWDD self-reported dataset, surpassing models like Llama3-8B-Chinese and GLM4-9B-Chat, but the recall is lower at 0.80. This suggests that while our models are highly accurate in identifying positive cases, they tend to be conservative, possibly missing some true positives to maintain higher precision. In practice, we prioritize precision over recall, and the smaller models' efficiency allows multiple evaluations as new user data emerges (e.g., updated Weibo posts).

We hypothesize that this conservative tendency arises from the textometry approach used in constructing the training data and fine-tuning phase, which likely emphasizes fine-grained semantic distinctions and promotes selective predictions. This will be examined further in Section 6.5.2 Ablation Studies.

6.3 Text Generation quality

We used our optimization pipeline to fine-tune two lightweight models, reducing textual hallucinations and improving text quality. Typically, generated text is assessed with n-gram-based metrics like BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004), but these methods are sensitive to expression or format differences. Moreover, our gold reference texts, which are structured analyses containing many annotations of depressive expressions (words/phrases), make n-gram matching unreliable. Since we focus on semantic similarity between generated and reference texts, we adopt BERTScore (Zhang et al., 2020) as our primary metric, and includ ROUGE-1 and BLEU as complementary references, with evaluation results detailed in Table 3.

Model	**BertScore**	**BLEU**	**ROUGE1**
Llama3-8B-CN-Chat	0.5925	1.2025	0.0944
Glm4-9B-chat	0.5705	1.8607	0.2059
Qwen2.5-7B-Instr.	0.6401	1.4009	0.2025
Qwen2.5-0.5B-Instr.	0.2857	3e-99	0.0000
Qwen2.5-1.5B-Instr.	0.5774	0.9306	0.0855
GPT-4o	0.6548	2.2707	0.2970
GPT-4o Mini	0.6360	1.7500	0.1012
DeepSeek-R1-671B	**0.7084**	10.8407	**0.4076**
TML-0.5B	0.6919	13.1084	0.0548
TML-1.5B	0.6866	**13.4499**	0.0702

Table 3: Quality Evaluation of Generated Text.

For the test samples, we randomly selected 15 positive and 15 negative users from the SWDD SR Test dataset. Two psychology experts compiled and cross-validated the gold reference texts by summarizing and analyzing the user data. They were required to make judgments about the users' risk of depression based on the data, annotate and categorize key phrases or expressions according to DSM-5 standards, and provide a brief evaluative summary for each category. Subsequently, we extracted the generated texts from all previously mentioned models and computed their BERTScores against each gold reference text. The results show that, aside from the extremely large-parameter DeepSeek R1-671B, our TML0.5B and 1.5B models achieved excellent scores on BERTScore.

	Tokens/S	Memory (Mb)
Glm4-9B-chat	32	18773
Llama3-8B-Chinese-Chat	34	20231
Qwen2.5-7B-Instruct	43	15588
TML-0.5B	56	1188
TML-1.5B	48	3428

Table 4: Efficiency Comparison of Different Models.

6.4 Efficiency Comparison

The efficiency comparison in Table 4 shows that our models, TML-0.5B and TML-1.5B, outperform larger models in both generation speed and memory usage. TML-0.5B generates 56 tokens per second, much faster than GLM4-9B-chat (32 tokens/s), Llama3-8B (34 tokens/s), and Qwen2.5-7B (43 tokens/s), while using only 1188 Mb of memory, considerably less than the larger models. TML-1.5B also performs efficiently, generating 48 tokens per second and using 3428 Mb of memory.

6.5 Ablation Studies

In our ablation studies, we focus on two key factors that potentially influence the model's performance:

6.5.1 Impact of Fine-Tuning Order

We are particularly interested in exploring the effects of an alternative fine-tuning order, as it aligns more closely with a human-like reasoning process. In this setup, the model first analyzes the six dimensions of the depressive text before making a binary decision about whether the text contains depressive emotions. This order reflects how humans typically approach such tasks: by first examining various aspects of the text (such as psychological state, medical expressions, and external causes) before drawing a final conclusion.

Table 5 highlights key differences between the two fine-tuning orders. For the 0.5B models, we

Model	SWDD self-reported				SWDD				WU3D			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
TML-0.5B_multi-dims_first	0.77	0.76	0.81	0.79	0.76	0.80	0.75	0.77	0.76	0.81	0.68	0.75
TML-1.5B_multi-dims_first	0.86	0.92	0.82	0.87	0.81	0.90	0.69	0.78	0.81	0.88	0.72	0.79
TML-1.5B_random	0.88	0.85	0.93	0.89	0.81	0.81	0.82	0.81	0.83	0.81	0.86	0.83
TML-1.5B_standard	0.83	0.78	0.92	0.84	0.80	0.75	0.91	0.82	0.72	0.96	0.46	0.62

Table 5: Results of Ablation Studies.

observe a clear accuracy gap: the classification first approach consistently performs better than the six dimensions first approach. For instance, on the SWDD self-reported dataset, TML 0.5B achieves 0.81 accuracy, outperforming TML 0.5B trained with six dimensions first, which reaches 0.77. This pattern holds across other datasets, suggesting that for smaller models, starting with the simpler classification task supports more robust learning.

In contrast, the accuracy gap narrows significantly with the 1.5B models. Although classification first still performs slightly better, the difference is minimal. On SWDD, TML 1.5B achieves 0.87 accuracy, while the six dimensions first version closely follows at 0.86. This indicates that larger models are less sensitive to fine-tuning order, although classification first retains a slight advantage.

For both model sizes, precision remains consistently higher with the classification first strategy. On SWDD, TML 1.5B reaches 0.94 precision compared to 0.92 with six dimensions first. Similarly, TML 0.5B achieves 0.83 precision versus 0.76 with six dimensions first.

6.5.2 Impact of user selection method

In our initial experimental setup, we used the Lafon index to identify users whose texts showed statistical features across each of the six dimensions, ensuring that selected samples accurately represented traits in these areas. This method helped the model emphasize the most relevant characteristics per dimension, whereas randomly selected training data might fail to capture these distinct patterns effectively. Therefore, we were particularly interested in assessing the impact of a randomly selected training set.

For a controlled comparison, we replaced the Lafon-selected depressive users in the original training set with an equal number of randomly selected depressive users, keeping all negative samples unchanged. We then analyzed these randomly selected users' texts using the same six-dimensional framework—covering psychological state, medical expressions, clinical symptoms, and more—to maintain consistency in data preparation. All other experimental parameters, including the fine-tuning method and order, remained the same.

The results showed that models trained on Lafon-selected users achieved higher precision but lower recall, making more conservative yet accurate judgments. In contrast, models trained on randomly selected users had higher recall, identifying more positive cases but at the cost of precision. This suggested that Lafon-selected users had clearer characteristic patterns across one or more dimensions, whereas random selection offered greater generalizability. The Lafon-based model tended to predict only when distinctive characteristics were present in the six dimensions, leading to fewer false positives but more false negatives, while the model trained on randomly selected users made broader decisions, with less focus on clear or typical cases.

A promising approach would be to combine both strategies, mixing Lafon-selected users with randomly chosen ones. This hybrid method could balance precision and recall, improving overall model performance by capturing both distinctive and general user characteristics.

7 Conclusion and Future Work

We introduced the **TMLs** family of lightweight depression detection systems for Chinese social media texts (currently 0.5B and 1.5B parameters). Despite their modest size, the models achieve performance comparable to much larger LLMs while operating efficiently on consumer-grade GPUs and requiring no cloud upload of user data. This makes TMLs well suited for scenarios with limited computational resources and heightened privacy constraints. Future work includes evaluating the TMLs models on additional social platforms and extending them to a wider range of mental health conditions.

Ethical considerations All experiments use publicly released, expert-annotated, and further anonymised Weibo corpora. The system is intended solely for risk screening and cannot replace professional diagnosis.

References

American Psychiatric Association et al. 2000. Diagnostic and statistical manual of mental disorders. *Text revision*.

Priyanka Arora and Parul Arora. 2019. Mining twitter data for depression detection. In *2019 international conference on signal processing and communication (ICSC)*, pages 186–189. IEEE.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. 1996. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, 67(3):588–597.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480.

Chengkun Cai, Xu Zhao, Haoliang Liu, Zhongyu Jiang, Tianfang Zhang, Zongkai Wu, Jenq-Neng Hwang, and Lei Li. 2024. The role of deductive and inductive reasoning in large language models. *arXiv preprint arXiv:2410.02892*.

Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, and Wei Gao. 2023. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, 217:119538.

Corinna Cortes. 1995. Support-vector networks. *Machine Learning*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephan Dreiseitl and Lucila Ohno-Machado. 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359.

Zhiyuan Du, Yuxian Qian, Xiao Liu, Ming Ding, Yuxian Qian, Xiaoyang Liu, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335. Association for Computational Linguistics.

Egle Eensoo and Mathieu Valette. 2012. Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des

sentiments. In *TALN 2012*, volume 2, pages 367–374. GETALP-LIG.

Shang Gao, Mohammed Alawad, M Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B Durbin, Jennifer Doherty, Antoinette Stroup, et al. 2021. Limitations of transformers on clinical text classification. *IEEE journal of biomedical and health informatics*, 25(9):3596–3607.

GLMTeam, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. 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*, pages 15–24, Online. Association for Computational Linguistics.

Jinpeng Hu, Tengteng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, and Meng Wang. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation.

Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6:1–12.

Shaoxiong Ji, Tianyu Zhang, Laith Ansari, Jie Fu, Piyush Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190. European Language Resources Association.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Xiaochong Lan, Yiming Cheng, Li Sheng, Chen Gao, and Yong Li. 2024. Depression detection on social media with large language models. *arXiv preprint arXiv:2403.10750*.

Lei Li. 2024. Cpseg: Finer-grained image semantic segmentation via chain-of-thought language prompting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 513–522.

Lei Li, Sen Jia, Wang Jianhao, Zhongyu Jiang, Feng Zhou, Ju Dai, Tianfang Zhang, Wu Zongkai, and Jenq-Neng Hwang. 2024. Human motion instruction tuning. *arXiv preprint arXiv:2411.16805*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Keshu Malviya, Bholanath Roy, and SK Saritha. 2021. A transformers approach to detect depression in social media. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 718–723. IEEE.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Josiane Mothe, Faneva Ramiandrisoa, and Md Zia Ullah. 2022. Comparison of machine learning models for early depression detection from users' posts. In *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*, volume 1018 of *Studies in Computational Intelligence book series (SCI)*, pages 111–139. Springer International Publishing.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bénédicte Pincemin. 2022. Sémantique textométrique. *La sémantique au pluriel. Théories et méthodes*, pages 373–396.

QwenTeam. 2024. Qwen2.5: A party of foundation models.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

André Salem. 1986. Segments répétés et analyse statistique des données textuelles. *Histoire & mesure*, pages 5–28.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Jingzhe Shi, Jialuo Li, Qinwei Ma, Zaiwen Yang, Huan Ma, and Lei Li. 2024. Chops: Chat with customer profile systems for customer service with llms. *arXiv preprint arXiv:2404.01343*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jifei Wang, Zhenping Zhao, Jing Yang, Limin Wang, Mei Zhang, and Maigeng Zhou. 2024a. The association between depression and all-cause, cause-specific mortality in the chinese population—china, 2010–2022. *China CDC Weekly*, 6(40):1022.

Shenzhi Wang, Yaowei Zheng, Guoyin Wang, Shiji Song, and Gao Huang. 2024b. Llama3-8b-chinese-chat (revision 6622a23).

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Yiding Wang, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. 2020. A multimodal feature fusion-based method for individual depression detection on sina weibo. In *2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC)*, pages 1–8.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024c. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Dong Xue* Xin Yan. 2023. Mindchat: Psychological large language model. `https://github.com/X-D-Lab/MindChat`.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024b. Mentallama: interpretable mental health analysis on

social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Xingwei Yang, Rhonda McEwen, Liza Robee Ong, and Morteza Zihayat. 2020. A big data analytics framework for detecting user-level depression from social networks. *International Journal of Information Management*, 54:102141.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

Wei Zhai, Hongzhi Qi, Qing Zhao, Jianqiang Li, Ziqi Wang, Han Wang, Bing Yang, and Guanghui Fu. 2024. Chinese MentalBERT: Domain-adaptive pre-training on social media for Chinese mental health text analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10574–10585, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.