Large Language Model-Based Automated Assessment of Depressive Texts on Social Media: Using the Depression Dimension of the DASS-21 as the Theoretical Framework

Authors: Tuo Xin, Zhang Ji, Jingwen Zhou, Zhu Tingshao, Zhu Tingshao

Date: 2026-01-24T07:23:57+00:00

Abstract

Objective This study aimed to develop and validate an automated evaluation framework for social media texts based on the depression dimension of the DASS-21. Methods A total of 304 Chinese social media posts were collected from Weibo, Zhihu, Bilibili, Douban, and the comment section of NetEase Cloud Music. Drawing on the seven core psychological constructs of the DASS-21, we developed an annotation guideline for depression severity tailored to the Chinese online context. After systematic training, three psychology-major annotators conducted a trial annotation of 50 texts, on the basis of which a human gold-standard dataset of 254 texts was established. Using a chain-of-thought prompting strategy, we compared the performance of two open-source large language models, Qwen3-14b and Llama3.1-6b. Results Inter-rater consistency among human annotators was good (ICC(2,1) = 0.84, Kendall' s W = 0.819). Qwen3-14b demonstrated higher consistency with the human gold standard than Llama3.1-6b (ICC = 0.827 vs. 0.801), with particularly strong performance in identifying both absence of depressive symptoms and severe depressive symptoms. Limitations The models still exhibit limitations in understanding Chinese online subcultures, poetic and metaphorical expressions, and context-specific psychological vulnerability. Conclusions This study verifies the effectiveness of large language models pretrained on Chinese corpora for depression assessment in social media, providing a methodological foundation and empirical evidence for building low-cost, high-coverage digital mental health monitoring systems.

Full Text

Automated Assessment of Depression in Social Media Texts Using Large Language Models: A DASS-21 Depression Dimension Framework

Tuo Xin[12], Zhang Ji[12], Zhou Jingwen[12], Zhu Tingshao[12]*

[1](Institute of Psychology, Chinese Academy of Sciences, Beijing 100101)
[2](Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049)

Abstract

Objective: This study aimed to construct and validate an automated assessment framework for depression in social media texts based on the DASS-21 depression dimension.

Methods: We collected 304 Chinese social media texts from Weibo, Zhihu, Bilibili, Douban, and NetEase Cloud Music comment sections. Based on seven core psychological constructs from the DASS-21, we developed annotation guidelines for depression severity tailored to Chinese online contexts. Following systematic training, three psychology-trained annotators performed trial annotations on 50 texts, establishing a gold-standard dataset of 254 manually annotated texts. Using chain-of-thought prompting strategies, we compared the performance of two open-source large language models, Qwen3-14b and Llama3.1-6b.

Results: Inter-annotator agreement was strong (ICC(2,1) = 0.84, Kendall' s W = 0.819). The Qwen3-14b model demonstrated superior consistency with the human gold standard compared to Llama3.1-6b (ICC = 0.827 vs. 0.801), particularly in identifying non-depressive and severely depressive symptoms.

Limitations: Models still struggle with understanding Chinese internet subcultures, poetic metaphorical expressions, and psychological vulnerability in specific contexts.

Conclusion: This study validates the effectiveness of Chinese-pretrained large language models for social media depression assessment, providing methodological foundations and empirical evidence for constructing low-cost, high-coverage digital mental health monitoring systems.

Keywords: Depression; DASS-21; Large Language Models; Social Media; Chinese Text Analysis

1.1 Research Background: Epidemiology of Depression and Assessment Challenges

Depression has become one of the most prevalent mental disorders globally. According to World Health Organization data, approximately 280 million people worldwide suffer from depression (World Health Organization, 2023). In China,

accelerated social transformation and rapid lifestyle changes have contributed to rising depression prevalence, particularly among adolescents and young adults. A meta-analysis encompassing 93,679 Chinese university students revealed a depression prevalence rate of 34.70% (Deng et al., 2024), significantly higher than the general population. The COVID-19 pandemic further exacerbated this trend, with prevalence rising to 38.7% during and after the pandemic (Li et al., 2023). Prolonged social isolation, fear of infection, and disrupted routines have triggered widespread death anxiety and existential distress (Zhang & Ma, 2022).

Traditional depression assessment relies primarily on standardized self-report scales such as the Beck Depression Inventory, Patient Health Questionnaire, and Depression Anxiety Stress Scales. These instruments offer structured, quantifiable, and standardized advantages, forming the cornerstone of epidemiological surveys and clinical screening. However, scale-based assessment faces numerous practical challenges. Self-report scales require respondents to possess adequate introspective ability and willingness to disclose their psychological states. In cultural contexts where mental health stigma remains prevalent, many individuals tend to conceal genuine symptoms, leading to social desirability bias and false-negative results (Eisenberg et al., 2009). Scales provide retrospective, static assessments that cannot capture dynamic fluctuations in mood states or early warning signals (Torous et al., 2016). Critically, individuals with severe depression who experience social withdrawal often lack motivation to actively seek help or complete questionnaires, creating assessment blind spots (Shen et al., 2018).

With the proliferation of mobile internet and social media, individuals leave extensive behavioral traces and linguistic expressions in digital spaces. "Digital phenotyping" refers to the method of inferring psychological states by analyzing naturally occurring digital behavioral data (Insel, 2017). Compared to traditional scales, natural language expressions on social media offer unique advantages: users spontaneously generate content in authentic life contexts without artificial measurement influences, better reflecting genuine psychological states and emotional experiences (Harrigian et al., 2020); the immediacy of social media enables continuous monitoring, capturing subtle emotional fluctuations and early warning signals (Guntuku et al., 2017); the relative anonymity and physical distance of online environments make individuals more willing to express suppressed feelings, particularly negative emotions and sensitive topics (Suler, 2004).

Research has demonstrated significant associations between linguistic features in social media texts and depressive symptoms. Depressed individuals tend to use more first-person singular pronouns, reflecting excessive self-focus (Rude et al., 2004); more absolute terms, manifesting cognitive distortion through overgeneralization (Eichstaedt et al., 2018); and higher frequencies of negative emotion words and death-related vocabulary (Coppersmith et al., 2015). However, existing research predominantly relies on English corpora, with insufficient

attention to unique linguistic features and cultural contexts of Chinese social media, limiting cross-cultural generalizability.

1.2 Theoretical Framework: Psychometric Foundations of DASS-21 Depression Dimension

To ensure theoretical validity of automated assessment, this study adopted the Depression Anxiety Stress Scales-21 (DASS-21) depression dimension as the core reference framework. The DASS-21 was developed by Lovibond and Lovibond (1995) based on the tripartite model of emotions, which was proposed by Clark and Watson (1991) to differentiate between depression and anxiety—two highly overlapping emotional disorders in clinical presentation.

The tripartite model deconstructs emotional disorders into three independent yet correlated dimensions: general negative affect as a non-specific factor common to both depression and anxiety, manifested as general subjective distress and unpleasantness; physiological hyperarousal as an anxiety-specific factor involving autonomic nervous system overreaction; and low positive affect or anhedonia as the core depression-specific factor. Unlike simple sadness or distress, depression is defined as the absence of happiness, vitality, interest, and confidence—representing functional inhibition of the positive affect system (Watson et al., 1995). This theoretical framework's key contribution lies in distinguishing shared from specific factors, providing a conceptual foundation for differential diagnosis and theoretical guidance for measurement tool construction.

Compared to DSM-based scales, the DASS-21 depression subscale offers unique advantages. It deliberately excludes somatic symptoms such as fatigue, sleep, and appetite disturbances, as these may arise from various non-pathological factors and lack specificity in certain populations, potentially causing false positives (Henry & Crawford, 2005). By focusing on core cognitive-affective symptoms —anhedonia, hopelessness, and self-deprecation—the scale demonstrates higher discriminant validity in distinguishing normal stress responses from pathological depression. Furthermore, DASS-21 employs continuous scoring rather than categorical diagnosis, better aligning with the epidemiological reality of depressive symptoms following a normal distribution in the population (Haslam et al., 2012), and facilitating identification of subclinical depression and early warning signals. The DASS-21 has demonstrated excellent reliability and validity in Chinese populations, with internal consistency coefficients typically above 0.90 and confirmatory factor analysis supporting its three-factor structure (Gong et al., 2010; Wang et al., 2016).

The DASS-21 depression subscale comprises seven items measuring anhedonia (“I couldn’ t seem to experience any positive feeling at all”), lack of initiative (“I found it difficult to take initiative”), hopelessness (“I felt that life was meaningless”), core emotional symptoms (“I felt sad and depressed”), loss of interest (“I was unable to become enthusiastic about anything”), self-deprecation (“I felt I was pretty worthless”), and existential emptiness (“I felt that life was

meaningless"). These seven items collectively constitute the cognitive-affective symptom cluster of depression, encompassing multiple levels from basic emotional experiences to complex cognitive evaluations. Anhedonia is considered a core neurobiological marker of depression, closely associated with dopaminergic dysfunction in the ventral striatum and prefrontal cortex (Treadway & Zald, 2011), while hopelessness serves as an important predictor of suicide risk (Beck et al., 1990). A core task of this study involves exploring how these abstract psychological constructs manifest in specific expression patterns within natural social media language.

1.3 Depression Discourse Characteristics in Chinese Social Media

China' s social media ecosystem exhibits platform diversity and functional differentiation. This study collected data from five mainstream platforms: Weibo, Zhihu, Bilibili, Douban, and NetEase Cloud Music comment sections. These platforms differ in user demographics, interaction modes, and content characteristics, collectively forming the ecological landscape of Chinese online depression discourse. Platform-specific architectural affordances shape differentiated forms of depression expression. Weibo' s immediacy and public nature make it a primary channel for emotional venting, where users frequently confide pain in "tree hole" posts—texts that are immediate, unpolished, and directly correspond to DASS-21' s hopelessness and meaninglessness dimensions (Huang et al., 2020). Zhihu' s Q&A community attributes encourage lengthy narratives and rationalized expressions, with depression discourse exhibiting rationalized features as users tend to review diagnostic experiences and analyze symptom mechanisms through extensive autobiographical accounts involving self-reflection and attribution analysis (Ji et al., 2021). Bilibili' s bullet comments and comment sections carry emotional resonance among youth groups; Douban groups provide relatively private spaces for confession; NetEase Cloud Music comment sections frequently feature emotional short sentences and poetic expressions, where users reveal late-night loneliness within musical contexts.

In Chinese online contexts, "Sang culture" constitutes a major confounding variable for depression identification. "Sang" originally referred to a decadent, negative life attitude, but has evolved in youth subculture into a defensive pessimism and group identity symbol (Sun & Wang, 2019). For instance, a user posting "I' m just trash" may be expressing clinical self-deprecation as measured by DASS-21, or participating in self-deprecating interactions within specific online communities to alleviate competitive pressure and gain group belonging. This pragmatic ambiguity requires assessment systems to possess deep contextual understanding capabilities to distinguish "Sang" as cultural performance from "depression" as pathological symptom. Key distinguishing criteria include: pathological depression manifests as long-term, recurrent negative experiences, while cultural "Sang" primarily constitutes transient complaints; genuine depression accompanies clear social withdrawal and decreased daily functioning,

whereas cultural expressions often do not affect actual life; pathological depression contains profound subjective pain and helplessness, with despair detectable even in calm tones, while cultural "Sang" may carry connotations of ridicule, irony, or detachment.

1.4 Large Language Models in Mental Health Assessment

Traditional natural language processing methods primarily rely on bag-of-words models, shallow semantic feature extraction tools, or machine learning-based classifiers. These approaches exhibit significant limitations when handling context-dependent structures, negations, irony, and metaphors, potentially misclassifying "I no longer feel sad" as containing sadness (Calvo et al., 2017). Moreover, traditional methods typically require large amounts of manually annotated data for supervised learning, with limited generalization capabilities in new domains or small-sample scenarios.

The emergence of large language models represents a paradigm shift. Through self-supervised pretraining on massive text corpora, these models acquire deep semantic representations, world knowledge, and commonsense reasoning capabilities (Brown et al., 2020). In mental health applications, large language models demonstrate unique advantages: they can perform complex psychological assessment tasks without large-scale supervised training for specific tasks, simply by providing task definitions, evaluation criteria, and few examples in prompts (Yang et al., 2023); through specific prompt engineering strategies, they can generate explicit reasoning processes, simulating step-by-step clinical analysis—identifying textual evidence, matching psychological constructs, integrating multiple cues, and providing justification (Wei et al., 2022); they can understand complex grammatical structures, recognize metaphors and metonymy, and infer implicit information, providing advantages in handling indirect expressions in depression discourse.

However, applying large language models to mental health assessment still faces challenges. These models may generate factually inaccurate or over-inferred content, manifesting in psychological assessment as "hallucinating"symptoms not present in the text or over-interpreting ambiguous expressions (Ji et al., 2023). Additionally, most large language models are pretrained on English corpora and may lack deep understanding of Chinese-specific expression habits and internet subcultures, leading to misjudgment of phenomena like "Sang culture" (Yang et al., 2024). Therefore, this study employs rigorous human gold standards as benchmarks to systematically evaluate the reliability and validity of large language models in DASS-21 depressive symptom identification tasks.

1.5 Research Objectives

Based on the above background, this study aims to construct and validate an automated assessment framework for social media texts based on the DASS-21 depression dimension. Specifically, this study will: (1) develop a set of depres-

sion severity annotation guidelines for Chinese social media texts based on the DASS-21 theoretical framework and clinical operational definitions; (2) establish a high-quality human-annotated gold-standard dataset through multi-round iterative training and consistency testing; (3) compare the performance of two mainstream open-source large language models—Qwen3-14b and Llama3.1-6b—on depressive symptom identification tasks using chain-of-thought prompting strategies; (4) evaluate their consistency with human gold standards using multidimensional reliability and validity indicators to comprehensively examine the reliability and construct validity of model scoring.

This study systematically applies the DASS-21 theoretical framework to Chinese social media text analysis, bridging the gap between standardized psychological measurement and ecological digital behavioral data. It explicitly incorporates discrimination criteria between cultural expressions and pathological symptoms in annotation guidelines, addressing unique Chinese internet phenomena like "Sang culture." The study employs standardized psychometric procedures to ensure reliability. The developed assessment framework and annotated dataset can provide technical support for applications including social media mental health monitoring, suicide risk warning systems, and digital psychological interventions, holding significant public health value.

2.1 Data Sources and Collection

Our corpus originated from five mainstream Chinese social media platforms: Weibo, Zhihu, Bilibili, Douban, and NetEase Cloud Music comment sections. Data collection was completed by three systematically trained research team members between October and November 2025.

Regarding sampling strategy, this study employed manual purposive sampling to avoid data noise issues associated with automated crawling. Researchers browsed and screened texts in emotion-related topic areas, mental health discussion zones, and "tree hole" comment sections across platforms. Inclusion criteria comprised: first-person narratives explicitly expressing the author's emotional state; text length between 100-500 characters; and semantically complete, comprehensible content. Exclusion criteria included: obvious commercial advertisements, help-seeking redirects, or emotion-unrelated political/entertainment comments; pure lyric excerpts or film dialogue quotes (unless accompanied by personal autobiographical elaboration). Notably, to ensure sample representativeness and scoring standard completeness, this study included both texts exhibiting depressive features and those without obvious depressive characteristics. Following this screening process, we initially obtained 304 valid texts. During the trial annotation phase, 50 texts were randomly selected for annotator training and guideline revision. The formal annotation phase utilized the remaining 254 texts, constituting the final analytical corpus for this study.

2.2 Human Rating Procedure

This study developed annotation guidelines for depression severity in Chinese social media texts based on the DASS-21 depression dimension (see Appendix). The guidelines operationalized the seven psychological constructs measured by DASS-21 depression subscale—anhedonia, lack of initiative, hopelessness, sadness/depression, loss of interest, self-deprecation, and meaninglessness—into operational definitions for social media texts, providing typical linguistic marker examples for each construct in Chinese online contexts. Ratings employed a 0-3 Likert scale: 0 indicating no depressive cues, 1 indicating mild depressive tendency, 2 indicating moderate depressive tendency, and 3 indicating severe depressive tendency. The guidelines clarified core characteristics and typical examples for each level to guide consistent judgment.

The guidelines established several core annotation principles. The "Pathology Priority Principle"emphasizes that raters should identify depressive pathological cores beneath surface emotional intensity, distinguishing performative venting from genuine symptomatic expression. "Context and Temporal Consideration" requires raters to focus on the duration of distress experience and degree of functional impairment rather than judging based solely on instantaneous reactions. The "Culture and Subculture Exclusion" principle addresses Chinese internet-specific phenomena like "Sang culture" and "hysterical literature," requiring raters to filter entertainment or self-deprecating components to avoid misclassifying cultural expressions as pathological symptoms.

Annotation was completed by three psychology graduate students with backgrounds in clinical or health psychology coursework. Before formal annotation, all raters underwent standardized training covering: theoretical foundations of DASS-21 and psychometric properties of the depression dimension; detailed interpretation of annotation guidelines and understanding of operational definitions; discussion and calibration of typical cases; and judgment strategies for common boundary situations.

The annotation process comprised trial and formal phases. In the trial phase, three raters independently scored 50 randomly selected texts using the initial guideline version to test operability and inter-rater consistency. Following trial annotation, the research team held calibration meetings to discuss cases with substantial scoring discrepancies, identify boundary issues requiring clarification, and revise the guidelines accordingly. The revised guideline (Version 2) clarified three main aspects: first, the definition of score 0 was tightened from "no obvious depressive cues or only minor cues"to "absolutely no depressive cues" to improve discriminability and consistency; second, guidance for discriminating depression from anxiety and stress was added, explicitly stating that texts primarily showing tension, worry, or pressure without core depressive experiences like hopelessness, anhedonia, or self-deprecation should not receive high depression scores; third, scoring weights for somatic symptoms and suicidal/self-harm ideation were specified, emphasizing that while these indicators are important

references, they should not serve as sole determinants for high scores but must be integrated with core depressive symptoms like hopelessness, worthlessness, and functional impairment.

In the formal annotation phase, the remaining 254 texts were independently blind-rated by two annotators using the revised guidelines. When ratings aligned, that score served as the final human gold standard; when discrepancies occurred, the two raters reached consensus through discussion, with the negotiated score serving as the final gold standard.

2.3 Large Language Model Evaluation

This study selected two open-source large language models for evaluation: Qwen3-14b and Llama3.1-6b. Selection criteria included: moderate parameter scale enabling local deployment and batch inference; Qwen series models' strong Chinese pretraining foundation, while Llama series serves as an internationally widely-used benchmark, with comparison between them helping examine performance differences between models with different linguistic backgrounds on Chinese depression text identification tasks.

For prompt design, this study employed chain-of-thought prompting strategies to construct the prompt framework, which improves accuracy on complex judgment tasks by guiding models through step-by-step reasoning. The prompt comprised three core components: role-setting positioned the model as "a professional mental health assessment expert specializing in analyzing depressive tendencies in user self-reported texts from social media platforms, strictly following DASS-21 depression dimension annotation guidelines" ; task definition transformed DASS-21' s seven items into detection instructions for texts, requiring the model to assess the degree of depressive emotion expression and provide 0-3 scores; the chain-of-thought component required the model to explicitly output its reasoning process before final scoring, including identified depressive cues, matched psychological constructs, and functional impairment evidence, to enhance score interpretability and facilitate subsequent error analysis.

During model inference, 254 texts were sequentially submitted to both large language models. To reduce output randomness and improve scoring stability, the temperature parameter was set to 0.2. Model outputs included reasoning processes and final 0-3 depression severity scores, which were automatically extracted from output texts using regular expressions for subsequent consistency analysis with human gold standards.

2.4 Data Analysis

This study employed multiple statistical indicators to evaluate annotation quality and large language model performance from reliability and validity perspectives. For reliability assessment, Intraclass Correlation Coefficient (ICC) served as the core indicator. ICC evaluates absolute agreement among raters, making it

more appropriate for scoring task reliability than correlation coefficients focusing solely on relative ranking. This study calculated multiple ICC types including ICC(1), ICC(2), and ICC(3) for comprehensive evaluation, where ICC(2,1) using a two-way random effects model reflects absolute agreement of individual raters and represents the most commonly used reliability indicator. ICC interpretation standards are: <0.50 poor reliability, 0.50-0.75 moderate reliability, 0.75-0.90 good reliability, >0.90 excellent reliability (Koo & Li, 2016).

In addition to ICC, this study calculated Cohen' s Kappa coefficients (including linear and quadratic weighting) to assess classification agreement. Weighted Kappa considers severity of disagreement between categories, with quadratic weighting imposing greater penalties on larger discrepancies, making it more suitable for ordinal data (Fleiss & Cohen, 1973). Kendall' s W coefficient of concordance was used to evaluate overall coordination among multiple raters, with values ranging from 0-1, where higher values indicate greater consistency in rating rankings.

Validity assessment employed Spearman rank correlation and Pearson correlation to evaluate association strength between model scores and human gold standards. Spearman coefficient is suitable for ordinal data, reflecting monotonic relationships; Pearson coefficient measures linear correlation. Correlation interpretation standards are: 0.10-0.29 weak, 0.30-0.49 moderate, 0.50-0.69 strong, $ $0.70 very strong (Cohen, 1988).

Additionally, confusion matrix analysis examined model accuracy and misclassification patterns across depression levels. Diagonal elements represent correctly classified samples, while off-diagonal elements reveal systematic bias directions (over- or underestimation). Based on confusion matrices, accuracy rates for each level were calculated as direct performance indicators.

All statistical analyses were completed in R 4.3.0 environment, using the psych package for ICC and Kappa calculations, irr package for Kendall' s W, with significance level set at $\alpha = .05$.

3.1.1 Inter-Annotator Agreement (50 Texts)

Before formal annotation, the research team systematically trained three annotators, clarifying operational definitions for the 0-3 four-level scoring system: 0 indicating no depressive symptom expression, 1 indicating mild depressive symptom expression, 2 indicating moderate depressive symptom expression, and 3 indicating severe depressive symptom expression. Training involved trial scoring of 50 randomly sampled texts from the corpus using non-replacement sampling to calculate inter-annotator consistency metrics, with results as follows.

As shown in Table 1 , this study calculated multiple types of Intraclass Correlation Coefficients to comprehensively evaluate annotator consistency. Single-rater absolute agreement ICC(1) = 0.83, 95% CI [0.75, 0.89], F(49, 100) = 16, $p < .001$; single random rater ICC(2,1) = 0.84, 95% CI [0.75, 0.89], F(49, 98) =

16, p < .001; single fixed rater ICC(3,1) = 0.84, 95% CI [0.75, 0.90], F(49, 98) = 16, p < .001. Average-rater ICC values reached 0.94, 95% CI [0.90, 0.96], p < .001, indicating very strong inter-annotator consistency.

Other consistency metrics showed Kappa coefficient $\kappa = 0.57$, p < .001, indicating moderate agreement, and Kendall' s coefficient of concordance W = 0.819, p < .001, indicating strong coordination among the three annotators.

Annotator correlation analysis is shown in Figure 1 [Figure 1: see original paper]. Pearson correlation results indicated r = .81 between Rater 1 and Rater 2, r = .88 between Rater 1 and Rater 3, and r = .85 between Rater 2 and Rater 3. The average Pearson correlation among three raters was 0.85, indicating high linear correlation.

Rater score distributions and inter-text rating differences are shown in Figure 2 [Figure 2: see original paper] and Figure 3 [Figure 3: see original paper]. The three raters' score distributions exhibited similar patterns, concentrated within the 0-3 range. Rater 1' s scores centered around 2, Rater 2' s scores clustered in the 1-2 range, while Rater 3' s distribution was relatively broader. Figure 3 shows that among 50 texts, the vast majority had low mean absolute differences between raters, with only occasional texts showing substantial scoring discrepancies.

In summary, consistency metrics from this trial scoring round reached acceptable levels, with ICC(2,1) = 0.84 indicating strong agreement, supporting subsequent formal annotation work.

Table 1 Inter-Annotator Consistency Coefficients (50 Texts)

ICC Type	ICC Value	95% CI	p-value
Single-rater absolute agreement	0.83	[0.75, 0.89]	<.001
Single random rater consistency	0.84	[0.75, 0.89]	<.001
Single fixed rater consistency	0.84	[0.75, 0.90]	<.001
Average-rater absolute agreement	0.94	[0.90, 0.96]	<.001

Figure 1 [Figure 1: see original paper] Annotator Score Correlation Analysis (50 Texts)
Figure 2 [Figure 2: see original paper] Annotator Score Distribution (50 Texts)
Figure 3 [Figure 3: see original paper] Annotator Score Differences (50 Texts)

3.1.2 Final Gold Standard Establishment

Based on operational definitions of the DASS-21 depression dimension, following multiple trial rounds and consistency calibration, three raters proceeded with formal annotation after achieving acceptable inter-rater agreement. During annotation, texts with substantial scoring discrepancies were discussed until consensus was reached, with operational definitions revised accordingly to improve

consistency. This process resulted in a gold-standard dataset of 254 manually annotated depression scores.

The gold-standard score distribution is shown in Figure 4 [Figure 4: see original paper]. The 254 texts exhibited an uneven distribution: 98 texts (38.58%) scored 0 (no depressive symptoms), representing the most frequent category; 72 texts (28.35%) scored 2 (moderate symptoms); 58 texts (22.83%) scored 1 (mild symptoms); and 26 texts (10.24%) scored 3 (severe symptoms). Overall, the gold-standard sample showed a right-skewed distribution, with low-depression texts (scores 0 and 1) comprising 61.41% of the sample, and moderate-to-high depression texts (scores 2 and 3) comprising 38.59%. This distribution aligns with the natural distribution of depression expression in social media texts while ensuring sufficient sample sizes across all levels to support subsequent large language model evaluation and statistical analysis.

Figure 4 [Figure 4: see original paper] Human Rating (Gold Standard) Distribution (254 Texts)

3.2.1 Qwen3-14b Evaluation Results

This study employed the Qwen3-14b large language model to automatically assess depression severity in 254 social media texts, with results compared against human gold standards. As shown in Table 3 , Qwen3-14b demonstrated strong consistency with human ratings, ICC(2) = 0.827, p < .001. Linear weighted Kappa $\kappa = 0.729$, p < .001; quadratic weighted Kappa $\kappa = 0.827$, p < .001, both indicating good agreement. Kendall' s tau $\tau = 0.777$, p < .001; Kendall' s W = 0.819, p < .001; Spearman rank correlation $\rho = 0.838$, p < .001. All metrics indicate high consistency between Qwen3-14b scores and human gold standards.

As shown in Figure 5 [Figure 5: see original paper], Qwen3-14b and human gold standards showed distribution differences across depression levels. For score 0 (no symptoms), model ratings accounted for 46.5%, higher than the gold standard' s 38.6%, indicating a tendency to misclassify mildly depressive texts as non-depressive. For score 1 (mild symptoms), model ratings accounted for 11.8%, lower than the gold standard' s 22.8%, suggesting insufficient sensitivity in identifying mild depressive symptoms. For score 2 (moderate symptoms), model ratings accounted for 26.8%, close to the gold standard' s 28.3%. For score 3 (severe symptoms), model ratings accounted for 15.0%, higher than the gold standard' s 10.2%, indicating some degree of over-pathologizing tendency.

Figure 6 [Figure 6: see original paper] presents the confusion matrix showing cross-distribution between Qwen3-14b scores and human gold standards. Diagonal values represent correctly identified samples: 93 correct at level 0, 25 at level 1, 48 at level 2, and 15 at level 3. Accuracy rates calculated from the confusion matrix showed highest accuracy (94.9%, 93/98) for gold-standard level 0 texts. Level 2 accuracy was 66.7% (48/72), and level 3 accuracy was 57.7% (15/26). Level 1 accuracy was relatively lowest at 43.1% (25/58).

The confusion matrix revealed primary misclassification patterns: at level 1, 22 cases (37.9%) were overestimated as level 2, showing tendency to overrate mild symptoms as moderate. At level 2, 16 cases were underestimated as level 0 and 10 as level 1, indicating some underestimation of moderate symptoms. At level 3, 10 cases were misclassified as level 2, suggesting limited discriminability between severe and moderate depression.

In summary, Qwen3-14b performed best in identifying non-depressive texts, while showing relatively weaker ability in identifying mild depressive symptoms (level 1) with tendency to overrate mild symptoms as moderate.

Figure 5 [Figure 5: see original paper] Qwen3-14b Score Stacked Histogram
Figure 6 [Figure 6: see original paper] Qwen3-14b Score Confusion Heat Matrix

3.2.2 Llama3.1-6b Evaluation Results

This study also employed the Llama3.1-6b large language model for automated depression assessment of 254 social media texts, with results compared against human gold standards. As shown in Table 3, Llama3.1-6b demonstrated good consistency with human gold standards, ICC(2) = 0.801, $p < .001$. Linear weighted Kappa $\kappa = 0.670$, $p < .001$; quadratic weighted Kappa $\kappa = 0.800$, $p < .001$, reaching good to excellent agreement levels.

Correlation metrics showed Kendall's $\tau = 0.745$, $p < .001$, Kendall's $W = 0.824$, $p < .001$, and Spearman rank correlation $\rho = 0.818$, $p < .001$, all indicating strong positive relationships between model scores and human gold standards.

As shown in figures, Llama3.1-6b and human gold standards showed distribution differences across four depression levels. For level 0, model ratings accounted for 39.8%, close to the gold standard' s 38.6%. For level 1, model ratings accounted for 27.2%, higher than the gold standard' s 22.8%. For level 2, model ratings accounted for 21.3%, lower than the gold standard' s 28.3%. For level 3, model ratings accounted for 11.8%, slightly higher than the gold standard' s 10.2%.

The confusion matrix showed diagonal values of correctly identified samples: 85 at level 0, 35 at level 1, 33 at level 2, and 6 at level 3. Accuracy rates were highest for level 0 (86.7%, 85/98), followed by level 1 (60.3%, 35/58), level 2 (45.8%, 33/72), and lowest for level 3 (23.1%, 6/26).

Misclassification patterns revealed: at level 2, 20 cases were overestimated as level 3 and 16 underestimated as level 1, showing bidirectional bias in moderate symptom identification. At level 3, 11 cases were misclassified as level 2 and 6 as level 1, indicating clear underestimation of severe depression. At level 1, 10 cases were underestimated as level 0, showing some underestimation of mild symptoms.

Figure 7 [Figure 7: see original paper] Llama3.1-6b Score Stacked Histogram
Figure 8 [Figure 8: see original paper] Llama3.1-6b Score Confusion Heat Matrix

3.2.3 Model Comparison

Compared to Qwen3-14b, Llama3.1-6b showed slightly weaker performance across all consistency metrics: ICC(2) (0.801 vs. 0.827), linear weighted Kappa (0.670 vs. 0.729), and Spearman correlation (0.818 vs. 0.838). In terms of per-level accuracy, Qwen3-14b outperformed Llama3.1-6b in level 0 (94.9% vs. 86.7%), level 2 (66.7% vs. 45.8%), and level 3 (57.7% vs. 23.1%), with particularly significant differences in severe depression identification. Llama3.1-6b only slightly exceeded Qwen3-14b in level 1 accuracy (60.3% vs. 43.1%).

Table 3 Consistency Between Large Language Models and Gold Standard

Metric	Qwen3-14b	Llama3.1-6b
ICC(2)	0.827***	0.801***
Kappa (Linear)	0.729***	0.670***
Kappa (Quadratic)	0.827***	0.800***
Kendall' s tau	0.777***	0.745***
Kendall' s W	0.819***	0.824***
Spearman ρ	0.838***	0.818***

Note: *** indicates $p < .001$.

In summary, Llama3.1-6b performed adequately in identifying non-depressive texts but showed significantly weaker ability than Qwen3-14b in identifying moderate and severe depressive symptoms, with particularly severe underestimation of severe depression.

3.3.1 Misjudgment of Exaggerated Expressions in Online Subculture Contexts

Large language models 容易产生误判 when processing texts with Chinese social media-specific expressive styles due to insufficient understanding of indigenous internet subcultures. For example: "Winter makes me mentally unhealthy. Thinking about spending autumn and winter here then flying ten hours to spend autumn and winter on the other side makes me want to die. What' s more divine is that after spending autumn and winter there until spring, I have to fly back to spend autumn and winter again, then fly back to spend autumn and winter. Where' s the 'life is short' they promised? I' m just having winter after winter after winter." This text employs typical Chinese social media "hysterical literature" style, using extreme vocabulary ("want to die"), repetitive rhetoric ("winter winter winter"), and self-questioning to create humorous effects. Human raters recognized it as expressing frustration and absurdity about uncontrollable life rhythms with limited emotional intensity, rating it 1. However, the model, triggered by the high-risk keyword "want to die," directly mapped literal semantics to severe depression signals, lacking capacity to understand irony and self-mockery in networked expression styles.

3.3.2 Misjudgment of Literal Interpretation of Rhetorical and Metaphorical Expressions

Large language models tend to underestimate depression risk when processing poetic expressions with implicit meaning, lacking sensitivity. For example: "I' ve been waiting for a moment of explosive crying. Intense emotions are too rare for me. More often I just sit there watching the rainstorm pour on me without ever truly getting wet. I detach too quickly···I start suspecting I might be a rotten egg. I' ve been waiting for that bottom that looks ready to collapse to actually break. Perhaps some things only reveal whether they' re life or stillbirth when shattered." This employs highly poetic, symbolic language (e.g., "rainstorm," "rotten egg," "stillbirth"), typical of restrained, highly introspective emotional expression common among well-educated individuals or those accustomed to suppressing emotions. Such groups often don' t cry for help but describe internal collapse in calm tones. Human raters can penetrate rhetoric to capture emotional undertones, while models, better at identifying straightforward, explicit distress statements, may miss this "quiet collapse" and misclassify it as non-depressive (0).

3.3.3 Insufficient Sensitivity to Specific Contexts

Large language models lack understanding of psychological vulnerability in specific contexts (e.g., perinatal period, long-term interpersonal trauma), underestimating situational impacts on mental health. For example: "Really exhausted, so tired. Became sensitive after pregnancy, constant fighting. Discovered my husband went to his ex-girlfriend' s place at 1 AM after a fight, returned home at 5 AM. What happened during those two hours? I really want to know. He said they just talked. I called his ex and she said he ran over to ask why she said hurtful things when breaking up. I don' t understand this psychology. So exhausted." Although lacking typical depression keywords, the text clearly conveys strong exhaustion, insecurity, and trust crisis, occurring during pregnancy —a highly sensitive, emotionally volatile period. Human raters integrated the background of "became sensitive after pregnancy," "really exhausted," and partner betrayal behavior to judge moderate emotional distress (2). The model, lacking understanding of perinatal psychological vulnerability, may classify it as ordinary interpersonal conflict based on surface narrative.

4 Discussion

This study explored using large language models to automatically identify depression in social media. Referencing DASS-21 depression criteria, we compared Qwen3-14b and Llama3.1-6b performance in Chinese contexts. Findings indicate that large language models show potential in depression identification, while also revealing significant differences between models in understanding exaggerated expressions in network subculture contexts, Chinese rhetoric and metaphor, and specific situational recognition accuracy.

4.1 Model Performance in Depression Identification

In Chinese social media depression assessment tasks, Qwen3-14b demonstrated superior performance over Llama3.1-6b. This advantage was evident not only in consistency metrics with human annotation but also in sensitivity to capturing core DASS-21 symptoms. Specifically, Qwen3-14b showed high accuracy in identifying non-depressive texts, effectively filtering daily emotional venting. In contrast, Llama3.1-6b exhibited marked accuracy decline in identifying severe depression. This means that applying Llama3.1-6b to suicide risk warning systems might miss many high-risk individuals, generating false-negative risks. This performance difference is not isolated; a study on China' s nursing qualification exam similarly found Qwen series models significantly outperformed GPT-4o and other models in tasks involving complex clinical decisions and Chinese medical terminology (Zhu et al., 2025). This suggests that in domains involving professional medical knowledge and indigenous culture (e.g., depression symptom descriptions, Chinese internet slang), the volume and distribution of Chinese data encountered during model pretraining are key determinants of identification accuracy.

The two models also showed distinct error patterns. Qwen3-14b exhibited clear "overestimation" tendency, easily identifying mild symptoms as moderate. This bias may stem from high sensitivity to negative vocabulary. In social media contexts, users frequently use high-arousal words like "collapse" or "want to die" to express temporary frustration from life stress. While Qwen3-14b can identify these words' literal meanings, it struggles to distinguish transient emotional reactions from pathological depressive states, thus adopting a "risk-averse" strategy to ensure identification sensitivity. From a public health screening perspective, this high sensitivity facilitates early warning (Ji et al., 2021), but may lead to overutilization of medical resources in actual diagnosis (Harrigian et al., 2020). In contrast, Llama3.1-6b showed "underestimation" tendency, frequently identifying severe depression as milder grades.

4.2.1 Interference from Online "Sang Culture" on Model Judgment

This study found that Chinese internet-specific "Sang culture" significantly interferes with model judgment. Models often cannot effectively distinguish "Sang" as subcultural performance from pathological "depression." "Sang" is not merely emotional expression but rather a self-protection mechanism (Li Xin & Peng Yi, 2020). Young people often use self-mockery (e.g., "I' m trash") to lower external expectations and alleviate anxiety (Du Junfei, 2017). Such expressions overlap literally with DASS-21' s "self-deprecation," but posters often maintain normal social function and even gain group belonging through such interactions (Guo Dongsheng & Wang Yuying, 2023).

Although large language models possess strong semantic understanding capabilities, when processing high-context texts, they still tend to prioritize explicit

negative literal meanings, establishing direct semantic mapping between high-arousal vocabulary like "don't want to live"and depression labels (Li et al., 2026). However, human experts incorporate contextual cues: texts filled with negative vocabulary but using numerous internet memes or humorous interactions are typically judged as "Sang" or "playing with memes" rather than depression. Llama3.1-6b, lacking accumulation of Chinese internet subculture data, can only perform literal interpretation, leading to misjudgment. In contrast, Qwen3-14b, exposed to extensive Chinese community corpora during pretraining, possesses some recognition capability for subcultural symbols like "hysterical literature," contributing to its higher accuracy in identifying non-depressive samples. However, this also reveals current technology' s limitations in processing high-context culture: large language models still struggle to fully understand the social-psychological motivations behind "irony" (Guo et al., 2025).

4.2.2 Differences in Understanding Metaphorical Expressions

The DASS-21 aims to assess core cognitive and affective dimensions of depression, yet in Chinese contexts, these abstract pathological experiences are often constructed through metaphor. This phenomenon is particularly prominent in online communities where depressed groups tend to use "deliberate metaphors" to concretize unspeakable psychological pain (Jing & Jiang, 2024). When metaphorical expressions are very close to literal meaning, models perform adequately; but when cross-domain inference is required (i.e., large literal distance but deep relevance), LLM performance declines significantly (Tong et al., 2024). In this study, Llama3.1-6b likely lacked Chinese contextual training and tended toward "literal interpretation" of metaphors, ignoring depression signals behind the text and failing to accurately assess patients' psychological states.

For example, one user described their state as "a potato slice repeatedly soaked and oxidized black." In Chinese context, this ontological mapping vividly points to DASS-21' s "self-deprecation" and "hopelessness." Model output showed Llama3.1-6b focused primarily on the entity noun "potato slice," misclassifying it as low-risk diet-related text. In contrast, Qwen3-14b' s reasoning records indicated it successfully penetrated literal appearance, explicitly identifying "persistent anhedonia" and "hopelessness" and recognizing implied "functional impairment" behind the metaphor.

However, existing models still face challenges with synesthetic metaphors possessing "high emotional granularity." For instance, a user wrote: "The thousands of pains I once wanted to describe, the moment I made a sound, slipped away like wind, becoming sand grains mixed in my throat, rendering me nearly speechless." This high-context narrative transforms psychological aphasia into physical foreign-body sensation (sand grains). Examination of Qwen3-14b' s reasoning process revealed that while the model could identify metaphors like "aphasia" and "sand grains in throat" suggesting severe cognitive and affective blockage, it

stopped at this general "blockage" qualification without further precisely mapping this complex cross-modal sensation to specific DASS-21 symptom dimensions (e.g., somatic swallowing difficulties in anxiety dimension or psychomotor retardation in depression dimension). This indicates that while current large language models have made progress in emotion classification tasks, they still exhibit subtle "cognitive granularity" gaps compared to human experts in understanding emotional texture and deep psychological mechanisms (Hu et al., 2025; Shu et al., 2025).

4.3 Research Limitations

Despite validating large language models' potential in depression assessment, several limitations must be considered when interpreting results:

First, human raters' scores possess subjectivity. This study's "gold standard" was established by three psychology graduate students. Although statistical results showed good inter-rater consistency, psychological state inference inherently involves unavoidable subjectivity. Therefore, the "gold standard" represents more of an "expert consensus" than absolute objective truth (Chancellor & De Choudhury, 2020).

Second, this study conducted "slice-based" assessment on single social media posts. This static analysis easily confuses temporary "emotional states" with long-term "depressive traits." For example, angry rants about daily hassles might be misidentified as severe depression, while long-term low mood might be missed due to calm expression on a particular day. Without temporal data, models struggle to capture depression' s fluctuation patterns and long-term course characteristics (Guntuku et al., 2017).

Third, ethical and privacy risks exist. Deep psychological profiling of social media data involves user privacy boundaries. Even public data analysis without explicit authorization may raise controversies (Benton et al., 2017). Additionally, if models harbor potential biases (e.g., heightened sensitivity to certain genders' emotional expressions), algorithmic discrimination may affect assessment fairness (Straw & Callison-Burch, 2020).

4.4 Future Directions

Based on current findings and limitations, future computational psychiatry research should seek breakthroughs in data construction, assessment dimensions, and clinical applications. First, the core task at the data level is constructing highly culturally-adapted specialized datasets. Since general large models often struggle to fully understand complex contexts in Chinese mental health domains (Chen et al., 2023), there is urgent need to establish large-scale Chinese mental health instruction fine-tuning datasets annotated by experts. Such datasets should focus on three content types: (1) subculture-specific samples targeting phenomena like "Sang culture," paired with "non-depressive" labels to

train models to accurately distinguish subcultural expression from pathological symptoms; (2) deep metaphor interpretation samples using chain-of-thought formats to guide models in understanding deep semantic mapping; (3) multi-turn dialogue data simulating real consultation scenarios to improve models' symptom assessment precision in continuous interaction.

Second, assessment frameworks should transform from single static text analysis to multimodal and longitudinal temporal tracking. Given depression' s comprehensive manifestations, reliance solely on single-post text slices easily leads to misjudgment. Future research should integrate visual signal analysis, using multimodal models to parse image tones, compositions, and environmental features (e.g., living environment disorderliness, potential self-harm traces) to capture psychological states difficult to express through text (Gui et al., 2019). Simultaneously, temporal analysis techniques must be introduced, using temporal attention mechanisms to track users' long-term posting trajectories (Sawhney et al., 2020). By detecting sudden changes in linguistic style, activity levels, and pronoun usage habits, models can more effectively capture pathological dynamic changes, accurately distinguishing temporary emotional fluctuations from persistent affective disorders and resolving state-trait confusion (Choudhury et al., 2013).

Finally, at the clinical application level, efforts should focus on constructing "human-machine collaborative" decision support systems. AI should serve as physicians' cognitive assistants rather than replacements (As' ad et al., 2025). Ideal systems should adopt hybrid intelligence architecture: front-end utilizes large language models'powerful semantic analysis to extract structured symptom clues from scattered social media data (e.g., "sleep disturbance: present," "anhedonia: absent"); middle-end combines traditional machine learning algorithms for stable risk classification based on extracted symptom features, effectively circumventing large models' "hallucination" risks (Aggarwal et al., 2025); back-end generates explainable diagnostic reports listing key evidence supporting judgments for physician verification (Zhang et al., 2025). Additionally, strict ethical intervention protocols must be established, mandating manual intervention and warning processes when high-risk suicide signals are detected, ensuring optimal balance between technical efficacy, cultural adaptability, and ethical safety.

5 Conclusion

This study confirms that Chinese-pretrained large language models like Qwen3 demonstrate significant validity advantages in social media depression text assessment, particularly in capturing indigenous symptom expressions and subcultural features. However, cultural context complexity and non-literal metaphorical meanings remain primary challenges for current technology. By integrating DASS-21 psychometric theory with advanced natural language processing technology, this study provides empirical foundations for constructing low-cost, high-coverage mental health monitoring systems through text analysis. Future research must further pursue breakthroughs in multimodal fusion, longitudinal

temporal analysis, and ethical safety alignment to achieve more precise digital mental health services.

Author Contributions

Tuo Xin: Participated in data collection and annotation, responsible for annotation guideline development, wrote Introduction and Methods sections.
Zhang Ji: Participated in data collection and annotation, responsible for large language model implementation, wrote Discussion section.
Zhou Jingwen: Participated in data collection and annotation, responsible for data analysis, wrote Results section.
Zhu Tingshao: Guided research design, formulated research questions, designed study framework.

References

Aggarwal, V., Thukral, S., Patel, K., & Chatterjee, A. (2025). Leveraging LLMs for mental health: Detection and recommendations from social discussions (No. arXiv:2503.01442). arXiv. https://doi.org/10.48550/arXiv.2503.01442

As' ad, M., Faran, N., & Joharji, H. (2025). AI-supported shared decision-making (AI-SDM): Conceptual framework. JMIR AI, 4, e75866. https://doi.org/10.2196/75866

Beck, A. T., Brown, G., Berchick, R. J., Stewart, B. L., & Steer, R. A. (1990). Relationship between hopelessness and ultimate suicide: A replication with psychiatric outpatients. American Journal of Psychiatry, 147(2), 190-195. https://doi.org/10.1176/ajp.147.2.190

Benton, A., Coppersmith, G., & Dredze, M. (2017). Ethical research protocols for social media health research. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (pp. 94-102). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1612

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ···Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649-685. https://doi.org/10.1017/S1351324916000383

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), 43-54. https://doi.org/10.1038/s41746-020-0233-7

Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q., & Xu, X. (2023). SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In H. Bouamor, J. Pino, & K. Bali (Eds.), Findings of the Association for Computational Linguistics:

EMNLP 2023 (pp. 1170–1183). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.83
Choudhury, M. D., 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
Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. Journal of Abnormal Psychology, 100(3), 316-336. https://doi.org/10.1037/0021-843X.100.3.316
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Coppersmith, G., Dredze, M., & Harman, C. (2015). Quantifying mental health signals in Twitter. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 51-60). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-1207
Deng, Y., Chen, Y., Zhang, B., & Wang, L. (2024). Prevalence of depression among Chinese university students: A systematic review and meta-analysis. Journal of Affective Disorders, 342, 45-56. https://doi.org/10.1016/j.jad.2024.01.045
Du, J. (2017). Sang culture: From learned helplessness to "self-irony." Editing Friend, 9, 109–112. https://doi.org/10.13786/j.cnki.cn14-1066/g2.2017.09.022
Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences, 115(44), 11203-11208. https://doi.org/10.1073/pnas.1802331115
Eisenberg, D., Downs, M. F., Golberstein, E., & Zivin, K. (2009). Stigma and help seeking for mental health among college students. Medical Care Research and Review, 66(5), 522-541. https://doi.org/10.1177/1077558709335173
Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33(3), 613-619. https://doi.org/10.1177/001316447303300309
Gong, X., Xie, X., Xu, R., & Luo, Y. (2010). Psychometric properties of the Chinese versions of DASS-21 in Chinese college students. Chinese Journal of Clinical Psychology, 18(4),
Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., & Chen, Z. (2019). Cooperative multimodal approach to depression detection in Twitter. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 110–117. https://doi.org/10.1609/aaai.v33i01.3301110
Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences, 18, 43-49. https://doi.org/10.1016/j.cobeha.2017.07.005
Guo, S., Jiang, S., He, Q., Xiao, Y., Liang, J., Yude, B., He, M., Tao, S., & Zhang, L. (2025). Do large language models truly understand cross-cultural differences? (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2512.07075
Guo, D., & Wang, Y. (2023). Exploring psychological motivations behind

"Sang culture" phenomenon. Public Relations World, 7, 78–79.

Harrigian, K., Aguirre, C., & Dredze, M. (2020). Do models of mental health based on social media data generalize? In T. Cohn, Y. He, & Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 3774-3788). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.337

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: A quantitative review of taxometric research. Psychological Medicine, 42(5), 903-920. https://doi.org/10.1017/S0033291711001966

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. British Journal of Clinical Psychology, 44(2), 227-239. https://doi.org/10.1348/014466505X29657

Hu, H., Zhou, Y., Si, J., Wang, Q., Zhang, H., Ren, F., Ma, F., Cui, L., & Tian, Q. (2025). Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling (No. arXiv:2505.15715). arXiv. https://doi.org/10.48550/arXiv.2505.15715

Huang, X., Zhang, L., Chiu, D., Liu, T., Li, X., & Zhu, T. (2020). Detecting suicidal ideation in Chinese microblogs with psychological lexicons. In 2014 IEEE 11th International Conference on Ubiquitous Intelligence and Computing (pp. 844-849). IEEE. https://doi.org/10.1109/UIC-ATC-ScalCom.2014.48

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. JAMA, 318(13), 1215-1216. https://doi.org/10.1001/jama.2017.11295

Ji, S., Li, X., Huang, Z., & Cambria, E. (2023). Suicidal ideation and mental disorder detection with attentive relation networks. Neural Computing and Applications, 35(13), 9309-9323. https://doi.org/10.1007/s00521-020-05589-4

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2021). Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems, 8(1), 214-226. https://doi.org/10.1109/TCSS.2020.3021467

Jing, Y., & Jiang, G. (2024). "No man is an island": How Chinese netizens use deliberate metaphors to provide "depression sufferers" with social support. Digital Health, 10, 20552076241228521. https://doi.org/10.1177/20552076241228521

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Li, T., Yang, S., Wu, J., Wei, J., Hu, L., Li, M., Wong, D. F., Oltmanns, J. R., & Wang, D. (2026). Can large language models identify implicit suicidal ideation? An empirical evaluation (No. arXiv:2502.17899). arXiv. https://doi.org/10.48550/arXiv.2502.17899

Li, W., Zhao, Z., Chen, D., Peng, Y., & Lu, Z. (2023). Prevalence and associated factors of depression and anxiety symptoms among college students: A systematic review and meta-analysis. Journal of Affective Disorders, 320, 596-606. https://doi.org/10.1016/j.jad.2022.09.156

Li, X., & Peng, Y. (2020). Symbolic performance: Critical discourse construction of Sang culture in cyberspace. International Press, 42(12), 50–67. https://doi.org/10.13495/j.cnki.cjjc.2020.12.003

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behaviour Research and Therapy, 33(3), 335-343. https://doi.org/10.1016/0005-7967(94)00075-U

Rude, S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. Cognition & Emotion, 18(8), 1121-1133. https://doi.org/10.1080/02699930441000030

Sawhney, R., Joshi, H., Gandhi, S., & Shah, R. R. (2020). A time-aware transformer based model for suicide ideation detection on social media. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7685–7697). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.619

Shen, Y., Zhang, W., Chan, B. S. M., Wu, Q., Meng, F., Liang, L., & Shi, L. (2018). Effectiveness of group cognitive behavioral therapy with student assistance program for Chinese college students with mental health problems: A randomized controlled trial. Behaviour Research and Therapy, 107, 133-140. https://doi.org/10.1016/j.brat.2018.06.003

Shu, B., Joshi, I., Karnaze, M., Pham, A. C., Kakkar, I., Kothe, S., Hovasapian, A., & ElSherief, M. (2025). Fluent but unfeeling: The emotional blind spots of language models (No. arXiv:2509.09593). arXiv. https://doi.org/10.48550/arXiv.2509.09593

Straw, I., & Callison-Burch, C. (2020). Artificial intelligence in mental health and the biases of language based models. PLOS ONE, 15(12), e0240376. https://doi.org/10.1371/journal.pone.0240376

Suler, J. (2004). The online disinhibition effect. Cyberpsychology & Behavior, 7(3), 321-326. https://doi.org/10.1089/1094931041291295

Sun, X., & Wang, H. (2019). The sang subculture and mental health among Chinese youth: A phenomenological analysis. Asian Journal of Social Psychology, 22(3), 289-297. https://doi.org/10.1111/ajsp.12367

Tong, X., Choenni, R., Lewis, M., & Shutova, E. (2024). Metaphor understanding challenge dataset for LLMs (No. arXiv:2403.11810). arXiv. https://doi.org/10.48550/arXiv.2403.11810

Torous, J., Staples, P., & Onnela, J. P. (2016). Realizing the potential of mobile mental health: New methods for new data in psychiatry. Current Psychiatry Reports, 18(6), 1-7. https://doi.org/10.1007/s11920-016-0704-y

Treadway, M. T., & Zald, D. H. (2011). Reconsidering anhedonia in depression: Lessons from translational neuroscience. Neuroscience & Biobehavioral Reviews, 35(3), 537-555. https://doi.org/10.1016/j.neubiorev.2010.06.006

Wang, K., Shi, H. S., Geng, F. L., Zou, L. Q., Tan, S. P., Wang, Y., Neumann, D. L., Shum, D. H. K., & Chan, R. C. K. (2016). Cross-cultural validation of the Depression Anxiety Stress Scale-21 in China. Psychological Assessment, 28(5), e88-e100. https://doi.org/10.1037/pas0000207

Watson, D., Clark, L. A., Weber, K., Assenheimer, J. S., Strauss, M. E., &

McCormick, R. A. (1995). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. Journal of Abnormal Psychology, 104(1), 15-25. https://doi.org/10.1037/0021-843X.104.1.15

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

World Health Organization. (2023). Depressive disorder (depression). https://www.who.int/news-room/fact-sheets/detail/depression

Yang, K., Ji, S., Zhang, T., Xie, Q., & Ananiadou, S. (2023). Towards interpretable mental health analysis with large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6056-6077). Association for Computational Linguistics. https

Note: Figure translations are in progress. See original paper for figures.