

RESEARCH ARTICLE

SentinelGPT: A Large Language Model for Detecting Suicide and Depression Tendencies among Adolescents Based on the Micro-blog's "Tree Hole" and GraphRAG

Shaojie Wang¹, Shaofu Lin^{1*}, Zhisheng Huang^{2,3,4}, Haoru Su¹, Chaohui Guo²

¹College of Computer Science, Beijing University of Technology, Beijing 100124, China

²Department of Computer Science, Vrije University Amsterdam, Amsterdam, Netherlands

³Clinical Research Center for Mental Disorders, Shanghai Pudong New Area Mental Health Center, Tongji University School of Medicine, Shanghai, China

⁴Deep Blue Technology Group, Shanghai, China

ARTICLE DATA

Article History

Received 25 October 2025

Revised 12 February 2026

Accepted 25 May 2026

Keywords

SentinelGPT

Micro-blog "Tree Hole"

Adolescent suicidal and depression tendencies

GraphRAG

LoRA fine-tuning

ABSTRACT

In recent years, suicide rates among Chinese adolescents have risen markedly. Moreover, the onset of suicidal behaviour has shifted toward increasingly younger age groups. As a result, expressions of suicidal ideation on social media have become an important public health concern. Although the deep integration of natural language processing and large language models offers new opportunities for adolescent mental-health detection, significant technical challenges remain in terms of model generalizability and detection accuracy. To address these challenges, we propose SentinelGPT, a large language model designed to detect suicidal and depressive tendencies in adolescents. First, we construct a Chinese social-media corpus for suicide and depression detection by performing multi-dimensional cleaning, screening, and preprocessing on original data collected from the Micro-blog "Tree Hole". Second, we adopt a collaborative GraphRAG architecture that combines a knowledge graph with a vector database to enhance the model's ability to integrate information across multiple dialogue turns and capture semantic relationships in complex contexts. Third, we fine-tune a dialogue-enhanced LLaMA 3.1 model using LoRA and deploy it privately via the Ollama framework to improve parsing of fragmented expressions, metaphorical language, and cross-platform contextual variations commonly found in online text. To further verify model effectiveness, we conduct comparative experiments against four representative baselines DeepSeek V3, Qwen 2.5, Fine-tuned Qwen 2.5 and ELECTRA using the same dataset and evaluation metrics including precision, recall, and F1-score. SentinelGPT achieves the highest overall F1-scores, outperforming all baselines and demonstrating superior robustness and generalization ability. The results demonstrate that SentinelGPT effectively detects suicidal and depressive tendencies in the complex context of Micro-blog comments, maintains semantic consistency across multi-turn conversations, and offers a reliable technical solution for adolescent mental-health monitoring and early warning.

1. INTRODUCTION

The adolescent mental-health crisis has become a global public-health challenge. According to the World Health Organization's 2024 annual report [1], suicide is the second leading cause of death among people aged 15–29. The Chinese National Mental Health Development Report (2019–2020) [2] published by the Institute of Psychology, Chinese Academy of Sciences, indicates that 27% of adolescents report 'often feel lonely'. In the digital era, social media are reshaping both emotional expression and help-seeking behaviours among young

people. The term 'Tree Hole', which originally described hollows in damaged trees where people might hide secrets and worries, has been extended to denote anonymous online spaces for sharing personal emotions, experiences and concerns. The Micro-blog 'Tree Hole' [3], as a representative anonymous communication space, provides a major channel through which adolescents vent emotions and psychological complaints; by circumventing real-world identity constraints and interpersonal pressures, teenagers can more freely express their inner feelings.

Micro-blog [4], one of China's largest mainstream social media platforms, reports 590 million monthly active users, nearly

*Corresponding author. Email: linshaofu@bjut.edu.cn

© 2026 The Authors. Published by Nexus Press B.V.

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

80% of whom are young people [5]. As such, it has become a natural outlet for adolescent emotional expression. However, as the user base expands, the volume of negative and suicidal posts in Micro-blog 'Tree Hole' has increased. A particularly concerning phenomenon is the frequent use of metaphorical language to express suicidal ideation. Phrases such as 'want to fall asleep forever' and 'eat sugar' obscure explicit intent, and the blurred boundary between figurative expressions and everyday language creates substantial challenges for early identification and timely intervention.

Traditional mental-health screening methods have long dominated the field, but practice and research reveal four major limitations. First, conventional scale-based screening tools such as the PHQ-9 exhibit high omission rates, which some studies report to exceed 56% because screening is not conducted frequently enough [6]. Second, there is a severe shortage of professional counsellors. Third, automated monitoring systems that rely on keyword matching produce high false-alarm rates. Fourth, questionnaire surveys and clinical interviews, despite their value, are subject to subjective biases from respondents and evaluators, and they are constrained in frequency and scope, making dynamic, comprehensive tracking of individual mental health difficult.

The application of natural language processing (NLP) to mental-health detection has therefore become a major research direction [7]. With the proliferation of social media and online health services, massive volumes of unstructured text now provide valuable resources for mental-health research. Previous studies have used NLP techniques such as sentiment analysis, semantic role labeling and topic modeling to extract emotional features, sources of stress and indicators of psychological problems. For example, in the task of depression detection, Chen et al. [8] proposed a hybrid deep-learning model that combines pretrained Sentence-BERT (SBERT) with a convolutional neural network (CNN) to analyse Reddit posts; when trained and evaluated on the Self-Reported Mental Health Diagnoses (SMHD) dataset, the model achieved an accuracy of 0.86.

Despite these advances, significant technical bottlenecks remain. First, the limited interpretability of deep-learning models constrains their clinical applicability. Transformer-based models [9], with complex attention mechanisms and large numbers of parameters, are difficult to interpret in terms of how they extract mental-health features from text. Another major issue is data scarcity and the high cost of annotation, which increases the risk of overfitting. Many mental-health datasets remain small. For instance, the widely used DAIC-WOZ depression dataset includes only 189 samples [10], which limits the model's ability to learn generalizable features. A further challenge is the semantic complexity of natural language. Metaphorical and indirect expressions reduce detection accuracy, because existing NLP systems perform much better on literal text than on figurative language.

Large language models (LLMs) [11] offer a promising avenue to address these challenges. Pretrained on vast corpora, LLMs acquire strong language understanding and generation capabilities and can leverage self-supervised learning to make efficient use of limited domain data. Their attention mechanisms also improve contextual semantic understanding, which is advantageous for processing complex language phenomena such as metaphors and puns. Nevertheless, when general LLMs are applied to the mental-health domain, notable deficiencies emerge. Such models typically lack specialised mental-health

knowledge, including terminology, diagnostic criteria and psychotherapeutic principles, and they may therefore produce inaccurate or unprofessional judgments in complex clinical scenarios.

Low-Rank Adaptation (LoRA) fine-tuning [12] can rapidly adapt pretrained models to new tasks by adding a small number of low-rank parameter matrices while leaving the original pretrained weights unchanged. LoRA helps avoid catastrophic forgetting during fine-tuning and can enhance a model's ability to understand and analyse conversational Chinese while preserving general language proficiency. The GraphRAG architecture [13] integrates a knowledge graph with a vector database in a retrieval-augmented generation framework, which substantially improves the model's ability to process complex context. A knowledge graph can store structured domain knowledge and provide explicit semantic relationships, enabling rapid access to related concepts; a vector database enables efficient similarity retrieval. Together, GraphRAG better addresses fuzzy semantics, metaphors and other complex online expressions typical of adolescent communications and improves the model's ability to capture cues related to suicidal ideation. The Ollama framework [14] provides a lightweight, extensible deployment environment that supports private deployment of the LLaMA 3.1 model [15], improving response latency while ensuring data privacy and security.

Built on these technologies, our SentinelGPT model demonstrates advantages over previous approaches. By combining GraphRAG with a LoRA-fine-tuned LLaMA 3.1 model, SentinelGPT more accurately interprets internet slang, emoticons and other nonstandard expressions, reducing detection errors arising from misinterpretation. Integrating professional psychological knowledge via the knowledge graph mitigates the lack of domain expertise in general LLMs, leading to more professional and accurate assessments of suicidal tendencies. Private deployment improves detection efficiency and preserves user privacy. In multi-turn conversational settings, the optimized architecture helps SentinelGPT maintain semantic consistency, track psychological changes over time and provide more reliable technical support for detecting adolescent suicide and depression tendencies.

The main contributions of this paper are as follows:

1. We propose SentinelGPT, a model specifically developed for detecting suicidal and depressive tendencies among adolescents on social media. Its main contribution lies in the domain-adapted integration of the GraphRAG architecture with LoRA-based fine-tuning. Experimental results indicate that SentinelGPT achieves superior overall detection performance compared with multiple general-purpose and fine-tuned baselines.
2. We develop a risk-aware corpus construction strategy that systematically organizes heterogeneous social-media expressions into fine-grained supervisory labels covering four levels of suicidal tendency and four levels of depressive tendency through schema-aligned annotation and step-wise filtering, thereby providing reliable structured signals for subsequent knowledge graph induction.
3. We propose an integrated GraphRAG architecture that combines knowledge graphs with vector retrieval. By establishing a synergistic mechanism between structured reasoning and semantic matching, our framework achieves interpretable understanding of complex linguistic

expressions in social media texts. This design is further operationalized within a psychological risk detection pipeline, thereby enhancing both the depth of semantic interpretation and the effectiveness of the detection process.

The remainder of this paper is organised as follows. [Section 2](#) describes the research methods. [Section 3](#) presents the dataset, experimental settings, case studies, ablation experiments, evaluation metrics, and discussions. [Section 4](#) concludes and outlines directions for future work.

2. METHODOLOGY

The SentinelGPT model is a large language model developed by fine-tuning LLaMA 3.1 with LoRA to enhance Chinese dialogue capabilities. Its social-media corpus for detecting suicidal and depressive tendencies was constructed from the ‘Tree Hole’ dataset on Micro-blog. The model further employs a GraphRAG architecture that integrates a knowledge graph with a vector database, enabling structured storage of domain knowledge and efficient semantic retrieval via vectorised representations. Private deployment ensures data security and user privacy. Following iterative integration and optimisation of these components, SentinelGPT can effectively detect adolescent suicidal and depressive tendencies and thereby facilitate timely psychological interventions to safeguard adolescent mental health. The SentinelGPT architecture is illustrated in [Figure 1](#).

2.1. Base Language Model and Fine-tuning with LoRA

LLaMA 3.1, a next-generation natural language processing model developed by Meta, supports context lengths of up to 128,000 tokens and demonstrates clear advantages in understanding complex contexts as well as in accurately capturing details and logical relationships. In multi-turn dialogue scenarios such as those encountered in intelligent psychological counselling services, LLaMA 3.1 is capable of retaining key information, recognizing changes in users’ emotions and intentions, and generating coherent and contextually appropriate responses.

In this study, we employ Low-Rank Adaptation (LoRA) as an efficient fine-tuning paradigm to tailor a pre-trained language model to the task of suicidal and depressive tendency detection. Rather than advancing LoRA as a methodological innovation, our approach situates it as a lightweight parameter-efficient mechanism that supports the core objective of task alignment. LoRA is employed to reduce computational cost and stabilize training, allowing us to concentrate on designing effective fine-tuning data and evaluation mechanisms that are consistent with the target task.

2.1.1. Task-oriented Fine-tuning Dataset Design

This study constructs a task-oriented fine-tuning dataset explicitly tailored to the detection of suicidal and depressive tendencies in adolescent social media discourse. Approximately 81,000 Chinese dialogue samples are curated from social media platforms, online forums, and professional psychological counselling corpora. The dataset is designed to reflect the linguistic and psychological characteristics of high-risk mental health expressions, guided by task-specific design considerations.

Suicidal and depressive tendencies are rarely manifested through isolated utterances, but instead emerge through the accumulation of negative self-perception, emotional exhaustion, and hopelessness across multiple conversational turns. To capture this phenomenon, dialogue samples are selected and segmented in a manner that preserves longitudinal emotional trajectories, enabling the model to differentiate sustained psychological risk from transient negative emotional states. High-risk psychological conditions in adolescent social media are frequently conveyed through metaphorical, abstract, and indirect language rather than explicit self-harm statements. Accordingly, the dataset deliberately incorporates non-literal expressions such as disappearance wishes, self-worth negation, and chronic fatigue metaphors.

Through these design considerations, the constructed dataset establishes a direct alignment between the fine-tuning process and downstream suicidal and depressive tendency detection tasks, providing a task-consistent foundation for subsequent model training and evaluation.

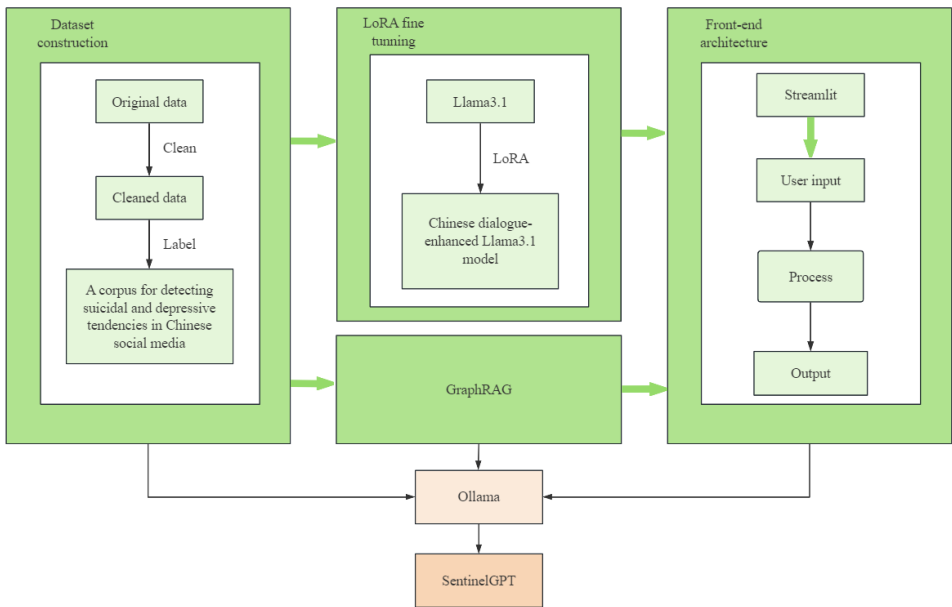


Figure 1 | SentinelGPT model frame diagram.

2.1.2. Fine-tuning Configuration and Training Framework

Fine-tuning is conducted on the LLaMA-Factory platform [16], using LLaMA 3.1-8B as the base model. The fine-tuning process is illustrated in Figure 2. Key hyperparameters, including learning rate and batch size, are adjusted through controlled experiments to balance convergence stability and task performance. LoRA adapters are inserted into attention and feed-forward layers, while the majority of base model parameters remain frozen. This configuration enables efficient adaptation while minimizing overfitting to domain-specific expressions.

Importantly, the effectiveness of fine-tuning is evaluated not only by general dialogue quality, but also by task-relevant indicators, including the model’s ability to correctly interpret emotionally ambiguous content and maintain consistent risk awareness across multi-turn interactions. As shown in Table 1, the fine-tuned model demonstrates clear improvements in Chinese dialogue understanding and task-specific response quality. The fine-tuned model is integrated into the Ollama framework to support lightweight deployment, efficient inference, and continuous iteration.

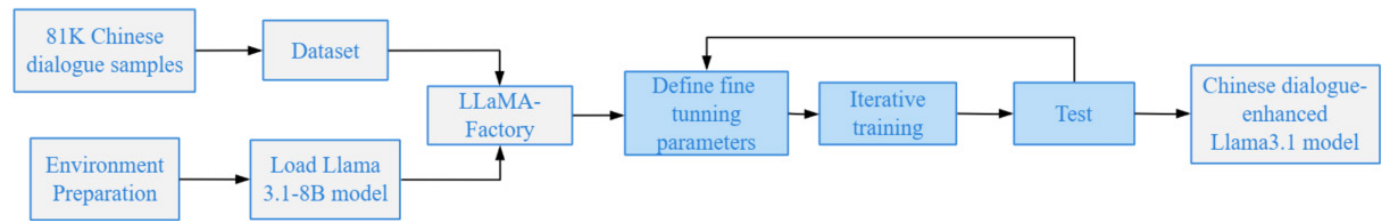


Figure 2 | LoRA fine tuning.

Table 1 | Comparison of fine-tuning results

Model	Predict bleu-4	Predict rouge-1	Predict rouge-2	Predict rouge-l	Predict runtime
Llama 3.1	1.33	16.70	4.57	5.27	3830.85
LoRA fine-tuned Llama 3.1	13.20	32.45	13.07	25.11	1622.78

2.2. Graph Retrieval-Augmented Generation Architecture

Figure 3 presents the overall architecture of the GraphRAG system, which integrates a knowledge graph, a vector data-base, and a retrieval and generation mechanism to enhance the semantic grounding and interpretability of mental health detection. The architecture operates by first constructing a structured knowledge graph from large-scale social media data, then embedding and clustering posts into a vector space, and finally using both sources of information to augment LLM based reasoning and response generation. Together, these components enable the model to perform context aware retrieval and produce outputs.

2.2.1. Data Sources and Construction of the Knowledge Graph

The knowledge graph used in the GraphRAG framework is constructed primarily from large-scale social media data collected from the Micro-blog ‘Tree Hole’ platform. The annotated corpus contains over 200,000 user-generated messages and serves as

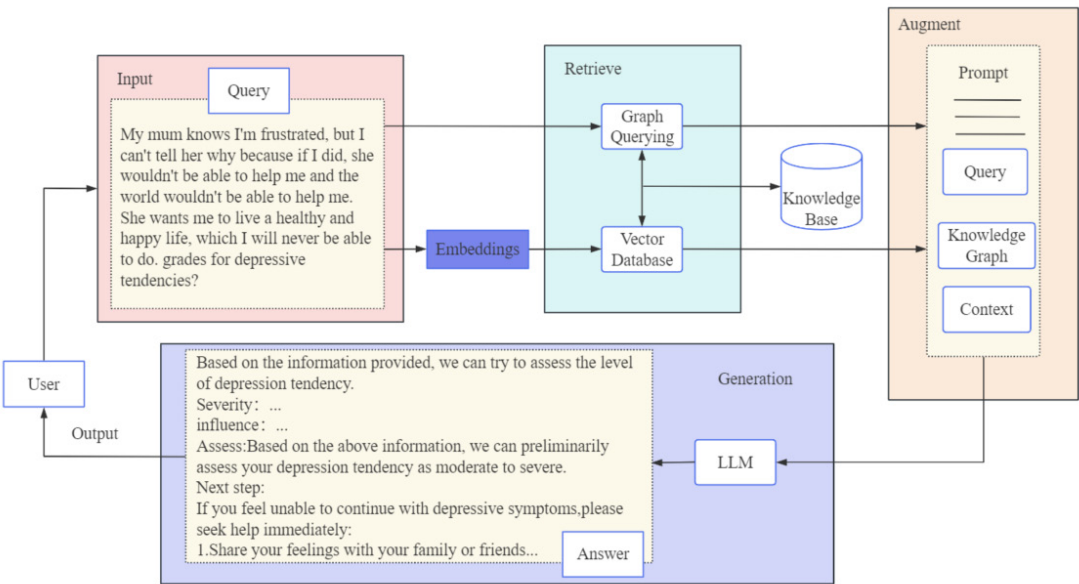


Figure 3 | GraphRAG architecture.

the sole data source for entity extraction and relationship induction, rather than relying on clinical manuals as direct inputs.

To ensure validity while avoiding purely data-driven bias, authoritative psychiatric knowledge including DSM-5 [17] diagnostic criteria and standardized assessment scales such as BDI [18], CES-D [19], SSI [20], is incorporated at the schema and concept normalization level. Specifically, social media texts provide candidate entities and linguistic expressions, such as emotions, symptoms, risk factors, and metaphorical phrases, whereas DSM-5 functions as a semantic anchor that constrains entity types and permissible relationships.

The knowledge graph is automatically constructed through a large-language-model-assisted pipeline that performs semantic parsing and co-occurrence analysis to extract mental-health-related entities and infer their relationships. Importantly, DSM-5 and assessment scales are not directly ingested into the GraphRAG pipeline, nor are they used during retrieval or generation at runtime; their role is limited to offline schema design and post-hoc concept alignment.

2.2.2. Sentence-to-Entity Semantic Grounding with Schema Constraints

A central challenge in applying retrieval-augmented generation to social media mental health data lies in mapping informal, implicit, and metaphorical expressions to structured clinical concepts. To address this challenge and reduce noise introduced by automatic extraction, sentence-to-entity alignment is achieved through an LLM-assisted, schema-constrained semantic grounding process rather than direct keyword matching.

Prior to graph construction, a predefined knowledge graph schema is specified, which includes a fixed set of entity types such as emotion, symptom, risk factor, and intervention, as well as relation types such as indicates, co-occurs with, increases risk of, and alleviated by. This schema is informed by authoritative psychiatric knowledge and defines clinically plausible concepts and relations.

Given an input sentence, the language model first performs contextual semantic parsing to identify candidate psychological expressions. These candidates are then mapped to normalized graph entities by considering semantic similarity and entity-type compatibility under the predefined schema. Only entity-relation pairs that satisfy schema constraints are retained, while incompatible or ill-formed outputs are filtered out.

To further mitigate spurious matches, grounded entities are validated against local graph structures to ensure consistency with existing relational patterns. Through this schema-constrained grounding mechanism, diverse social media expressions are systematically linked to structured mental health concepts, providing a reliable foundation for subsequent graph-based retrieval and reasoning.

2.2.3. Vector Database Integration and Latent Theme Modeling

In parallel with graph construction, each preprocessed post is encoded into a 1,024-dimensional vector using the text2vec-base-chinese sentence-transformer model. These embeddings are indexed in a FAISS-based vector database to enable efficient similarity search.

For each query, the system retrieves semantically similar posts and constructs a similarity graph, which is further processed using the Leiden community detection algorithm to identify coherent semantic clusters. These clusters correspond to latent psychological themes, such as self-blame, hopelessness, or family conflict. Each cluster is subsequently summarized by the language model into a concise semantic representation, providing high-level contextual information that complements explicit graph relations.

2.3. Construction of SentinelGPT Model

The objective of SentinelGPT is to assess suicidal and depressive tendencies among adolescents and to assist psychological intervention volunteers in risk identification and response prioritization. This is achieved by integrating the LoRA-fine-tuned language model with the GraphRAG architecture into a unified system. During inference, when an input utterance is received, SentinelGPT initiates a dual-path retrieval process. The vector-based path retrieves semantically similar social media posts and latent thematic summaries, while the graph-based path queries the knowledge graph to obtain structured entities and their relational context relevant to the grounded expressions. The retrieved information from both paths is jointly provided as augmented context to the fine-tuned LLaMA 3.1 model, enabling generation that is both contextually informed and clinically grounded.

Based on the generated analysis, SentinelGPT categorizes suicidal and depressive tendencies into four risk levels: none, low, medium, and high. This categorical assessment supports rapid risk screening and response prioritization by intervention volunteers. Beyond static classification, SentinelGPT models the temporal evolution of psychological states across multi-turn conversations. By tracking changes in emotional vocabulary, discourse patterns, and references to daily experiences, the system identifies trends indicating deterioration, improvement, or stability in mental condition. For instance, a progression from mild negative emotions to repeated expressions of hopelessness and helplessness is interpreted as an escalation in psychological risk. An overview of the integrated SentinelGPT system is shown in Figure 4.

While the overall framework follows the general GraphRAG paradigm, this work introduces task-specific adaptations for adolescent mental health detection, including graph induction from noisy social media data, schema-guided clinical grounding without using DSM-5 as runtime input, and the integration of vector clustering with graph-based retrieval. These adaptations enable GraphRAG to operate effectively in high-uncertainty and linguistically implicit mental health detection scenarios.

3. EXPERIMENTS

3.1. Dataset Source and Introduction

The data used in this study are primarily sourced from the ‘Tree Hole’ section of the Micro-blog platform. The specific tree hole under investigation refers to the comment section associated with the suicide note posted by the Micro-blog user known by the nickname ‘Zoufan’. At present, this tree hole contains more than two million messages, with a follower base of approximately 620,000 and over 5.53 million combined reposts, comments, and likes.

SentinelGPT

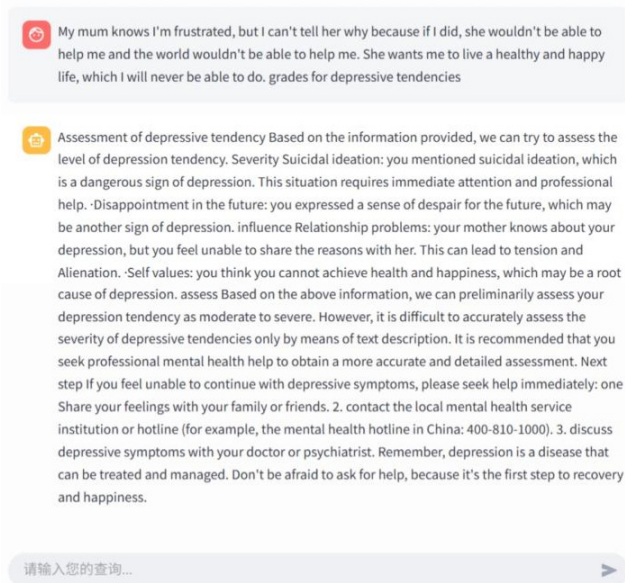


Figure 4 | SentinelGPT Q&A display.

Between October 2012 and February 2014, the 'Zoufan' tree hole experienced a relatively stable and active period of engagement, attracting a large number of users who expressed and exchanged their emotions. During this period, we selected posts related to psychological states, emotional distress, and life pressures. The messages concerning these themes are rich in content and highly representative, offering a comprehensive reflection of adolescents' psychological conditions, emotional struggles, and life stressors at that time. These factors are closely linked to adolescent psychological issues such as suicidal and depressive tendencies and thus hold substantial reference value for constructing a corpus dedicated to the detection of suicidal and depressive tendencies.

In accordance with strict screening criteria, duplicate, invalid, and irrelevant messages were removed. Ultimately, more than 330,000 original messages were retained. Each message contains six attributes: timestamp, commenter ID, commenter name, comment content including emoticons and other symbols, the user being replied to, and the hyperlink to that user's page.

To construct a high-quality Chinese social media corpus for suicidal and depressive tendency detection, the research team conducted multi-dimensional data cleaning and annotation of the raw dataset. Initially, natural language processing techniques were applied to remove special characters, advertisement links, invalid emoticons, and other extraneous information. Subsequently, sentiment analysis and keyword extraction algorithms were employed to identify texts containing potential indicators of suicidal or depressive tendencies, resulting in a base corpus comprising more than 200,000 messages.

On this foundation, and drawing upon the core assessment dimensions of authoritative domestic and international scales for evaluating suicidal tendencies such as the Suicidal Ideation Inventory (SSI), depression assessment instruments such as the Beck Depression Inventory (BDI) and the Centre for Epidemiological Studies Depression Scale (CES-D), as well as the Diagnostic and Statistical Manual of Mental Disorders, Fifth

Edition (DSM-5), an annotation guideline was developed to align with the linguistic and expressive characteristics of social media texts. Following this guideline, each message was annotated with both a suicidal tendency level and a depression level.

The classification of suicidal and depressive tendencies is defined as follows:

1. Suicidal Tendency Annotation

- 1) No Suicidal Tendency: The text expresses only transient or situationally induced negative emotions, for example, 'I am a bit sad because I was criticised by the teacher today.' There are no indications of suicidal thoughts, intentions, or behaviours.
- 2) Low Risk of Suicidal Tendency: This level includes expressions of persistent negative emotions such as 'I live in great pain every day', vague self-denial such as 'I seem to be unable to do anything well', or occasional questioning of the value of life. However, there are no explicit references to suicidal ideation or self-harming behaviour.
- 3) Medium Risk of Suicidal Tendency: Texts at this level contain explicit expressions of suicidal ideation, such as 'There is no meaning in living; it would be better to die', or descriptions of minor self-harming behaviour such as 'I couldn't help but cut myself again today'. Nevertheless, there is no evidence of a concrete or planned suicide attempt.
- 4) High Risk of Suicidal Tendency: This level is characterised by the clear articulation of a suicide plan such as 'I am going to take sleeping pills this weekend', references to previous suicide attempts, or strong expressions of the will to die such as 'I have already chosen the way to leave'.

2. Depression Level Annotation

- 1) No Depressive Tendency: The text conveys predominantly positive emotions or only mild and occasional negative feelings, without displaying any characteristic symptoms of depression.
- 2) Mild Depressive Tendency: The individual exhibits some negative emotions, such as sadness or disappointment, but these are short-lived. Mild depressive symptoms may appear, such as occasional insomnia or minor changes in appetite, though the severity remains below the diagnostic threshold for a depressive episode.
- 3) Moderate Depressive Tendency: Negative emotions occur frequently and persistently, including prolonged low mood and loss of interest. Noticeable symptoms such as sleep disturbances including insomnia or hypersomnia, significant appetite fluctuations including either increase or decrease, reduced energy, and difficulty concentrating are evident, but their severity and duration do not yet meet the criteria for severe depression.
- 4) Severe Depressive Tendency: Texts at this level reveal intense negative emotions, such as extreme despair and helplessness. The individual may present serious sleep disturbances, marked weight changes, psychomotor agitation or retardation, and recurrent thoughts

of death or suicide. These symptoms cause significant impairment in daily functioning and social interaction.

The labelling process adopted a two-person independent annotation mechanism, in which two annotators independently determined the categories of suicidal ideation and depression level for each text. In cases where the annotation results were inconsistent, a third annotator intervened to adjudicate and ensure consensus. The annotation framework was continuously refined through intra-group case discussions to maintain consistency and accuracy. Consequently, a corpus, as shown in Table 2, was established, providing clear annotations of suicidal risk levels and degrees of depression.

The dataset was strictly partitioned in a 7:3 ratio. Specifically, 70% of the data were utilised for constructing the knowledge graph within the GraphRAG architecture, training the vector database, and fine-tuning the model. The remaining 30% constituted an independent test set, employed to evaluate the generalisation capability and accuracy of SentinelGPT in detecting suicidal and depressive tendencies among adolescents.

3.2. Experimental Setup

The experiment was conducted on the Ubuntu 22.04 operating system, employing PyTorch 2.3.1 and other deep learning libraries. Using Python, data were retrieved from the comment section of the Micro-blog user ‘Zoufan’, yielding a substantial amount of textual information relevant to this study. The GraphRAG architecture integrates knowledge graphs and vector databases to construct a semantically enriched knowledge system. On this basis, the SentinelGPT model was designed and implemented within the Ollama framework, utilising the Llama3.1 large language model fine-tuned with the LoRA technique, and visualised through Streamlit.

For benchmarking, SentinelGPT was compared against four representative baselines: DeepSeek V3 [21], Qwen 2.5 [22], a fine-tuned variant of Qwen 2.5 and ELECTRA [23]. DeepSeek V3 has recently demonstrated strong zero-shot and few-shot capabilities in depression detection and severity assessment across clinical and multilingual datasets, making it a relevant large language model (LLM) baseline for depressive-tendency analysis. Qwen 2.5 represents a Chinese-optimized LLM that has been evaluated in mood and crisis detection tasks, showing strong suitability for mental-health-related conversational applications. Fine-tuned Qwen 2.5 was adapted using LoRA on

70% of the annotated dataset, following the same train–test split as SentinelGPT, and without GraphRAG integration or structured knowledge enhancement. ELECTRA, a widely adopted Transformer-based pretraining framework, is frequently employed as a robust baseline for social media mental health classification, including suicidal ideation detection. All models were evaluated on the same dataset and assessed using identical metrics such as precision, recall, and F1-score to ensure fair and consistent comparison.

3.3. Results

3.3.1. Ablation Experiment

This study conducts ablation experiments to further verify the performance of the SentinelGPT model in the absence of key components. The GraphRAG architecture plays an essential role in enhancing the model’s capability to detect suicidal and depressive tendencies among adolescents. Together, GraphRAG and the LoRA fine-tuning technique improve the model’s semantic understanding, information retrieval, and task adaptation performance.

According to the experimental results presented in Table 3, the model without these key components primarily generates structured plans for improving mental health, offering general strategies and suggestions but failing to evaluate depressive tendencies. In contrast, the complete SentinelGPT model is capable of assessing depressive tendencies by considering warning signs such as suicidal ideation and hopelessness about the future, and by analysing their effects on interpersonal relationships and self-worth. Based on these factors, it preliminarily categorises depressive tendencies as moderate to severe. Furthermore, the model provides practical recommendations for subsequent actions, emphasising that depression is both treatable and manageable, while encouraging individuals to seek professional help.

Whereas the model lacking key components merely proposes improvement measures, the complete SentinelGPT model not only performs evaluative analysis but also offers tailored recommendations and a more comprehensive understanding of depression. The incorporation of the GraphRAG and LoRA fine-tuning mechanisms significantly enhances the model’s analytical depth and diagnostic precision. These mechanisms enable the model to make more accurate judgements regarding suicidal and depressive tendencies, to produce more

Table 2 Presentation of corpus data			
ID	Comment (including emoticons)	Suicidal tendency annotation	Depression level label
1	My mum knows I'm frustrated, but I can't tell her why because if I did, she wouldn't be able to help me and the world wouldn't be able to help me. She wants me to live a healthy and happy life, which I will never be able to do.	Low risk of suicidal tendencies	Moderate depressive tendency
2	If I feel uncomfortable in my heart, I will come to see you. It's not that I'm doing well, I'm just pretending to be doing well. I'm also unhappy in my heart, and I'll never be happy. I'm a stupid and slow reacting person, with endless insecurities.	No suicidal tendency	Moderate depressive tendency
3	My biggest hobby is sleeping, but God has deprived me of it, leaving me extremely exhausted and unable to fall asleep, struggling to turn over and over.	Low risk of suicidal tendencies	Severe depressive tendency

Table 3 | Ablation experiment results

User statement	Model	Depressive tendency analysis	Interpretable reasoning trace	Intervention suggestions
	Llama 3.1	Common Depressive Symptoms: 1. Persistent feelings of sadness or hopelessness; 2. Loss of interest in activities once enjoyed; 3. Changes in appetite or sleep patterns ...	None	None provided
My mum knows I'm frustrated, but I can't tell her why because if I did, she wouldn't be able to help me and the world wouldn't be able to help me. She wants me to live a healthy and happy life, which I will never be able to do.	Llama 3.1 + LoRA fine-tuning	None provided	None	Focusing on your mental health can be challenging but is an important journey towards improvement. Here's a structured plan based on your reflections: 1. Self-Care Practices:- Ensure adequate ...
	SentinelGPT	Based on the information provided, we can try to assess the level of depression tendency. Severity: ... Influence: ... Assess: Based on the above information, we can preliminarily assess your depression tendency as moderate to severe.	1. Phrase Analysis: helplessness ("wouldn't be able to help me"), and hopelessness ("will never be able to"). 2. Knowledge Graph Query: KG validation established 'Helplessness' & 'Hopelessness' as concepts associated with a 'Moderate-Severe Depressive Episode' (DSM-5). 3. Vector DB Retrieval: Retrieved posts with analogous expressions of incommunicable distress & pessimism, previously annotated as Moderate depressive tendency.	Next step: If you feel unable to continue with depressive symptoms, please seek help immediately: 1. Share your feelings with your family or friends. 2. Contact the local mental health service institution or hotline.

reasonable analyses of users' emotional trajectories, and to deliver intervention strategies that are both targeted and actionable. This evidence further underscores the crucial importance of these components in improving the overall performance of the SentinelGPT model and provides a solid foundation for its future optimisation and refinement.

This study conducts ablation experiments to further verify the performance of the SentinelGPT model in the absence of key components. The GraphRAG architecture plays an essential role in enhancing the model's capability to detect suicidal and depressive tendencies among adolescents. Together, GraphRAG and the LoRA fine-tuning technique improve the model's semantic understanding, information retrieval, and task adaptation performance.

According to the experimental results presented in Table 3, the model without these key components primarily generates structured plans for improving mental health, offering general strategies and suggestions but failing to evaluate depressive tendencies. In contrast, the complete SentinelGPT model is capable of assessing depressive tendencies by considering warning signs such as suicidal ideation and hopelessness about the future, and by analysing their effects on interpersonal relationships and self-worth. Critically, as shown in the 'Interpretable Reasoning Trace' column, the model's decision-making is transparently grounded in a multi-step process: it identifies key risk phrases from the user's input, validates the inferred clinical concepts such as Helplessness and Hopelessness against the DSM-5-aligned knowledge graph, and retrieves corroborating evidence from the annotated corpus. Based on these factors, it preliminarily categorises depressive tendencies as moderate to severe. Furthermore, the model provides practical recommendations for subsequent actions, emphasising that depression is both treatable and manageable, while encouraging individuals to seek professional help.

Whereas the model lacking key components merely proposes improvement measures, the complete SentinelGPT model not only performs evaluative analysis but also offers tailored recommendations and a more comprehensive understanding of depression. The incorporation of the GraphRAG and LoRA fine-tuning mechanisms significantly enhances the model's analytical depth and diagnostic precision. These mechanisms enable the model to make more accurate judgements regarding suicidal and depressive tendencies, to produce more reasonable analyses of users' emotional trajectories, and to deliver intervention strategies that are both targeted and actionable. This evidence further underscores the crucial importance of these components in improving the overall performance of the SentinelGPT model and provides a solid foundation for its future optimisation and refinement.

3.3.2. Comparative Experiment

In the task of detecting suicidal and depressive tendencies among adolescents, the functional performance of the SentinelGPT model relies on the integrity of its architectural components. Ablation experiments have demonstrated that when key components such as the GraphRAG architecture, the LoRA fine-tuning module, or specific annotated datasets are removed, the model is unable to effectively assess the levels of suicidal and depressive tendencies. Consequently, the evaluation of performance metrics, including precision, recall, and the F1 score, is conducted only for the fully configured SentinelGPT

model. The corresponding definitions of these indicators are provided in Equations (1)–(3).

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{FN} + \text{TP}} \quad (2)$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Among them, TP denotes the number of samples in which the actual category corresponds to a specific risk level such as medium risk and the model correctly predicts that same level. FP represents the number of samples that do not actually belong to a given risk category but are incorrectly classified by the model as belonging to it, while FN refers to the number of samples that genuinely belong to a particular risk category but are incorrectly predicted as other categories.

A two-dimensional evaluation framework was established, encompassing suicidal tendency with four risk levels and depression severity with four levels. A random partitioning strategy was employed to divide the dataset into training and test subsets in a 7:3 ratio, ensuring the generalisability of the evaluation results. The test subset was subsequently used to assess the performance of five models: SentinelGPT, DeepSeek V3, Qwen 2.5, Fine-tuned Qwen 2.5 and ELECTRA.

The comparative results, summarised in Table 4, report the average precision, recall, and F1-scores for each model across both evaluation dimensions. Among them, SentinelGPT achieves the best overall performance, with an average F1-score of 0.7114 for suicidal-tendency detection and 0.7541 for depression-tendency detection. Fine-tuned Qwen 2.5 ranks second, with DeepSeek V3 following closely in third place. Fine-tuned Qwen 2.5 outperforms the base Qwen 2.5 but still lags behind SentinelGPT due to the absence of GraphRAG integration. The base Qwen 2.5 delivers a moderate performance, whereas ELECTRA yields the weakest results—particularly in identifying higher-risk categories. This demonstrates that while fine-tuning enhances model performance, the knowledge graph (KG)-enhanced GraphRAG architecture provides additional gains in semantic understanding and detection accuracy for mental health risk assessment.

Overall, SentinelGPT demonstrates the most balanced and robust predictive capability across both dimensions, outperforming the other models in nearly all indicators. The performance gap is especially evident in suicidal-tendency classification, where the F1-score difference between SentinelGPT and ELECTRA exceeds 0.23. This superiority verifies that the integration of the GraphRAG architecture, LoRA fine-tuning, and the annotated corpus enhances the model’s generalisation ability in mental-health text analysis. These findings confirm that large language models equipped with domain-specific knowledge and adaptive fine-tuning strategies can provide technical support for early psychological-risk detection and adolescent mental-health monitoring.

3.4. Discussion

The experimental evaluation demonstrates that SentinelGPT exhibits strong detection capability in complex social-media contexts and consistently outperforms the baseline models DeepSeek V3, Qwen 2.5, Fine-tuned Qwen 2.5 and ELECTRA across both suicidal-tendency and depressive-tendency detection tasks. Compared with these models, SentinelGPT achieves higher precision–recall balance and superior F1-scores, indicating stronger generalisation and robustness in identifying psychological-risk features. Its performance advantage is particularly evident in suicidal-tendency classification, where it maintains stable accuracy even across subtle risk-level distinctions.

In addition, ablation studies confirm that the synergistic integration of the GraphRAG architecture and LoRA fine-tuning enables SentinelGPT to maintain semantic coherence throughout multi-turn dialogue interactions. The model effectively distinguishes between normal emotional expressions and varying levels of psychological risk, while its retrieval–generation mechanism renders the detection process interpretable and traceable.

SentinelGPT addresses several major bottlenecks inherent in conventional mental-health assessment systems. First, it alleviates data-sparsity issues by enhancing its knowledge base through over 200,000 annotated corpus entries and the GraphRAG framework. Second, by combining knowledge-graph-based structured reasoning with vector-database-driven

Table 4 | Comparative performance of five models on two-dimensional mental-health risk detection

Model	Dimension	Avg. precision	Avg. recall	Avg. F1
Knowledge-enhanced task-aligned models				
SentinelGPT	Suicidal tendency	0.6902	0.7545	0.7114
	Depression tendency	0.7538	0.7585	0.7541
General pre-trained LLMs				
DeepSeek V3	Suicidal tendency	0.6018	0.6827	0.6279
	Depression tendency	0.6833	0.6814	0.6808
Qwen2.5	Suicidal tendency	0.5403	0.6209	0.5656
	Depression tendency	0.6312	0.6235	0.6261
ELECTRA	Suicidal tendency	0.4511	0.5268	0.4729
	Depression tendency	0.5321	0.5464	0.5373
Task-fine-tuned LLMs (without KG)				
Fine-tuned Qwen 2.5	Suicidal tendency	0.6156	0.6927	0.6414
	Depression tendency	0.7065	0.6897	0.6968

semantic retrieval, the model improves its ability to interpret metaphorical and fragmented expressions, thereby reducing semantic ambiguity. Third, its private deployment via the Ollama framework ensures localised processing of sensitive adolescent data, guaranteeing data security and user privacy.

4. CONCLUSION

Focusing on the critical public-health issue of detecting suicidal ideation among adolescents, this study proposes SentinelGPT, a large language model designed to assist professionals in monitoring and assessing mental-health risks. The research centres on three core components: constructing a professional and domain-specific corpus, designing the GraphRAG architecture, and fine-tuning and deploying the model using LoRA within the Ollama framework.

Experimental results demonstrate that SentinelGPT outperforms existing baseline models across multiple detection tasks, exhibiting strong overall performance and robustness. The model not only effectively mitigates issues such as data sparsity, semantic ambiguity, and privacy concerns inherent in traditional mental health assessments, but also enhances interpretability through its structured retrieval-generation mechanism.

Despite these promising outcomes, several areas warrant further exploration in future research:

- 1) Expanding knowledge sources on adolescent suicidal ideation by incorporating professional research findings and clinical case data;
- 2) Investigating more efficient fine-tuning strategies such as QLoRA to enhance adaptability while reducing resource consumption;
- 3) Integrating multimodal data including speech and visual cues to better capture emotional tone and behavioural signals, thereby improving detection accuracy and reliability.

It must be emphasized that SentinelGPT is designed as an assistive tool, and its outputs must be reviewed by qualified mental health professionals before informing any intervention decisions. To protect user privacy, all data in this study were anonymised and obtained from publicly available sources. Future work will focus on improving confidence calibration and fairness evaluation to mitigate these issues. By incorporating these ethical safeguards, SentinelGPT aims to contribute positively to adolescent mental-health detection while upholding the standards of privacy, fairness, and accountability.

STATEMENTS AND DECLARATION

Conflict of Interest

The authors declare that they have no conflicts of interest.

Authors' Contribution

SW, ZH, and CG contributed to data collection and resource preparation. SW led the study conception and design, performed data analysis, developed the proposed SentinelGPT framework, conducted the experiments, and drafted the original manuscript. ZH assisted with data processing, validation,

and manuscript revision. CG contributed to methodology refinement and technical review. SL provided project administration, funding acquisition, supervision, and critical revisions of the manuscript. HS provided study guidance and supervision and reviewed the final manuscript. All authors reviewed the manuscript and approved the final version for publication.

Acknowledgments

The authors would like to thank all reviewers and editors for their valuable comments and suggestions, which helped improve the quality of this manuscript.

Funding

This study was funded by the Beijing Natural Science Foundation (Grant No. L222048).

REFERENCES

- [1] World Health Organization (WHO), World health statistics 2024: monitoring health for the SDGs, sustainable development goals, WHO, Geneva, Switzerland, 2024. <https://www.who.int/publications/i/item/9789240094703>.
- [2] Institute of Psychology, Chinese Academy of Sciences, Report on the Development of National Mental Health in China (2019–2020), Social Science Literature Press, Beijing, China, 2021.
- [3] X. Jing, S. Lin, Z. Huang, Extraction of characteristics of time in "Tree Hole" data. *J. Artif. Intell. Med. Sci.* 1 (2021), 43–48.
- [4] Micro-blog, 2024 Micro-blog user consumption trend report, (2024-12-17) [2025-04-29]. <https://finance.sina.com.cn/roll/2024-12-19/doc-inczyxct6793995.shtml>.
- [5] Micro-blog, Q4 2024 financial results, (2025-03-13) [2025-04-29]. <https://weibo.com/ttarticle/p/show?id=2309405143789030473936>.
- [6] B. Levis, A. Benedetti, B.D. Thombs, Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis, *BMJ.* 365 (2019), l1476.
- [7] B. Zhou, G. Yang, Z. Shi, S. Ma, Natural language processing for smart healthcare, *IEEE Rev. Biomed. Eng.* 17 (2024), 4–18.
- [8] Z. Chen, R. Yang, S. Fu, N. Zong, H. Liu, M. Huang, Detecting Reddit users with depression using a hybrid neural network SBERT-CNN, *arXiv preprint, arXiv:2302.02759*, 2023.
- [9] J.C. Hu, H. Chen, K.J. Chen, Y. Gao, J. Ye, X. Wang, et al., Transformer doctor: diagnosing and treating vision transformers, In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, Vancouver, BC, Canada, 2024.
- [10] K. Ellis, V. Rozgić, DAIC-WOZ: dataset for affective interpretation in conversational speech - WOZ scenario, Carnegie Mellon University. Retrieved May 4, 2025, <https://dcapswoz.ict.usc.edu/>
- [11] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, et al., 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*, ACM, New York, NY, United States, 8 (2024), 1–32.
- [12] B. Fan, X. Su, S. Tarkoma, P. Hui, HeLoRA: LoRA-heterogeneous federated fine-tuning for foundation models, *ACM Trans. Internet Technol.* 25 (2025), 1–22.

- [13] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, et al., From local to global: a graph RAG approach to query-focused summarization, arXiv preprint, arXiv:2404.16130, 2024.
- [14] Q. Mao, Y. Luo, Q. Zhang, Y. Luo, Z. Cao, J. Zhang, et al., XRAG: eXamining the core—benchmarking foundational components in advanced retrieval-augmented generation, arXiv preprint, arXiv:2412.15529, 2024.
- [15] Y. Lu, W. Zhao, X. Zhou, C. An, C. Wang, S. Li, et al., Effective length extrapolation via dimension-wise positional embeddings manipulation, arXiv preprint, arXiv:2504.18857, 2025.
- [16] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, et al., LlamaFactory: unified efficient fine-tuning of 100+ language models, arXiv preprint, arXiv:2403.13372, 2024.
- [17] D.J. Kupfer, E.A. Kuhl, D.A. Regier, DSM-5—The future arrived, *JAMA*. 309 (2013), 1691–1692.
- [18] A.T. Beck, C.H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, *Arch. Gen. Psychiatry*. 4 (1961), 561–571.
- [19] L.S. Radloff, The CES-D scale: a self-report depression scale for research in the general population, *Appl. Psychol. Meas.* 1 (1977), 385–401.
- [20] A.T. Beck, M. Kovacs, A. Weissman, Assessment of suicidal intention: the scale for suicide ideation, *J. Consult. Clin. Psychol.* 47 (1979), 343–352.
- [21] L. Xian, J. Ni, M. Wang, Leveraging large language models for cost-effective, multilingual depression detection and severity assessment, arXiv preprint, arXiv:2504.04891, 2025.
- [22] G. Deng, S. Rao, T. Lin, A. Dai, P. Wang, J. Xie, et al., Evaluating large language models in crisis detection: a real-world benchmark from psychological support hotlines, *IEEE J. Biomed. Health Inform.* (2026), 1–14.
- [23] N.P. Shetty, Y. Singh, V. Hegde, D. Cenitta, D. K, Exploring emotional patterns in social media through NLP models to unravel mental health insights, *Healthc. Technol. Lett.* 12 (2025), e12096.