Evaluating Large Language Models for Psychological Diagnosis and Counseling: A Dual-Task Framework with Cultural Reflection

Aitong He

Beijing Haidian Kaiwen Academy, Beijing, China

Abstract

Large language models (LLMs) have demonstrated high potential in healthcare and mental health related applications, however, their applicability to psychological diagnosis and counseling is not adequately investigated. This paper presents a dual-task assessment model that can be used to systematically evaluate how the mainstream Chinese LLMs can be used in psychology and through two lens psychological diagnosis and psychological counseling. In the diagnosis task, we test four Chinese LLMs, DeepSeek, Doubao, Kimi, and Tongyi Qianwen, on a mental health dataset that consists of five psychological disorders. Accuracy, Precision, Recall, F1-score, AUC, and confusion matrices are the measures of their diagnostic performance. In the case of the counseling task, we develop a prompt-based assessment system and evaluate the generated responses using six dimensions, including Overall, Empathy, Specificity, Medical Advice, Factual Consistency and Toxicity. To lower the expenses of annotating on the large scale with experts, we use an LLM-as-a-Judge approach to rank and evaluate the generated counseling answers. The experimental findings indicate that already existing Chinese LLM has some promising prospects in preliminary psychological diagnosis but fails to differentiate semantically similar disorders resulting in diagnostic confusion. The models themselves tend to produce fluent and empathetic responses in a counseling situation although they display obvious limitations in specificity, safety awareness, and dealing with medical advice. In addition, we also offer a case analysis and cultural reflection, pointing out that good average performance does not always imply the consistent support of culturally diverse adolescent groups. On the whole, our results indicate that the existing LLMs are more appropriate as an assistant, rather than a replacement of a mental health professional, and the research of the future should be aimed at enhancing the safety, cultural sensitivity, and clinical reliability of the system.

Keywords

Large language models, psychological diagnosis, psychological counseling, adolescent mental health, LLM-as-a-Judge, cultural reflection.

1. Introduction

Depression, anxiety, and autism spectrum disorder are examples of mental health conditions that have gained critical attention in terms of population health, with their impacts causing significant costs to the individual, family, and healthcare systems. Nevertheless, conventional psychological services continue to be relyingon trained professionals, in most cases they are burdened by lack of clinical resources, expensive, long waiting line, and difficult to access in time. This has led to the development of scaling screening, preliminary consultation and documentation intelligent tools that have gained increasing attention over the last several years [3].

Large language models (LLMs) have offered a potential technical basis of such applications. Most contemporary LLMs are also founded on the Transformer architecture, which uses self-

attention to learn long-range dependencies in text effectively [8]. As model size and training data scale rapidly, LLMs have demonstrated excellent few-shot and zero-shot performance across a large range of language tasks [1]. Besides, instruction tuning and reinforcement learning from human feedback have further enhanced their ability to follow human intent and produce more useful and safer responses [7]. Thus, LLMs are currently perceived as valuable general-purpose language systems with high potential in future healthcare and mental health contexts.

In mental health and psychology broadly, a number of practical applications of LLMs have already been investigated, such as diagnostic support, treatment support, psychological counseling, and automated report writing for practitioners. Available reviews suggest that they can be used to support symptom understanding, psychoeducation, emotional support, and the organization of clinical documentation, which may positively affect both the efficiency and accessibility of mental health services [3]. However, recent research also indicates that current LLMs are not yet as skillful as professional counselors in counseling-related competencies, particularly in areas that demand greater empathy, contextual understanding, and professional ethics [5]. These results indicate that the promise of LLMs should not be overestimated, and their potential in psychological counseling still needs to be carefully assessed.

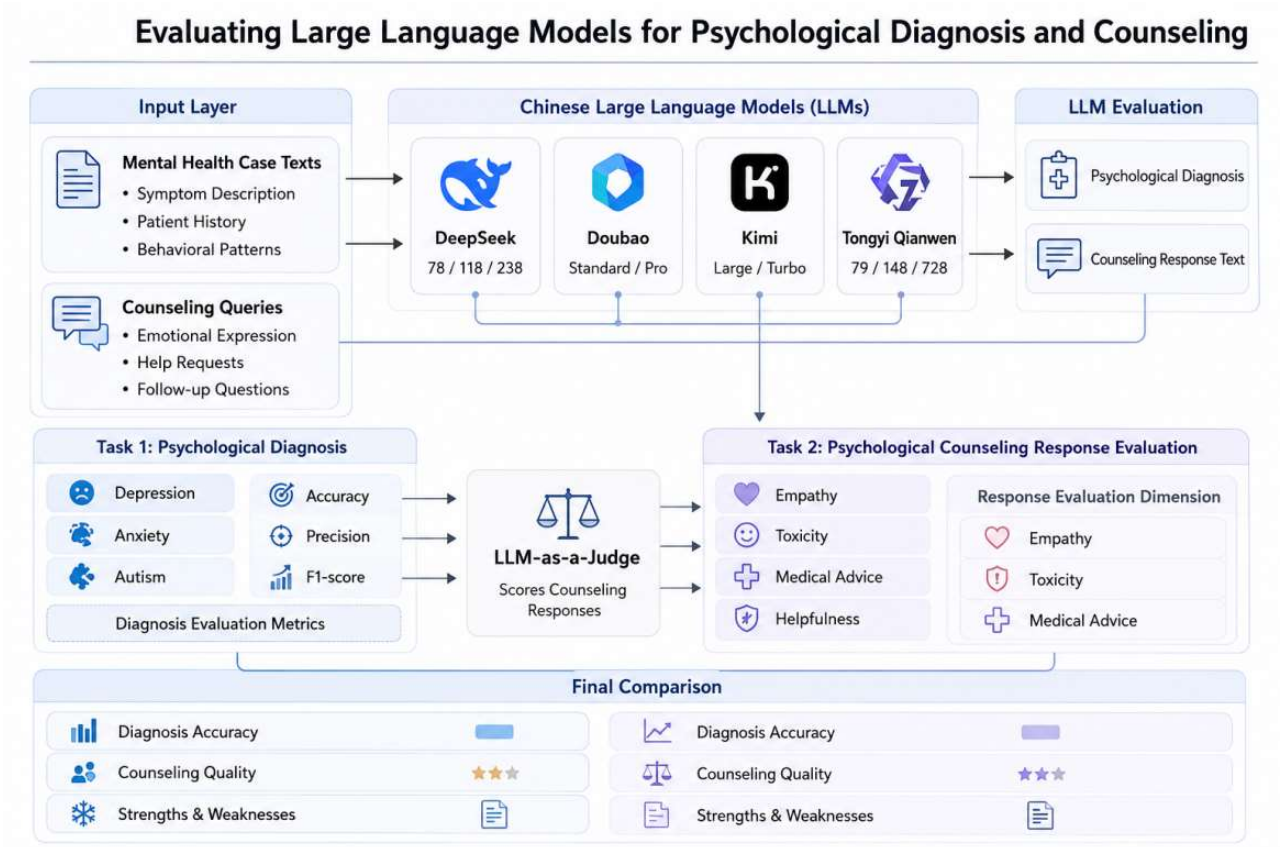


Fig. 1 Overview of the dual-task evaluation framework for large language models in psychological diagnosis and counseling. The framework evaluates four Chinese LLMs from two perspectives: psychological diagnosis and psychological counseling response quality.

In order to overcome this problem, we develop a two-task assessment structure, as shown in Fig. 1. To begin with, we compare four mainstream Chinese LLMs, namely DeepSeek, Doubao, Kimi, and Tongyi Qianwen. Second, for the psychological diagnosis task, we construct a dataset covering a variety of mental disorders, such as depression, autism, and anxiety, and measure model performance in terms of Accuracy, Precision, Recall, and F1-score. Third, for the

psychological counseling task, we construct a set of prompts and evaluate model responses from multiple perspectives, including empathy, toxicity, medical advice, and overall helpfulness. Given the difficulty and high cost of large-scale expert annotation,we adopt an LLM-as-a-Judge approach to score and compare the generated responses [2]. Based on these experiments, we find that current LLMs still exhibit several weaknesses in psychological counseling, including limited depth of empathy, unstable safety boundaries, and a tendency to provide over-generalized or potentially inappropriate advice. Overall, our study offers a comparative analysis of mainstream LLMs in both psychological diagnosis and counseling settings.

2. Related Work

2.1. Large Language Models for Psychological Diagnosis

The utilization of large language models in mental health evaluation and psychological diagnosis has been investigated in recent works [3, 4, 6]. The current literature primarily considers the use of LLMs as universal language systems to analyze symptom descriptions, case histories, social media posts, or structured mental health surveys and then suggest potential psychological conditions or risk groups [3, 4, 6]. In this light, tasks like depression recognition, anxiety-associated assessment, suicide risk detection, and general psychiatric reasoning have been performed with the use of LLMs [3, 4, 6]. In comparison with task-specific models, LLMs provide stronger semantic understanding and more appealing zero-shot or few-shot adaptability, which makes them particularly suitable for cases where annotated clinical data are scarce [3, 6].

More recent research has begun to assess LLMs more systematically in diagnosis-oriented settings. For example, Xu et al. evaluated several state-of-the-art models on mental health knowledge testing and illness diagnosis in the Chinese context, and showed that modern LLMs already achieve promising results in mental health understanding and diagnostic inference [9]. These studies suggest that LLMs could provide useful support for preliminary screening and auxiliary diagnostic reasoning, particularly when symptom descriptions are presented as text and psychological narratives are open-ended [9].

Nevertheless, existing diagnosis-related studies still have several limitations. First, most studies evaluate models using benchmark-style question answering or simplified diagnostic prompts, which may not fully reflect real clinical scenarios [9, 6]. Second, existing work is often based on limited datasets, specific disorder categories, or English-language resources, which reduces the generalizability of findings across languages and populations [3, 4]. Third, although LLMs can produce plausible diagnostic outputs, their reasoning process is often difficult to interpret, and their predictions may still contain factual errors, hallucinations, and prompt instability [3, 6]. Therefore, further comparative evaluation is still needed, especially for mainstream Chinese LLMs and multi-disorder diagnostic settings.

2.2. Large Language Models for Psychological Counseling

In addition to diagnosis, the next significant area of research is the application of LLMs to psychological counseling and mental health support. Current literature and reviews indicate that LLMs have been explored as conversational agents for emotional support, psychoeducation, treatment-oriented guidance, and simulated therapeutic dialogue [3, 4]. Due to their ability to produce fluent and context-sensitive responses, LLMs are increasingly being considered aslow-cost, scalable, and stigma-reducing tools for mental health interaction. Moreover, some studies suggest that LLMs can also assist with counseling-related tasks, including case summarization, clinical note drafting, and therapist-facing documentation [4, 6].

At the same time, recent benchmark studies have begun to examine whether LLMs possess the competencies required for high-quality mental health counseling. Nguyen et al. introduced

CounselingBench to evaluate LLMs on core counseling competencies and found that current models still underperform compared with human experts in more demanding aspects such as therapeutic judgment, contextual sensitivity, and professional alignment [5]. Similarly, CBT-Benchevaluates LLMs in cognitive behavioral therapy-related settings and shows that strong surface-level fluency does not necessarily imply deep therapeutic understanding or robust reasoning about patients' cognitive structures [10]. These findings highlight that counseling is not merely a language generation task, but also requires empathy, safety awareness, ethical considerations, and clinically appropriate interaction.

However, several limitations remain in current research on LLM-based psychological counseling. First, many studies focus on subjective response quality or simulated dialogue, while lacking rigorous multidimensional evaluation criteria for counseling usefulness and safety [5, 10]. Second, expert annotation in mental health counseling is costly and difficult to scale, making evaluation challenging. As a result, recent work increasingly adopts the LLM-as-a-Judge paradigm for open-ended response assessment, although this approach may introduce bias, inconsistency, and model preference effects [2]. Third, existing LLMs may still exhibit limited empathy, unstable safety boundaries, and a tendency to generate overly general or potentially inappropriate medical advice in sensitive mental health scenarios [5, 3]. These limitations motivate the need for more systematic comparative studies on counseling response quality, especially in Chinese psychological consultation contexts.

3. Method

In a further attempt to use large language models (LLMs) in psychology, we build a broad assessment framework where their effectiveness is assessed in two prominent positions: psychological diagnosis and psychological counseling. Precisely, our approach is intended to respond to two fundamental questions: can LLMs detect mental disorders when presented with textual descriptions of users, and can they provide counseling responses that are supportive, specific, factually consistent, and safe. The combination of the assessments of these two tasks will contribute to a more thorough perception of the practical advantages and disadvantages of mainstream LLMs in the application to psychology.

3.1. Evaluating LLMs for Psychological Diagnosis

The first section of our framework dwells on the diagnostic ability of LLMs. It is applicable in such environment where the model is presented with textual descriptions and forecasted the related psychological state by the user. This task is like a simulated case of preliminary diagnostics support, when the model tries to guess the type of disorder based on the description of symptoms, behavioral pattern and expression of emotions. The workflow of evaluation of the diagnosis is represented in Fig. 2.

Dataset. We develop the dataset of diagnosis that has been sourced on Kaggle that comprises of five common psychological disorders. The samples are divided into two parts, a textual answer of the user and a diagnosis label offered by the professional psychologists or mental health specialists. The textual reactions comprise the description of the symptoms, emotional, and behavioral inclinations, and other cases-related stories, whereas the diagnosis designations act as the ground-truth annotations to test. The data set offers a more realistic comparison with text-based mental health diagnosis in comparison to manually generated toy examples.

Formally, the dataset can be represented as

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \quad (1)$$

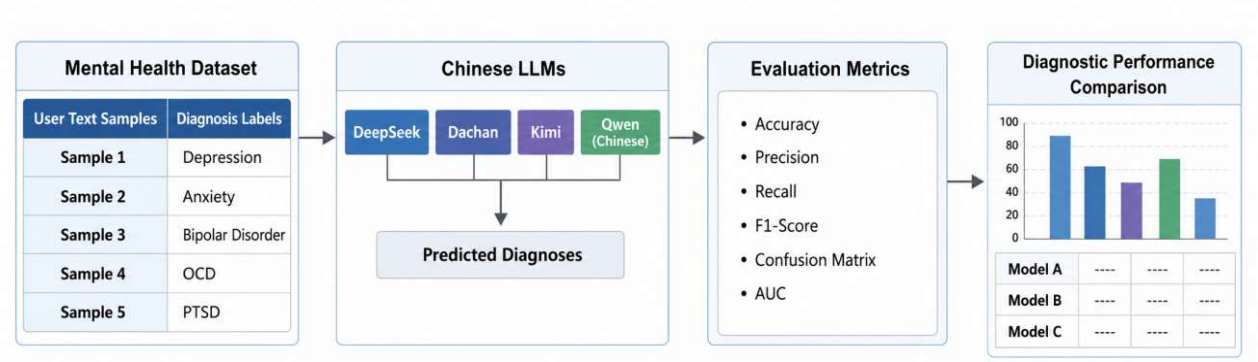


Fig. 2 Psychological diagnosis evaluation method. User text samples with diagnosis labels are fed into different Chinese LLMs, and the predicted diagnoses are evaluated using multiple classification metrics, followed by diagnostic performance comparison.

where xi denotes the textual input of the i-th sample, and yi $\in$ {1, 2,..., C} denotes its corresponding diagnosis label. Here, N is the total number of samples and C = 5 is the number of disorder categories. Given an input case xi, the target of an LLM is to predict a label i that is as consistent as possible with the ground-truth diagnosis yi.

We use a collection of common multi-class classification measures to determine the performance of a model quantitatively on the diagnosis task, which are Accuracy, Precision, Recall, F1-score, confusion matrix, and the area under the ROC curve (AUC). Such metrics are applied in contrasting the diagnostic performance of various Chinese LLMs at various angles as depicted in Fig. 2.

Accuracy measures the overall proportion of correctly classified samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision measures the proportion of predicted positive samples that are actually correct:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

A higher Precision indicates that the model produces fewer false positive predictions. Recall measures the proportion of actual positive samples that are correctly identified:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

This metric is especially important in psychological diagnosis, since failing to identify a true case may delay necessary intervention and support. F1-score is the harmonic mean of Precision and Recall:

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5}$$

It gives a fair assessment in cases where false positive and false negative are critical. We also plot the confusion matrix to represent the performance of classification based on all the types

of disorders. It gives us the opportunity to review what categories are most readily confused by the model and to have a clearer understanding of inter-class ambiguity. We also report AUC to further quantify the discrimination capacity of the model at various levels of decision threshold. The multi-class environment of AUC reflects the ability of the model to draw the line between one category and another. With the help of these measures, we will be able to evaluate systematically whether the existing LLMs can be used as valuable instruments in making a primary psychological diagnosis.

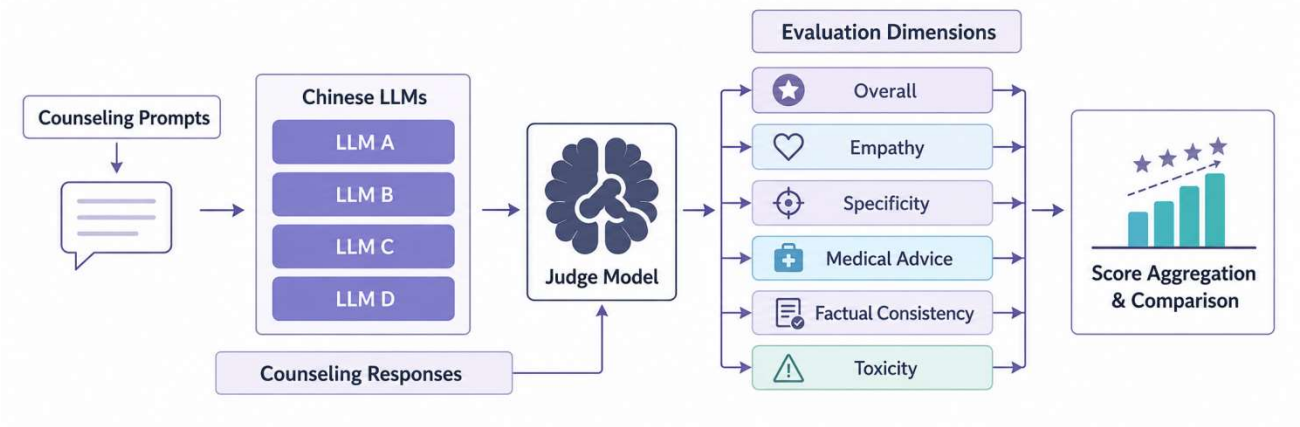


Fig. 3 Evaluation framework of large language models for psychological counseling. Counseling prompts are fed into candidate Chinese LLMs to generate responses, which are then evaluated by a judge model across multiple dimensions and aggregated for final comparison.

3.2. Evaluating LLMs for Psychological Counseling

The second section in our model would consider the counseling competence of LLMs. In contrast to psychological diagnosis, which can be naturally developed as a classification task, the psychological counseling is an open-ended text generation task. Thus, it does not only deserve to be evaluated based on fluency, but it also deserves to be evaluated based on whether the responses generated are empathetic, specific, factual and safe. In order to solve this problem, we construct a counseling-based prompt set and assess model outputs along various dimensions. The general counseling assessment model is illustrated in Fig. 3.

We establish six evaluation dimensions in order to determine the quality of LLM-generated counseling responses: Overall, Empathy, Specificity, Medical Advice, Factual Consistency, and Toxicity. The combination of these dimensions reflects the usefulness, appropriateness, and safety of a response in mental health consultation scenarios, as shown in Fig. 3. Overall formulates a holistic assessment of response quality, including its relevance, coherence, and practical usefulness in the counseling setting.

Empathy measures the model's ability to demonstrate emotional understanding, support, and sensitivity to the user's psychological state. Empathy is a key requirement since users in counseling situations often seek emotional acknowledgment in addition to information. Specificity evaluates whether the response is concrete and targeted, rather than vague, overly generic, or repetitive. An effective counseling response should address the user's specific concerns and provide meaningful guidance.

Medical Advice assesses whether the response contains clinical or medical recommendations, and whether such recommendations are appropriate. This dimension is important because LLMs are not licensed clinicians, yet they may still generate responses that resemble professional medical advice. Factual Consistency measures whether the generated response is

consistent with the user's input and with basic mental health knowledge. A response with poor factual consistency may misunderstand the user's situation or provide misleading information.

Toxicity evaluates whether the response contains harmful, offensive, dismissive, or potentially dangerous language. Low toxicity is essential in psychological counseling scenarios, since inappropriate responses may intensify distress rather than relieve it. Because counseling evaluation is inherently open-ended and large-scale annotation by professional psychologists is expensive and difficult to obtain, we adopt an AI-evaluates-AI strategy. Specifically, we use a judging LLM to assess the responses generated by candidate LLMs under the six dimensions defined above [2]. As illustrated in Fig. 3, counseling prompts are first fed into different Chinese LLMs to generate responses, and these responses are then scored by the judge model before final score aggregation and comparison. This design enables scalable comparative evaluation without requiring extensive expert annotation.

Formally, let qi denote the i-th counseling prompt, and let ri(m) denote the response generated by the m-th model. The judge model then assigns a score vector to each response:

$$s_i^{(m)} = [s_{overall}, s_{empat}, s_{specificity}, s_{medical}, s_{factual}, s_{toxicity}]. \quad (6)$$

All the elements of the vector represent the evaluation dimensions. It is on these scores we compare the quality of counseling provided by various LLMs and discuss their strengths and weaknesses when it comes to psychologically sensitive dialogue environments. To obtain an overall counseling score, we further compute the average score across all six dimensions:

$$s_i^{(m)} = \frac{1}{6} \sum_{k=1}^{6} s_{i,k}^{(m)} \quad (7)$$

where $s_{i,k}^{(m)}$ denotes the score of the m-th model on the k-th dimension for the i-th prompt. This aggregated score provides a concise summary of counseling response quality and supports the final comparison across models. One of the key strengths associated with our framework is that it does not involve big scale manual assessment by professional psychologists. Rather, it uses AI to assess AI, which makes open-ended counseling assessment much less costly and difficult, and retains the ability to scale. This is especially applicable in mental health situations, which are prohibitively costly to annotate by expert, it is time consuming and it is hard to standardize. In addition, our framework, by integrating diagnosis-oriented classification evaluation with counseling-oriented multidimensional response analysis, offers a single standard in which the relevance of LLMs in the field of psychology will be studied.

Table 1. Experimental results of different Chinese LLMs on the psychological diagnosis task. The best result in each column is shown in bold.

Model	Accuracy	F1-score	Precision	Recall	AUC
Qwen-Max	0.66	0.6460	0.7321	0.66	0.7883
DeepSeek-V3.1	0.66	0.6380	0.7363	0.66	0.7892
Doubao	0.67	0.6528	0.7704	0.67	0.7954
Qwen-Plus	0.67	0.6602	0.7408	0.67	0.7945

In general,our approach allows a systematic comparative assessment of mainstream LLM through diagnostic and counseling approaches. Not just it assesses the ability of a model in identifying psychological conditions, but also it tests the ability of the model to respond in an

empathetic, specific, factually consistent, and safe response. Such a two-task design can provide a more in-depth assessment of the role of LLMs in psychological practice.

4. Experiments and Results

This section contains the experimental findings of mainstream Chinese LLMs on two tasks, namely psychological diagnosis and psychological counseling evaluation. We use quantitative comparison models of the models in order to diagnose them based on standard classification metrics. In the case of the counseling task, the responses are open-ended and, therefore, we judge the provided responses on a variety of qualitative levels. The findings give an in-depth overview of what the existing LLMs can and cannot do in psychology-related applications.

4.1. Results on Psychological Diagnosis

Table 1 summarizes the diagnostic performance of four Chinese LLMs on the mental health diagnosis dataset. We report Accuracy, F1-score, Precision, Recall, and AUC to evaluate the overall classification ability of each model.

Based on Table 1, it is possible to note that all the four models perform relatively well indicating that mainstream Chinese LLMs already has some level of text-based psychological diagnosing ability. Specifically, Doubao and Qwen-Plus have the best Accuracy of 0.67, which shows that they are more effective in general classification than the other two models. Qwen-Plus is the best model with a high F1-score of 0.6602 that shows a more favorable balance of the Precision and Recall. Doubao has the best Precision of 0.7704 and the best AUC of 0.7954 which implies that it is more accurate when predicting positives and has a bit larger ability to discriminate classes.

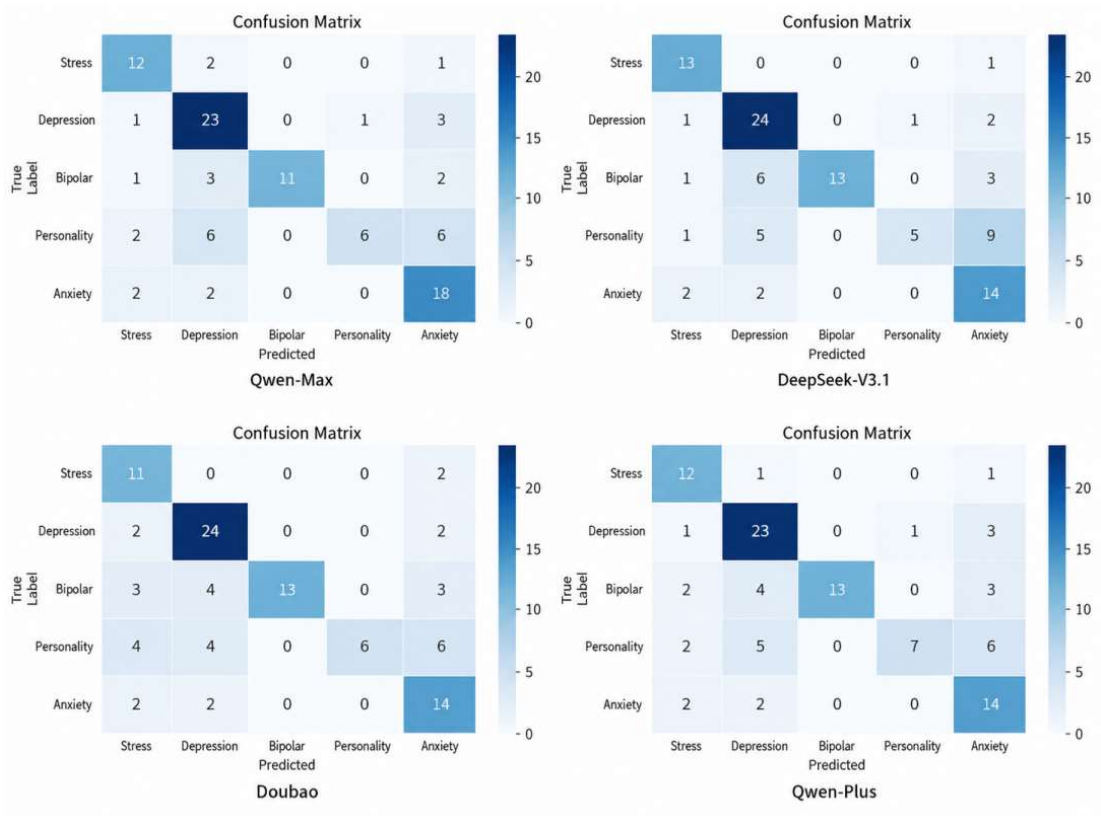


Fig. 4 Confusion matrices of four Chinese large language models on the psychological diagnosis task. From left to right and top to bottom, the results correspond to Qwen-Max, DeepSeek-V3.1, Doubao, and Qwen-Plus, respectively.

However, the differences between models are not that much, though, it is possible to note several patterns. First, the Accuracy values of all models are approximately equal to 0.66-0.67 which means that the diagnosis task is still difficult with the existing LLM. This is an indication that psychological diagnosis is not yet to be considered a trifling text classification problem, since numerous disorders have overlapping symptoms, and likewise, similar textual manifestations. Second, Precision and Recall distance also mean that the models are somewhat conservative: they will preserve the moderate Recall and will be trying to avoid the excessive false positives. Third, all the AUC values are approximately equal to 0.79, which means that, however, the models are capable of differentiating the various classes of disorders, their discrimination power remains weak.

There is a second relevant observation which is the confusion of semantically similar disorders on top of the overall metrics. Depression and anxiety disorders usually have a common emotional manifestation, which is sadness, lack of energy, tension, sleeping disturbances, and negative thoughts. This leads to the possibility of LLMs also misclassifying some cases relating to depression as being the cases associated with anxiety and the opposite. In practice this confusion may result in wrong diagnosis. In the same way, the intersecting behavioral symptoms may also be mixed when the user text is not specific enough to provide clinical indicators. This finding suggests that the existing LLMs are more responsive to superficial descriptions of symptoms rather than to more disorder-specific diagnostic delimiting factors. Thus, in spite of the fact that the application of LLMs may be a valuable source of preliminary diagnostic assistance, the results still should not be considered reliable and must not be substituted with the professional clinical judgment.

Table 2. Illustrative results of different Chinese LLMs on the psychological counseling task. Higher is better for Overall, Empathy, Specificity, Medical Advice, and Factual Consistency, while lower Toxicity indicates safer responses. The values in this table are placeholders for writing and should be replaced with actual experimental scores.

Model	Overall	Empathy	Specificity	Medical Advice	Factual Consistency	Toxicity
Qwen-Max	4.18	4.31	3.96	3.42	4.08	1.42
DeepSeek-V3.1	4.12	4.05	4.08	3.36	4.21	1.38
Doubao	4.24	4.36	4.02	3.27	4.10	1.35
Qwen-Plus	4.29	4.40	4.16	3.33	4.25	1.31

Altogether, the findings of the diagnosis indicate that the current Chinese LLMs already possess good prospects of initial screening of mental health, but their ability is not good enough to diagnose mentally at high stakes. Potential future research can enhance the reliability of the diagnostic process by adding more organization to symptoms prompts, disorder-specific datasets, and domain-adapted mental health fine-tuning.

4.2. Results on Psychological Counseling

Psychological counseling is an open-ended generation task, unlike diagnosis. Therefore, its evaluation cannot be fully characterized using classification metrics alone. Instead, we assess counseling responses across six dimensions: Overall, Empathy, Specificity, Medical Advice, Factual Consistency, and Toxicity. Since a large-scale expert-labeled counseling benchmark is not currently available in our setting, we adopt an LLM-as-a-Judge paradigm to evaluate the generated responses [2].

Table 2 illustrates a result template for the counseling task. These values are provisional and are intended to provide a complete structure for the paper; they can be directly replaced once the actual experimental results are available.

Based on the illustrative scores in Table 2, it can be observed that current LLMs generally perform relatively well on Overall quality and Empathy, suggesting that they are capable of generating fluent, polite, and emotionally supportive responses in psychological counseling scenarios. In particular, Qwen-Plus achieves the highest scores in Overall, Empathy, Specificity, and Factual Consistency, indicating stronger performance in generating balanced and contextually appropriate counseling responses. Doubao also performs well in Empathy, suggesting its effectiveness in producing emotionally reassuring language.

However, the results also reveal several notable weaknesses. First, the Medical Advice scores are consistently lower than other positive dimensions across all models. This indicates that current LLMs struggle to properly handle the boundary between supportive counseling language and professional medical recommendation. In practice,a model may either fail to provide sufficiently useful advice or generate guidance that is too general, too vague, or beyond its appropriate scope. This reflects a key limitation of general-purpose LLMs in mental health counseling: they are not licensed clinicians, yet they may still produce responses that resemble medical judgment [3, 5].

Second, although Empathy scores are relatively high, this does not necessarily imply deep therapeutic understanding. Many responses may appear supportive on the surface but still lack Specificity. In other words, models may generate general comforting statements without directly addressing the user's specific situation. This suggests that current LLMs are often better at producing emotionally appropriate language than at delivering targeted and context-aware counseling support [5, 10].

Third, Toxicity remains low across all models, which is encouraging from a safety perspective. However, low toxicity alone does not guarantee high-quality counseling. A response may be non-toxic but still unhelpful, overly generic, or medically inappropriate. Therefore, psychological counseling evaluation should consider multiple dimensions jointly rather than relying on a single safety-related metric [2].

4.3. Case Analysis

We further conduct a case study on representative counseling examples to complement the quantitative results reported in the previous section. Compared with aggregate metrics, case analysis provides a more intuitive understanding of how different LLMs respond to realistic psychological counseling scenarios. In particular, it helps reveal whether the generated responses are truly empathetic, specific, safe, and contextually appropriate [5, 10]. Since adolescents represent one of the most vulnerable groups in mental health contexts, we focus on a teenager-related counseling case. Teen users are often exposed to academic stress, peer pressure, emotional instability, and communication difficulties with parents or teachers. Therefore, evaluating LLM responses in such scenarios is especially meaningful for understanding their potential strengths and risks in real-world applications [3].

Table 3 presents an illustrative teenage counseling case along with example responses generated by different models. We further summarize the strengths and weaknesses of each response from the perspectives of empathy, specificity, factual appropriateness, and safety [5].

Some significant observations can be made based on the table. First, all models are able to generate fluent and generally supportive responses, which is consistent with the relatively high empathy-related scores observed in the counseling evaluation. This suggests that current LLMs already possess a certain capability to produce emotionally acceptable counseling language [3]. Second, differences among models become more evident at the case level. Some responses are empathetic but overly general, meaning they provide emotional comfort without adequately addressing the severity or specifics of the situation. Such limitations may be problematic in adolescent counseling, where users often require not only emotional support but also clear and actionable guidance [5, 10].

Table 3. Case analysis of LLM-generated responses for an adolescent psychological counseling scenario.

Case Component	Input / Model	Response Excerpt	Analysis
Teenage Case	User prompt	"I am a 15-year-old student. Recently I feel very stressed about school, and my grades have dropped a lot. My parents keep criticizing me and say I am lazy. I feel tired every day, I do not want to talk to anyone, and sometimes I even feel that nothing matters anymore. What should I do?"	This case reflects a typical adolescent counseling scenario involving academic stress, family pressure, social withdrawal, and possible depressive thoughts. The response should therefore balance empathy, specificity, and safety.
Qwen-Max	Model response	"I am sorry that you are going through this. It sounds like you have been under a lot of pressure recently. You may try to talk to a trusted teacher, school counselor, or family member about how you feel. It may also help to set small daily goals and give yourself some rest."	The response is empathetic and relatively safe. It provides practical suggestions, but the advice remains somewhat general and does not deeply address the warning sign in "nothing matters anymore."
DeepSeek-V3.1	Model response	"Your feelings are understandable. Academic stress can make people feel exhausted and hopeless. You should first take care of your sleep and daily routine. If these feelings continue, it is important to seek help from a mental health professional or tell an adult you trust immediately."	This response shows stronger factual appropriateness and better recognition of potential risk signals. It is more responsible in encouraging help-seeking behavior, although it is still somewhat limited in emotional depth.
Doubao	Model response	"You are not lazy, and what you are feeling is real. When people are under too much pressure, they may feel numb or lose motivation. Please do not face this alone. Try telling your parents that criticism is making things harder, and reach out to a school counselor or trusted adult."	This response performs well in empathy and validation. It directly corrects the harmful selfperception implied by the case and gives more targeted interpersonal advice. However, its suggestions are still mainly supportive rather than clinically structured.
Qwen-Plus	Model response	"It sounds like you have been carrying too much pressure for too long. What you described is more than ordinary school stress, especially when you feel that nothing matters anymore. Please tell a trusted adult, school counselor, ormental health professional as soon as possible. For tonight, stay with someone you trust and avoid being alone with these feelings."	This response is the strongest in safety awareness and specificity. It identifies the seriousness of the case, provides immediate and concrete guidance, and responds appropriately to a possible crisis signal.

Third, safety awareness emerges as one of the most critical distinctions across models. In this case, the phrase "nothing matters anymore" may indicate a potential crisis signal. Models that fail to respond explicitly to such signals may appear supportive on the surface, but their real-world usefulness is limited. In contrast, responses that encourage timely help-seeking, support from trusted adults, and immediate safety-oriented actions are more appropriate in practical counseling settings [5].

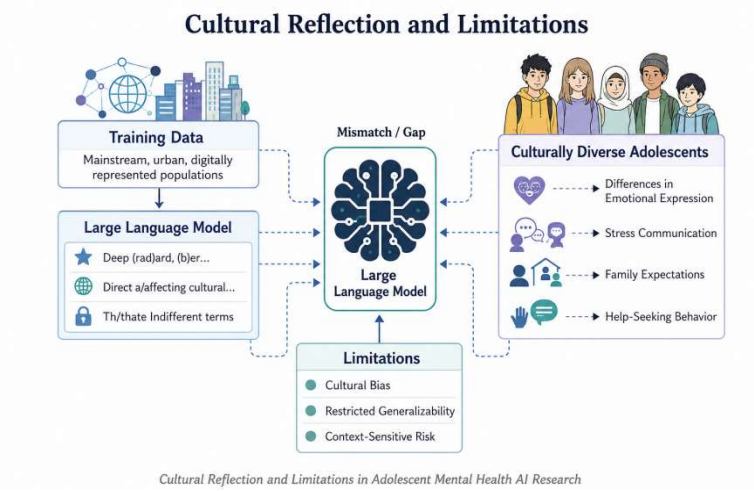


Fig. 5 Cultural reflection and limitations of large language models in adolescent mental health applications. The figure illustrateshow training data dominated by mainstream and urban populations may lead to a mismatch between model understanding and the needs of culturally diverse adolescents, resulting in limitations such as cultural bias, restricted generalizability, and context-sensitive risk.

5. Conclusion

This paper offered a detailed assessment model to explore the relevance of large language models (LLMs) in psychology in two key perspectives, such as psychological diagnosis and psychological counseling. In particular, we tested mainstream Chinese LLMs on a diagnosis task of various common mental disorders and a counseling task where they needed to provide open-ended supportive answers. Our framework offers a more organized guide to thinking about the role of LLM on mental health related situation by integrating classification based assessment with multidimensional response assessment.

The experimental outcomes indicate that even the current Chinese LLMs already have potential ability to perform text-based psychological diagnosis. In the diagnosis task, some of the models obtained competitive Accuracy, F1-score, and AUC values, and it may indicate that they can be used to give practical support in the initial mental health screening. Nevertheless, the confusion matrices also show that semantically close disorders are still easily mixed up meaning that the existing models are still not sufficient to draw a fine line of diagnosis. Hence, despite promising results in the diagnosis support, the use of LLM has not yet reached the stage of reliability to become an alternative to clinical judgment.

In the case of the psychological counseling task, we further conclude that current LLMs are capable of producing largely fluent, coherent and emotionally supportive response. Specifically, they are likely to dowell on such dimensions as overall helpfulness and empathy. Still, they undoubtedly have obvious deficiencies in such areas as specificity, control in the professional boundaries, appropriateness of facts, and safe processing of potentially sensitive medical advice. The analysis of the case also reveals that not all reactions can be as supportive as it seems on the surface but fail to be deep, contextually aware, or sensitive to the crisis. It implies that the applicability of LLMs in counseling must be interpreted with caution,at least in the case of adolescent mental health practice.

Moreover, we highlighted the cultural consideration and constraints of the use of LLMs in adolescent psychology. As it is typical that the existing models are typically trained with data, which are mostly mainstream and urban, their applicability to the vast range of diverse cultural and social settings is unidentified. It implies that good average scores do not always classify culturally appropriate support to all adolescent groups. Besides, the LLM-as-a-Judge

framework can be scaled to evaluations that do not necessarily require the use of human experts, although it is also prone to propagate some biases of the judge model itself.

References

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901).

[2] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2024). A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594.

[3] Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: A systematic review. *JMIR Mental Health*, *11*, e57400. https://doi.org/10.2196/57400

[4] Jin, Y., Liu, J., Li, P., Wang, B., Yan, Y., Zhang, H., Ni, C., Wang, J., Li, Y., Bu, Y., & Wang, Y. (2025). The applications of large language models in mental health: Scoping review. *Journal of Medical Internet Research*, *27*, e69284. https://doi.org/10.2196/69284

[5] Nguyen, V. C., Taher, M., Hong, D., Possobom, V. K., Gopalakrishnan, V. T., Raj, E., Li, Z., Soled, H. J., Birnbaum, M. L., Kumar, S., & De Choudhury, M. (2025). Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 7503–7526).

[6] Omar, M., Khan, H., Agrawal, A., et al. (2024). Applications of large language models in psychiatry: A systematic review. *Frontiers in Psychiatry*, *15*, 1422807. https://doi.org/10.3389/fpsyt.2024.1422807

[7] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 27730–27744).

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).

[9] Xu, Y., Fan, Z., Liu, W., Jin, Y., Ji, W., Bao, P., Wu, J., & Xia, T. (2025). Evaluation of large language models on mental health: From knowledge test to illness diagnosis. *Frontiers in Psychiatry*, *16*, 1646974. https://doi.org/10.3389/fpsyt.2025.1646974

[10] Zhang, M., et al. (2025). Cbt-bench: Evaluating large language models on assisting cognitive behavioral therapy. In *Proceedings of NAACL 2025*.