JUSTC

Article

http://justc.ustc.edu.cn

Received: May 16, 2023; Accepted: November 13, 2023

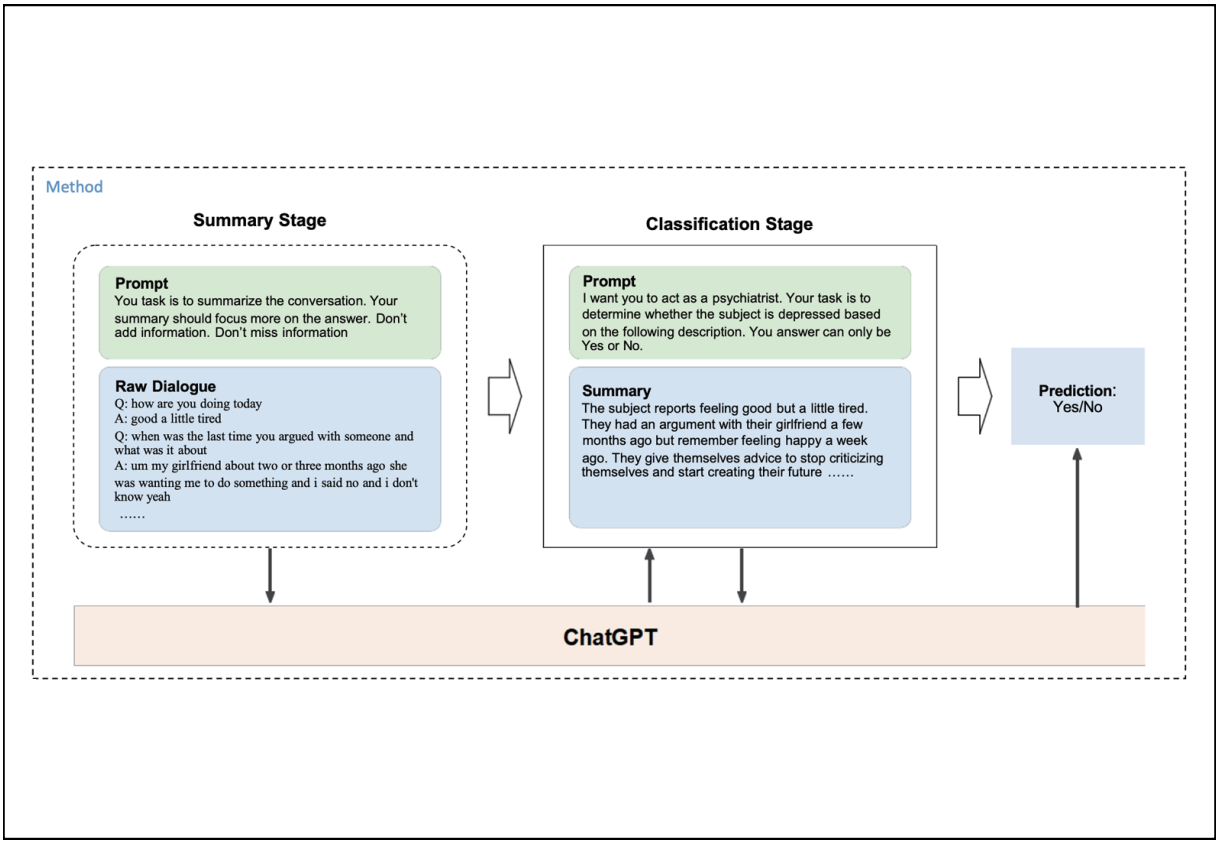
Large language model for interview-based depression diagnosis: an empirical study

Pengbo Hu[1], Hong Li[2], Xingyu Li[3], Chunxiao Li[4], and Yi Zhou[1] ✉

[1]*School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, China;*
[2]*School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China;*
[3]*Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai 201602, China;*
[4]*Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 138602, China*

✉Correspondence: Yi Zhou, E-mail: yi_zhou@ustc.edu.cn

Graphical abstract



Schematic diagram of our two-stage method.

Public summary

- We propose a two-stage method which uses the ChatGPT to summarize the data and then make classifications.
- We test our method with different setting on two distinct interview-based dataset, and achieve comparable performance with deep learning-based SOTA model.
- We systematically analyze the results and perform a detailed ablation study.

Citation: Hu P B, Li H, Li X Y, et al. Large language model for interview-based depression diagnosis: an empirical study. *JUSTC*, **2025**, 55(7): 0705. CSTR: 32290.14.JUSTC-2023-0088. DOI: 10.52396/JUSTC-2023-0088

JUSTC

Article

http://justc.ustc.edu.cn

Received: May 16, 2023; Accepted: November 13, 2023

Large language model for interview-based depression diagnosis: an empirical study

Pengbo Hu[1], Hong Li[2], Xingyu Li[3], Chunxiao Li[4], and Yi Zhou[1] ✉

[1]*School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, China;*
[2]*School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China;*
[3]*Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai 201602, China;*
[4]*Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 138602, China*

✉Correspondence: Yi Zhou, E-mail: yi_zhou@ustc.edu.cn

Cite This: *JUSTC*, **2025**, 55(7): 0705 (8pp)

Read Online

Abstract: The automatic diagnosis of depression plays a crucial role in preventing the deterioration of depression symptoms. The interview-based method is the most wildly adopted technique in depression diagnosis. However, the size of the collected conversation data is limited, and the sample distributions from different participants usually differ drastically. These factors present a great challenge in building a decent deep learning model for automatic depression diagnosis. Recently, large language models have demonstrated impressive capabilities and achieved human-level performance in various tasks under zero-shot and few-shot scenarios. This sheds new light on the development of AI solutions for domain-specific tasks with limited data. In this paper, we propose a two-stage approach that exploits the current most capable and cost-effective language model, ChatGPT, to make a depression diagnosis on interview-based data. Specifically, in the first stage, we use ChatGPT to summarize the raw dialogue sample, thereby facilitating the extraction of depression-related information. In the second stage, we use ChatGPT to classify the summarised data to predict the depressed state of the sample. Our method can achieve approximately 76% accuracy with a text-only modality on the DAIC-WOZ dataset. In addition, our method outperforms the performance of the state-of-the-art model by 6.2% in the D^4 dataset. Our work highlights the potential of using large language models for diagnosis-based depression diagnosis.

Keywords: depression diagnosis; language model; deep learning

CLC number: R749 **Document code:** A

1 Introduction

Depression is a mental disorder that can lead to suicidal behavior in severe cases[1,2]. The symptoms of depression usually become more serious and last a long time without early intervention or treatment. To ameliorate this situation, a critical strategy is to develop an automatic diagnostic method and timely intervention. Therefore, designing a reliable automatic depression detection algorithm is an urgently needed. Clinical interviews are commonly used in the diagnosis of depression[3]. During the interview process, the attending physician asks participants predesigned questions and judges their emotional state through facial expression and tone. By combining the participants' responses with the observed emotional state, the physician can make an accurate diagnosis of depression severity in participants.

To improve the efficiency of depression diagnosis, researchers have explored the automation of the process mentioned above via deep learning models. Most existing approaches utilize deep learning techniques, typically designing a multi-modal neural network and training it end-to-end[4–7]. These methods rely heavily on large-scale training data in order to achieve robust performance. However, the collected responses usually do not meet the requirements. First, to make participants feel comfortable and natural, the attending physician needs to adjust the number and order of questions. Furthermore, the attending physician may provide "irrelevant but important" questions to ensure dialog fluency. Consequently, the data collected from each participant usually demonstrate distinct distributions, affecting the generalization of deep learning models[3,8]. Second, the size of an interview-based dataset is typically small owing to the high cost of data collection. Without sufficient data, deep learning models can barely learn useful features and tend to overfit training samples, which constrains their practical use.

Recently, large language models (LLMs) have become a central topic. Powered by their zero-shot and few-shot generalization capacities, LLMs have found their applications in various fields. The most powerful and cost-effective model available is ChatGPT[9,10], which is featured by its close-to-human-level performance. Therefore, numerous studies have explored the use of ChatGPT for complex tasks. For example, Fu et al.[11] proposed an evaluation framework, the GPTScore, which uses ChatGPT to score the generated text on 4 types of text generation tasks. They reported that ChatGPT is able to perform multi-aspect, customized, and training-free evaluation without human-annotated samples. Wang et al.[12] systematically tested ChatGPT on 18 sentiment benchmark

CSTR: 32290.14.JUSTC-2023-0088
DOI: 10.52396/JUSTC-2023-0088
JUSTC, **2025**, 55(7): 0705

datasets. They observed that ChatGPT achieves impressive zero-shot performance in sentiment classification tasks and is even better than the fully fine-tuned BERT Model. Rao et al.[13] proposed a general framework to assess the personalities of a large language model. On the basis of the Myers–Briggs type indicator (MBTI) tests, their experiments indicated that ChatGPT can well understand human personality with appropriate prompts. The ever-growing list of applications on diverse domains demonstrates the generality of ChatGPT.

In this study, we present a systematic exploration of the use of the ChatGPT for depression diagnosis through interview-based assessments. Our investigation focuses on the application of ChatGPT in scenarios involving small datasets, where no model training is conducted. Specifically, we design a two-stage pipeline to handle the depression state prediction task. The first stage is called the summary stage, which applies ChatGPT to summarize the raw conversation data sample. This significantly reduces the sample length while retaining critical information related to depression features. The stage is of practical importance, as the conversation samples in real-world datasets are normally redundant and often exceed the maximum length limit of ChatGPT. In the second stage, the summarized data are fed into ChatGPT for depression diagnosis as a classification task. We carry out experiments under both zero-shot and few-shot settings with two distinct datasets. The results show that our pipeline significantly improves the diagnostic accuracy compared to the method that directly use raw conversation samples. In addition, we observe that the performance outperforms the SOTA model in D^4[8] dataset by more than 6.2%. To the best of our knowledge, this work is the first to systematically apply ChatGPT to the interview-based depression diagnosis, highlighting the potential of using a large language model for domain-specific tasks with limited data.

Our contributions are threefold:

(Ⅰ) We propose a two-stage method that uses the ChatGPT to summarize the data and then make classifications.

(Ⅱ) We test our method with different settings on two distinct interview-based datasets, and achieve comparable performance with the deep learning-based SOTA model.

(Ⅲ) We systematically analyze the results and perform a detailed ablation study.

2 Methods

2.1 Dataset

Dataset description. We test our method on two distinct datasets: the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) dataset[3] and the Chinese Dialog Dataset for Depression Diagnosis-oriented Chat (D^4) dataset[8]. Both of them are collected through clinical interviews and are generally used to aid in diagnosing psychological distress conditions for depression. The main difference between the two datasets lies in their inherent data structures. First, the DAIC-WOZ dataset has a well-structured organization in which questions are predesigned, resulting in question-answer pairs between physicians and participants that can be automatically segmented. In contrast, the D^4 dataset reveals an open-and-free dialog structure, with no direct ability to automate the extraction of question-answer pairs. Second, the DAIC-WOZ dataset contains more irrelevant quesiton-answer pairs whereas almost all dialogs in D^4 are focus on the query of depression symptoms. Therefore, the D^4 dataset contains less redundant information than the DAIC-WOZ dataset does. Third, the linguistic diversity provided by the two datasets is as follows: the DAIC-WOZ dataset is collected in English, whereas the D^4 dataset is collected in Chinese. These distinctions between the two datasets allow us to evaluate and validate our proposed methodology across a broader spectrum of scenarios, thereby demonstrating the robustness and generalizability of our approach.

The DAIC-WOZ dataset contains 107 training samples, 35 development samples, and 47 testing samples. Overall, there are 56 positive samples and 133 negative samples. The D^4 dataset contains 1056 training samples, 151 development samples, and 132 test samples. In total, there are 430 negative samples and 909 positive samples. In our experiments, we use ChatGPT for classification tasks under both zero-shot and few-shot settings. Therefore, we do not need to train the model. To balance the test dataset, we select the same number of negative and positive samples. Specifically, we select 56 positive and 56 negative samples (randomly selected) from the total 189 samples of the DAIC-WOZ dataset. Among them, 46 positive and 46 negative samples are used for testing, and the other 10 positive and 10 negative samples serve as few-shot examples (named the few-shot set). For the D^4 dataset, we randomly select 160 positive and 160 negative samples from the total 1339 samples. Among them, 150 positive and 150 negative samples are used for testing, and the remaining 10 positive and 10 negative samples serve as few-shot examples. Note that we do not use all samples due to the high cost of the ChatGPT API[10].

Data preprocessing. For the DAIC-WOZ dataset, we rank the questions in the dataset according to their frequency of occurrence, and only the top half of the most frequent questions are used as our test data. This procedure resulted in 26 high-frequency questions, all related to the diagnosis of depression. The question types and sample frequencies are shown in Table 1. The reason for this step is that when collecting data, the attending physician may need to casually add questions that are not in the predesigned question list to make the participant feel comfortable and natural. As a result, the sample may contain irrelevant question-answer pairs. The data preprocessing helps to reduce the irrelevant information. For the D^4 dataset, since the dialogue data contain less irrelevant information, we directly use the raw dialog to test our method without any extra pre-processing.

2.2 Two-stage classification

Directly performing depression diagnosis with large language models on raw conversation data often yields unsatisfactory results. As shown in Table 2, we directly input the raw conversation data to the ChatGPT and instruct it to make a prediction, resulting in an accuracy of 0.576 and an F1-score of 0.290. We speculate that even with filtered data, there is still much irrelevant information that deteriorates the language model's predictions. Therefore, we propose a

CSTR: 32290.14.JUSTC-2023-0088
DOI: 10.52396/JUSTC-2023-0088
JUSTC, **2025**, 55(7): 0705

Table 1. Frequency of questions.

Num.	Frequency	Question
1	184	How are you doing today?
2	182	When was the last time you argued with someone and what was it about?
3	179	When was the last time you felt really happy?
4	177	What advice would you give to yourself ten or twenty years ago?
5	176	How are you at controlling your temper?
6	173	How easy is it for you to get a good night's sleep?
7	171	What are you most proud of in your life?
8	167	Have you been diagnosed with depression?
9	165	Have you ever been diagnosed with PTSD?
10	164	How would your best friend describe you?
11	162	What did you study at school?
12	161	What do you do to relax?
13	158	What are some things you really like about Los Angeles?
14	151	Is there anything you regret?
15	146	What are some things you don't really like about Los Angeles?
16	144	What's your dream job?
17	144	Who's someone that's been a positive influence in your life?
18	126	What's one of your most memorable experiences?
19	124	Have you noticed any changes in your behavior or thoughts lately?
20	124	What are you like when you don't sleep well?
21	119	What do you enjoy about traveling?
22	117	Do you consider yourself more shy or outgoing?
23	111	Tell me about a situation that you wish you had handled differently.
24	109	Tell me about the hardest decision you've ever had to make.
25	103	What are some things you wish you could change about yourself?
26	101	What would you say are some of your best qualities?

two-stage approach, as shown in Fig. 1. Specifically, in the first stage, we use ChatGPT to summarize raw conversation samples to extract depression-related information and reduce data redundancy. In the second stage, we employ ChatGPT to classify the summarized text. Our experiments include both zero-shot and few-shot settings, and compare the performance of different summarization strategies. As shown in Tables 2 and 3, compared with using the raw conversation sample, the accuracy significantly improves when using summarized data.

2.2.1 Summary stage

In the summarization stage, we use ChatGPT to summarize the raw conversation data. The purpose is to extract features associated with depression while reducing irrelevant information. For comparison, we employ four types of prompts for different levels of summary generation. The corresponding prompt templates are shown in Figs. 2 and 3. (ⅰ) Plain summary. We simply instruct the language model to summarize a given sample without providing any depression-related words. In this setting, language models typically do not pay extra attention to features associated with depression. (ⅱ) Context summary. As inspired by Ref. [14] which added several principles to induce LLMs to focus more on principle-related information, we add some extra information, namely, *"I want you to act as a psychiatrist. Your task is to summarize the conversation to help the psychiatrist determine if the subject is currently depressed."* on top of the plain summary prompt. We aim to encourage the language model to focus more on content related to depression when making a summary. (ⅲ) Sentence-level summary. This method does not ask the model to summarize the entire sample directly. Instead, we summarize each question-answer pair in the sample separately and then aggregate the summarized content. For example, *"Question: How easy is it for you to get a good night's sleep. Answer: It's pretty good somewhat."* can be summarized as *"The subjects ability to get a good night's sleep is somewhat good."* In this setting, we can greatly reduce the sample length while maximizing the retention of its complete information. (ⅳ) Aspect summary. As shown in Fig. 3, This method adds extra six aspects of depression symptom to the prompt, which try to help model focus on these symptom when making summary.

2.2.2 Classification stage

In this stage, ChatGPT classifies the depressed states on the basis of the summary samples generated in the previous stage. We refer to the current classification scheme as summary-based classification, and test the performance in both

Table 2. Main results for DAIC-WOZ.

DAIC-WOZ	Zero-shot		Zero-shot-CoT		Few-shot	
	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
BiLSTM[7]	–	–	–	–	–	0.83
Raw conversation	0.576	0.290	0.641	0.645	–	–
Plain summary	0.630	0.499	0.706	0.682	0.708	0.655
Sent-level summary	0.706	0.689	0.717	0.723	0.721	0.761
Context summary	**0.739**	**0.684**	**0.750**	**0.747**	**0.765**	**0.764**
Aspect summary	0.695	0.702	0.728	0.736	0.713	0.727

We evaluate the accuracy (Acc.) and F1-score of different summary methods under three evaluation settings. Bold text indicates that this method performs best. Raw conversation represents the original sample without any Summary operation, and Sent represents sentence. We place the results of the deep learning based SOTA model on the few-shot column for convenience.

CSTR: 32290.14.JUSTC-2023-0088
DOI: 10.52396/JUSTC-2023-0088
JUSTC, **2025**, 55(7): 0705

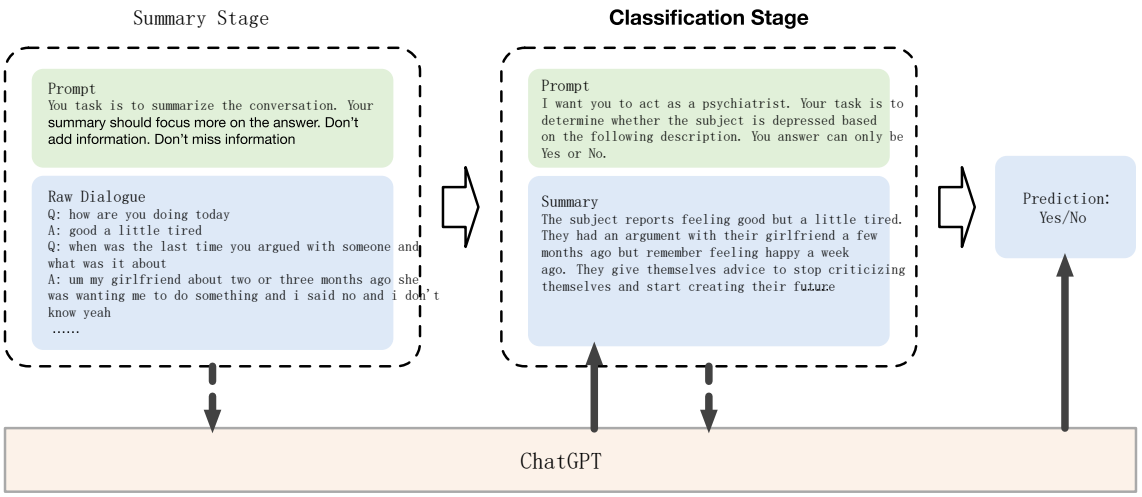


Fig. 1. Schematic diagram of our two-stage method. The dashed line refers to ChatGPT input and the solid line refers to ChatGPT output.

Table 3. Main result for D^4.

D^4	Zero-shot		Zero-shot-CoT		Few-shot	
	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
CPT[8]	–	–	–	–	–	0.820
Raw conversation	0.723	0.783	0.776	0.817	–	–
Plain summary	**0.800**	**0.833**	**0.826**	**0.852**	**0.876**	**0.886**
Context summary	0.760	0.806	0.790	0.826	0.806	0.837
Aspect summary	0.660	0.744	0.713	0.777	0.683	0.758

The setting of the table is same as that of Table 2. We place the results of the deep learning based SOTA model on the few-shot column for convenience. Bold text indicates that this method performs best.

zero-shot and few-shot settings. The prompt templates for different classification settings are illustrated in Fig. 2. In zero-shot classification, we directly instruct the model to predict whether the subject is depressed or not based on the given summary. Chain-of-thought (CoT) has been extensively proven that can significantly improve the performance of language models in downstream tasks[15]. Here, we use the most common setting of CoT, namely Zero-shot-CoT[16], to instruct the model to make predict. It simply adds a sentence *"Let's think step by step."* after the "Answer" token in the prompt, which can induce the model to "think" before answering the question, thereby improving the model's accuracy. In few-shot classification, we randomly select n samples from the few-shot set (see Section 2.1) as examples and concatenate them together with the zero-shot prompt[17]. Since depression diagnosis is a binary classification task, to reduce the bias effect of few-shot samples, we select the same number of samples from the positive and negative samples and randomly shuffle the order. For example, if we set $n = 4$, we will randomly select 2 negative and 2 positive samples as few-shot examples.

3 Experiments

3.1 Setup

We use the ChatGPT (GPT-3.5-Turbo-030[10]), the most capable and cost-effective language model, for all of our experiments. To generate our outputs, we employ greedy decoding with temperature $\tau = 0$, which results in deterministic outputs for identical inputs. For few-shot experiments, we use five random seeds for sampling examples and report the average performance. For evaluation metrics, we use Acc. to assess the overall classification accuracy. Additionally, we use F1-score, which takes into account both precision and recall, to provide a more comprehensive measure of the model's performance.

3.2 Main results

Results of classification on the raw preprocessed samples (raw conversation) are included as the baseline. Note that it is infeasible to test the few-shot setting in this case because of the input length limit of ChatGPT. The length of the original samples is too long to be included as the few-shot examples together with the actual input question-answer pair. In the DAIC-WOZ dataset, as shown in Table 2, all summary-based methods outperform the methods that use raw conversation samples. Our method still falls short of the performance achieved by deep learning-based SOTA model BiLSTM[7]. Nonetheless, it is worth highlighting that our approach does not involve any training procedures. Consequently, the obtained results surpass our initial expectations. In Table 3, the plain summary and context summary method outperforms the method using raw conversation. In addition, we observe that the utilization of plain summary and few-shot classification approach yields results that surpass the SOTA model CPT (An end-to-end trained deep learning model) by a margin of

CSTR: 32290.14.JUSTC-2023-0088
DOI: 10.52396/JUSTC-2023-0088
JUSTC, **2025**, 55(7): 0705

Plain Summary Prompt: Stage-1: Summary

""" You task is to summarize the conversation. Your summary should focus more on the answer. Don't add information. Don't miss information\n\n{dialogue sample} """

Context Summary Prompt:

""" I want you to act as a psychiatrist. Your task is to summarize the conversation to help the psychiatrist determine if the subject is depressed currently. Your summary should focus more on the subject's answer. Don't add information. Don't miss information\n\n{dialogue sample} """

Sentence-level Summary Prompt:

"""Summarize the following dialogue into one sentence from the perspective of the respondent. Using "The subject" to refer to the respondent. Don't add information. Don't miss information.
Question:{question}\nAnswer:{answer}"""

Zero-shot Classification Prompt: Stage-2: Classification

""" I want you to act as a psychiatrist. Your task is to determine whether the subject is depressed based on the following description. You answer can only be Yes or No.\n\n Summary:{summary}\n Answer: """

Zero-shot-CoT Classification Prompt:

""" I want you to act as a psychiatrist. Your task is to determine whether the subject is depressed based on the following description. You answer can only be Yes or No.\n\n Summary:{summary}\n Answer: Let's think step by step. """

Few-shot Classification Prompt:

""" I want you to act as a psychiatrist. Your task is to determine whether the subject is depressed based on the following description. You answer can only be Yes or No.\n\n {examples} \n Summary:{summary}\n Answer: """

Fig. 2. Prompt template for our two-stage method.

Prompt Template:

""" I want you to act as a psychiatrist. Your task is to summarize the conversation to help the psychiatrist determine if the subject is depressed. You should focus more on {aspect of depression symptom}. Your summary should focus more on the answer. Don't add information. Don't miss information\n\n """

Aspect

1. if the subject has the symptoms of low mood, anxiety, helplessness.
2. if the subject has experienced physical symptoms such as sleep disorders, changes in appetite.
3. if the subject has impaired attention, memory, decision-making abilities.
4. if the subject has lost interest in hobbies or social activities.
5. if the subject has negative emotions such as low self-esteem, self-blame.
6. if the subject has had suicidal thoughts, plans, or behaviors.

Fig. 3. Prompt template for aspect summary.

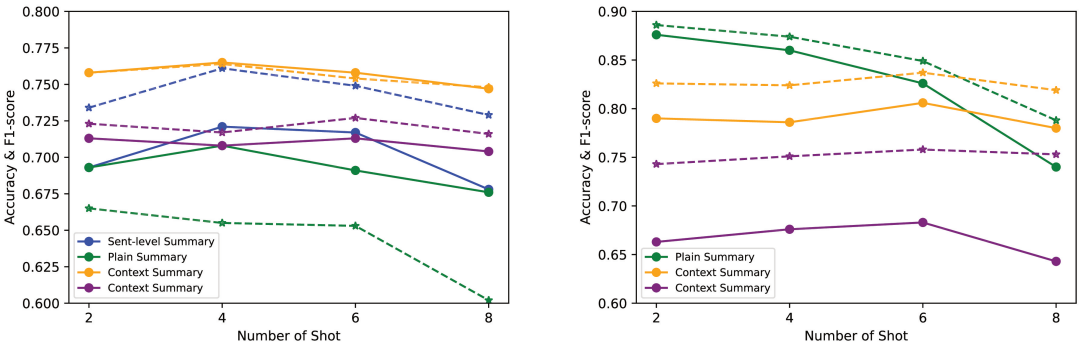
6.2%. This demonstrates the effectiveness of our two-stage approach, where we first generate a summary from the raw input and then carry out classification using the summary. We notice that the aspect summary in the D^4 dataset is worse than the method using raw conversation. We speculate that the main reason is that most samples do not contain all the content related to the six aspects of the symptoms, and thus this approach distracts ChatGPT, resulting in confusing summaries.

Based on Table 2, the context summary method performs best in zero-shot, zero-shot-CoT, and few-shot settings. This is in line with our assumption. First, plain summary provides only a summary of the sample without paying extra attention to features related to depression. Consequently, its summary may overlook many features related to depression. Second, the sentence-level summary almost retains all the information but contains much redundant information. This caused interference with the model, leading to inaccurate judgments. Finally, the context summary can guide ChatGPT in extracting more depression-related features, thereby leading to more accurate predictions. However, for D^4 dataset as shown in Table 3, a different phenomenon is observed, where plain summary achieves the best results. The main reason is that the dialog content in the sample is mainly about inquiring depressive symptoms, and thus contains less redundant information. Thus, summarizing it without any attention guidance can better preserve the original information and thus help to make better predictions.

We observe that the few-shot setting (with 4 examples) consistently performs better than the zero-shot and zero-shot-CoT settings in both DAIC-WOZ (see Table 2) and the D^4 (see Table 3). This phenomenon suggests that providing additional examples in the prompts can improve the model performance on the depression diagnosis task. Nonetheless, the aspect summary exhibits a distinct behavior, whereby the results for the zero-shot-CoT, in both datasets, exceed those produced under other settings. This may suggest that zero-shot-CoT can potentially optimize the reasoning capacity of ChatGPT on confusing summary samples.

Furthermore, we evaluate the effectiveness of the model by testing its performance with 2, 4, 6, and 8 examples. As shown in Fig. 4, we find that the model consistently achieves the highest performance when using four examples across different summary methods (excluding of Aspect Summary, which performs best with 6 examples) in the DAIC-WOZ dataset. Conversely, in the D^4 dataset, we observe that both the context summary and the aspect summary achieved optimal results with six examples, whereas the plain summary exhibits its highest performance with two examples. This observation implies that there is no definitive number of examples that can guarantee optimal performance. Hence,

CSTR: 32290.14.JUSTC-2023-0088
DOI: 10.52396/JUSTC-2023-0088
JUSTC, **2025**, 55(7): 0705



(a) DAIC-WOZ (b) D^4

Fig. 4. Few-shot classification. We conduct tests using 2, 4, 6, and 8 demonstrations under four summary methods (The D^4 dataset does not have sentence-level Summary) in the DAIC-WOZ and D^4 datasets. The solid line refers Accuracy, dashed line refers to F1-score.

few-shot approaches necessitate empirical evaluation. Moreover, we observe a consistent downward trend in the performance of all methods across both datasets when 8 examples are used. The main reason is that the sensitivity of the language model to the prompt.

3.3 Ablation study

Which questions are more important? According to the above experiment, the sentence-level summary, which retains most of the information, does not perform well. The reason may be that there is too much irrelevant information, which may affect the model predictions. We want to know which questions are more important for diagnosing depression, and whether there are questions with low correlation. To answer this question, we design an ablation experiment. Specifically, on the DAIC-WOZ dataset, we use the sentence-level summary samples as the basis, and delete one question-answer pair each time to observe the impact of this deletion on the model performance. After that, we measure the importance of the corresponding question-answer pairs by the degree of F1-score descent in the zero-shot classification setting. In addition, since not every sample contains all 26 questions, we normalize the F1-score drop based on the frequency of questions in the overall sample. We define decline rate for question i as $\frac{n_i}{N}(f1_{\text{score}} - f1^{i}_{\text{score}})100$, where n corresponds to the number of question i that appear in the overall sample, N represents the total number of samples, and $f1_{\text{score}}$ is the F1-score of the Sentence-level Summary samples under the zero-shot setting as shown in Table 2. $f1^{i}_{\text{score}}$ represents the F1-score of the Sentence-level Summary samples after removing the i question-answer pair, under the zero-shot setting. Note that we multiply the final result by 100 for better visualization.

As shown in Fig. 5, we observe that there are indeed many questions, such as 2, 5, 7, 9, 13, and 15, and these variables have a weak impact on the model performance. In addition, we surprisingly find that, compared with other questions, question 8 has a more significant effect on performance. Intuitively, it makes sense for the model to rely more on whether the subject has a history of depression when assessing for depression. We also find that question 6 significantly impacts the model's performance. This may indicate that sleep quality is a key feature of depression diagnosis. These phenomena are consistent with Ref. [18] which also observes that sleep quality and the history of depression are vital features for depression prediction.

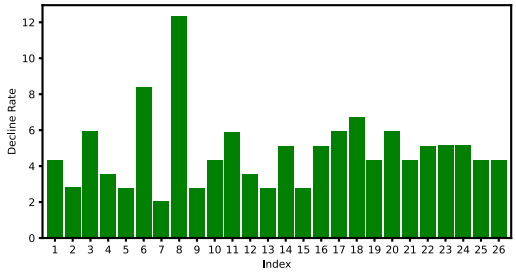


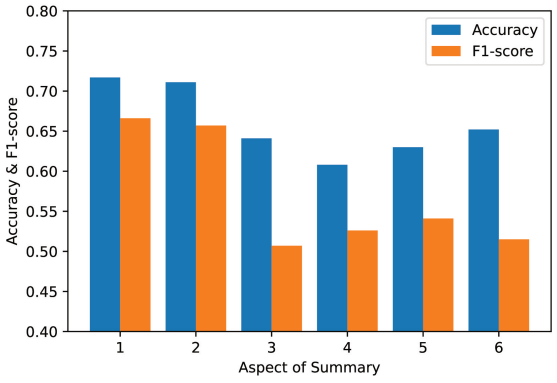
Fig. 5. Ablation study. The degree of F1-score decline caused by removing a specific question-answer pair on sentence-level summary samples.

Which aspect of information is important? In the above experiment, we have demonstrated that context summary can achieve better results on the DAIC-WOZ dataset. The primary explanation is that by incorporating task-related information into the prompt, the model can better concentrate on features related to diagnosing depression when summarizing. Using this method, we can analyze which aspects of feature are more important in diagnosing depression. Therefore, we manually analyze the dataset samples and identify six aspects of features related to depression. We include these aspects of feature in the prompt respectively to instruct the model to emphasize to the corresponding aspect when generating a summary. We then evaluate these summaries to determine which features are more important for depression. The aspects of the depression feature and the prompt template are presented in Fig. 6. We observe that the symptoms of low mood, anxiety, helplessness, and sleep disorders are most important for depression diagnosis in the DAIC-WOZ dataset. This result is consistent with the above ablation study. In addition, we find that performance of summary with aspects 1 and 2 in the zero-shot setting is better than the plain summary, which indicates that inducing the model to make a summary along with task-related information is an essential skill for feature extraction. However, the D^4 dataset presents distinct behavior as shown in Fig. 6b, where each aspect of summary displays relatively comparable performance. This further demonstrates that the dialogue content in the sample contains more queries about depression symptoms, thereby facilitating impressive performance in generating summaries for each aspect separately.

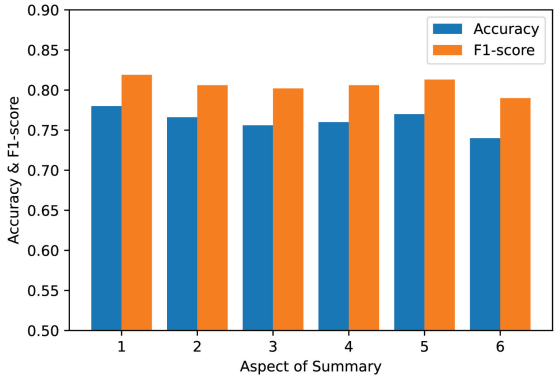
3.4 Limitation

Despite the outstanding zero-shot and few-shot performance

CSTR: 32290.14.JUSTC-2023-0088
DOI: 10.52396/JUSTC-2023-0088
JUSTC, **2025**, 55(7): 0705



(a) DAIC-WOZ



(b) D^4

Fig. 6. Ablation study. The Accuracy and F1-score corresponding to the summary with different aspect added in the prompt.

of our proposed pipeline, there is still room to improve. First, we note that the training corpus of ChatGPT does not include a wide range of medical data[9]. We believe that fine-tuning of the model on medical data, especially on psychiatric diagnostic data, could further boost its performance. Second, the current work only uses text modality and ignores the speech and visual modalities, which usually possess complementary information that can further improve the accuracy of depression diagnosis. Fortunately, recent studies indicated that transforming different modalities into text formats enables large language models to accomplish multimodal tasks[19,20]. In the future, we will explore the use of a large language model for depression diagnosis under the multimodel situation.

4 Conclusions

In this study, we systematically explore the potential of using ChatGPT for depression diagnosis tasks. Our results show that ChatGPT achieves outstanding performance with the text-only modality under zero-shot and few-shot settings. Moreover, we observe that adding task-related information in the prompt can guide the model to focus more on the relevant features of the underlying task, thus leading to better performance. In addition, the experiments under the few-shot setting demonstrate that providing the model with extra classification examples helps it make more accurate judgments. In conclusion, we believe that the large language models, such as ChatGPT, have great potential and can be widely applied to the interview-based diagnosis of mental disorders, helping to relieve the burden on medical resources and greatly promoting the progress of automatic depression diagnosis. We call for more adoption of the current powerful language models to help solve these types of tasks.

Acknowledgements

This work was supported by the Science and Technology Innovation 2030 Project of China (2021ZD0202600).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Pengbo Hu is currently pursuing a Ph.D. degree in Applied Mathematics at the University of Science and Technology of China. He received his B.E. degree in Computer Science from North China Electric Power University in 2014 and the M.S. degree in Neural Science at the University of Chinese Academy of Sciences in 2019. His research mainly focuses on brain-inspired intelligence, multi-modal algorithm, and language model.

Yi Zhou is currently a Professor in the School of Information Science and Technology, University of Science and Technology of China (USTC). He received his Ph.D. degree in Computer Science from USTC in 2011. His research mainly focused on cognitive artificial intelligence, brain-inspired intelligence, AI-based automated math question answering for International Math Olympiad (IMO), and AI solutions for real-life applications.

References

[1] Friedli L. Mental health, resilience and inequalities. Copenhagen, Denmark: WHO Regional Office for Europe, **2009**.
[2] Ringeval F, Schuller B, Valstar M, et al. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. New York: ACM, **2019**: 3–12.
[3] Gratch J, Artstein R, Lucas G M, et al. The Distress Analysis Interview Corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association, **2014**: 3123–3128.
[4] Lam G, Huang D, Lin W. Context-aware deep learning for multi-modal depression detection. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE, **2019**: 3946–3950.
[5] Yin S, Liang C, Ding H, et al. A multi-modal hierarchical recurrent neural network for depression detection. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. New York: ACM, **2019**: 65–71.
[6] Shalu H, Sankar CN H, Das A, et al. Depression status estimation by deep learning based hybrid multi-modal fusion model. arXiv: 2011.14966, **2020**.
[7] Shen Y, Yang H, Lin L. Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, **2022**: 6247–6251.
[8] Yao B, Shi C, Zou L, et al. D^4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. arXiv: 2205.11764, **2022**.
[9] Ouyang L, Wu J, Jiang X, et al. Training language models to follow

CSTR: 32290.14.JUSTC-2023-0088
DOI: 10.52396/JUSTC-2023-0088
JUSTC, **2025**, 55(7): 0705

instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM, **2021**: 27730–27744.

[10] OpenAI. ChatGPT. **2022**. https://openai.com/blog/openai-api. Accessed January 01, 2023.

[11] Fu J, Ng S-K, Jiang Z, et al. Gptscore: Evaluate as you desire. arXiv:2302.04166, **2023**.

[12] Wang Z, Xie Q, Ding Z, et al. Is ChatGPT a good sentiment analyzer? a preliminary study. arXiv: 2304.04339, **2023**.

[13] Rao H, Leung C, Miao C. Can ChatGPT assess human personalities? A general evaluation framework. arXiv: 2303.01248, **2023**.

[14] Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI feedback. arXiv: 2212.08073, **2022**.

[15] Wei J, Wang X, Schuurmans D, et al. Chain of thought prompting elicits reasoning in large language models. arXiv: 2201.11903, **2022**.

[16] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners. arXiv: 2205.11916, **2022**.

[17] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, **2020**: 1877–1901.

[18] Sun B, Zhang Y, He J, et al. A random forest regression method with selected-text feature for depression assessment. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challeng. New York: ACM, **2017**: 61–68.

[19] Wu C, Yin S, Qi W, et al. Visual ChatGPT: Talking, drawing, and editing with visual foundation models. arXiv: 2303.04671, **2023**.

[20] Shen Y, Song K, Tan X, et al. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. arXiv: 2303.17580, **2023**.

CSTR: 32290.14.JUSTC-2023-0088
DOI: 10.52396/JUSTC-2023-0088
JUSTC, **2025**, 55(7): 0705