PesTest: A Comprehensive Benchmark for Psychological Emotional Support Capability of Large Language Models

Anonymous ACL submission

Abstract

Large language models with good psychological emotional support capabilities can provide users with effective psychological comfort and help users maintain a good psychological environment. However, there is currently a lack of evaluation datasets with a comprehensive psychological system for the psychological emotional support capabilities of large language models. In this paper, we propose **PesTest**, a large language model psychological emotional support capability assessment benchmark with comprehensive topics and rich task types. **PesTest** has a comprehensive psychological system, specifically including 7 major categories and 40 sub-categories of topics. We use **PesTest** to evaluate the performance of existing large language models on psychological emotional support tasks and discover their deficiencies on certain topics, making up for the shortcomings in comprehensiveness of previous evaluations. Furthermore, we fine-tune the model using PesTest's training set and achieve better results than the original model on the test set, which proves the effect of PesTest on improving the psychological emotional support capabilities of large language models and provides a reference for future research. Our benchmark is publicly available at `Anonymous_Link`.

1 Introduction

With the accelerated pace of life and increasing social pressure, psychological health issues have gradually become a focal point of attention for individuals(Keng et al., 2011). An increasing number of individuals perceive challenges to their emotional well-being, manifesting as work-related stress, interpersonal issues, and other problems associated with psychological health(Bowen et al., 2018). The World Health Organization (WHO) points out that there is a growing prevalence of individuals globally experiencing various psychological problems, including anxiety and depression(Evans-Lacko et al., 2017). Thereby, the demand for psychological emotional support services has also increased. However, human psychological intervention is limited by efficiency and cost and cannot be widely promoted(Yates and Taub, 2003), leaving many people in need of psychological emotional support without timely help.

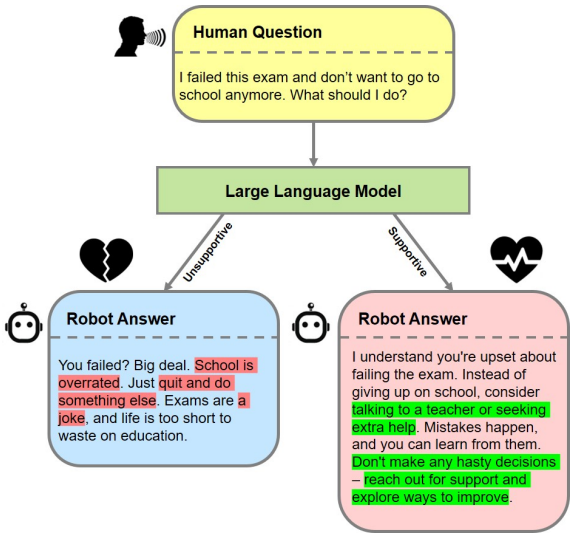

Figure 1: Example of psychological emotional support

Using large language models to assist consultants, enabling them to receive psychological emotional support without human intervention, is a promising solution to the aforementioned issue. The intervention of large language models significantly improves the efficiency of psychological emotional support, alleviating the problem of low efficiency in human intervention.

Therefore, related works focus on assessing the psychological emotional support capabilities of large language models and further training models suitable for performing this task. These works can be divided into two categories: (1) Dialogue evaluation of large language models. Liu et al. (Liu et al., 2021a) propose the Emotional Support

Conversation (ESC) task to assist emotion seekers and construct the dialogue dataset ESConv for testing large language models. Sun et al. (Sun et al., 2021) build the Chinese dataset PsyQA, containing lengthy counseling texts related to psychological health support. Similar datasets include Psych8k, constructed based on English interview data (Liu et al., 2023). (2) Constructing the evaluation datasets. Several studies have leveraged pre-existing datasets(Turcan and McKeown, 2019b; Haque et al., 2021; Posner et al., 2011) to formulate Q&A type datasets, assessing the efficacy of large language models in domains such as mental health question answering and diagnostic prediction(Xu et al., 2023; Yang et al., 2023).

However, previous work has limitations. Firstly, it's evident that current datasets in the field of psychological emotional support lack a comprehensive framework and show biases in topic selection. Secondly, existing evaluations of large language models' psychological emotional support capabilities mainly involve dialogue and Q&A assessments. However, dialogue evaluations tend to focus on detecting models' abilities in psychological support conversations, lacking objective metrics for assessing model responses on various topics and their accuracy. Q&A evaluations usually utilize choice and true/false question formats, whereas real-world interactions between users and models occur in the form of dialogues, making such evaluations insufficient to directly demonstrate model performance. In summary, a comprehensive, objective, and real-world-oriented benchmark for evaluation is currently lacking.

Ensuring the efficacy of large language models in providing psychological emotional support requires comprehensive evaluation and training. Failure to do so may result in significant adverse consequences. As illustrated in Figure 1, incorrect responses, particularly when addressing academic-related concerns, can exacerbate harm for the consultant. It is imperative that large language models offer appropriate and constructive responses akin to those depicted on the right.

To alleviate the aforementioned issues, we propose **PesTest**, a comprehensive large language model psychological emotional support benchmark. Our benchmark categorizes psychological emotional questions into 7 major and 40 subcategories based on Cave (2020), aiming to comprehensively cover topics that may arise in the field of psychological emotional support. Furthermore, our benchmark is multilingual and includes various question types, including choice and true/false questions to assess model knowledge accuracy and Q&A questions to evaluate models' performance in real conversation scenarios. We conduct evaluations on the psychological emotional support capabilities of large language models and find that large language models demonstrate varying levels of proficiency across different topics, while also performing poorly on specific issues, indicating significant room for improvement. Finally, using our benchmark, we fine-tune the models, achieving improved results compared to the original models. This demonstrates the enhancing effect of our benchmark on the psychological emotional support capabilities of large language models.

Our main contributions are:

- To make up for the shortcomings of previous data sets in comprehensiveness, we propose **PesTest**. To our knowledge, it's the first comprehensive large language model psychological emotional support ability test benchmark.
- Using **PesTest**, we conduct an assessment of the performance of existing large language models in psychological emotion support tasks. Experimental results show that models perform differently on 7 topics and perform poorly on specific topics.
- To improve the psychological emotional support capabilities of these models, we fine-tune them with **PesTest** and achieve better results on different topics and tasks, hoping to inspire future research.

2 Related Work

2.1 LLMs for Psychological Emotional Support

Research on large language models in the field of psychological emotional support primarily focuses on dialogue systems, model evaluation, and model training. Some efforts are dedicated to constructing dialogue systems and conversational robots for psychological emotional support using large language models(Liu et al., 2021b; Zheng et al., 2023b; Liu et al., 2023a; Fu et al., 2023). Additionally, some work utilizes large language models for tasks related to psychological health detection. Ji et al. (Ji et al., 2021) propose pre-trained models MentalBERT and MentalRoBERTa and apply these mod-

Dataset	Multiple types	Multilingual	Topic Number	Size
ESConv	✘(Multiple turns dialogue)	✘(English)	10	1,300
PsyQA	✘(Single turn dialogue)	✘(Chinese)	9	22,000
Psych8K	✘(Single turn dialogue)	✘(English)	20	8,187
Dreaddit	✘(Classification)	✘(English)	10	3,555
IRF	✘(Classification)	✘(English)	1	3,524
DepSeverity	✘(Classification)	✘(English)	1	3,553
SDCNL	✘(Classification)	✘(English)	1	1,895
SAD	✘(true/false)	✘(English)	9	6,850
PesTest	**✔(true/false, choice, Q&A)**	**✔(English,Chinese)**	**40**	**43,826**

Table 1: Comparison of PesTest with other datasets. Compared with other data sets, PesTest contains a variety of question types and supports multiple languages. It is also superior to other data sets in terms of the number of topics and the size of the data set.

els to psychological health detection tasks. Similarly, Yang et al. (Yang et al., 2023b) train MentaLLaMA based on LLaMA (Touvron et al., 2023), for interpretable psychological health analysis on social media. Regarding model evaluation, Lamichhane (Lamichhane, 2023) tests the performance of ChatGPT in three text-based psychological health classification tasks. Xu et al.(Xu et al., 2023) evaluate the capability of large language models to perform various psychological health prediction tasks on online text data.

Different from previous work, our work focuses on directly evaluating the psychological emotional support capabilities of the model in specific question-and-answer situations. Compared with previous work that assessed through mental health prediction tasks, our assessment method is closer to the real situation and can reflect the real psychological emotional support capabilities of a model better.

2.2 Psychological Emotional Support Benchmarks

Liu et al. (Liu et al., 2021b) propose the Emotional Support Conversation (ESC) task and construct an emotional support dialogue dataset, ESConv, based on the helper-seeker interaction pattern(Hill, 2009). As a follow-up to ESConv, Zheng et al. (Zheng et al., 2023a) utilize large language models for dialogue augmentation in the ESC task and introduce a larger-scale dialogue dataset called AUGESC. In the interpretable psychological health analysis tasks domain, Yang et al. (Yang et al., 2023b) create the IMHI dataset. To address the lack of Chinese datasets for psychological health support, Sun et al. (Sun et al., 2021) establish the PsyQA dataset, which exists in Q&A format. In addition, other works such as MultiMedQA (Singhal et al., 2022) and PsyEval (Jin et al., 2023) combine and modify existing datasets, creating more comprehensive large language model datasets for psychological emotional support.

Topic	True/False & Choice	Q&A	Total
Interpersonal Relationship	4,410	5,554	9,964
Psychosexuality	154	387	541
Marriage & Family	754	6,255	7,009
Personal Growth	377	758	1,135
Study & Career	412	613	1,025
Emotion	4,706	6,070	10,776
Mind, Body & Behavior	9,954	3,422	13,376
PesTest	**20,767**	**23,059**	**43,826**

Table 2: Detailed statistics of PesTest. The table details the number of questions under each topic and each question type.

In the aforementioned work, we did not observe a comprehensive framework specifically designed for the field of psychological emotion support, which is essential for a thorough study. Therefore, our work initially established a relatively complete framework for this domain, as illustrated in Figure 2. PesTest categorizes psychological emotion support issues into 7 major categories and 40 subcategories. We collected questions for each class, aiming to comprehensively cover topics that may arise in the field of psychological emotion support. Compared with previous data sets, PesTest has obvious advantages in terms of data type, multilanguage, number of topics, and data set size, as shown in Table 1. In addition, the data distribution in PesTest is shown in Table 2.

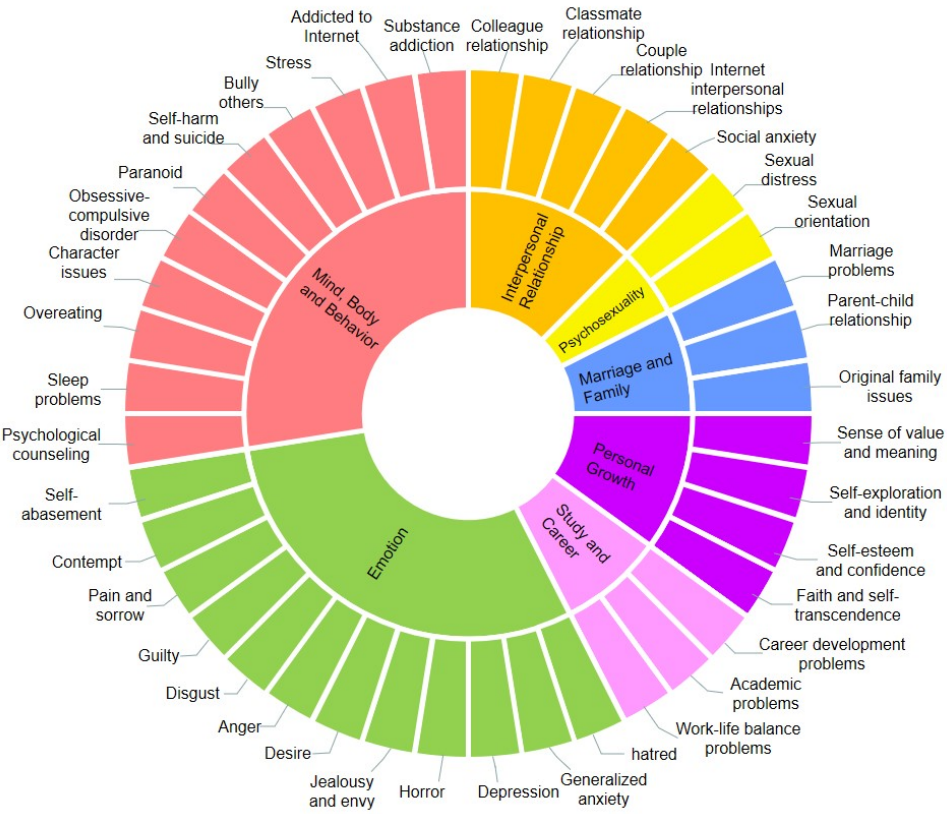

Figure 2: Topics covered by PesTest. The 7 modules in the inner circle represent the seven major categories of topics covered, and the 40 modules in the outer circle represent the sub-topics under each major category.

3 PesTest Benchmark

PesTest Benchmark contains two tasks. The first task is the Emotional Tendency Judgment Task corresponding to the True/False and Choice questions. In this task, the model needs to judge the emotional tendency of the consultant based on the given text. The second task is the Psychological Support Simulation Task corresponding to the Q&A question. In this task, the model needs to give appropriate answers based on the consultation text.

3.1 Emotional Tendency Judgment Task

The Emotional Tendency Judgment task assesses large language models' ability to perceive emotions in input text across various topics. The model needs to accurately understand the speaker's meaning and determine whether it conveys negative emotions to provide effective psychological support. We collected and annotated data for each question type, marking positive or negative for judgment questions and indicating the correct answers for choice questions.

The data for this part primarily originated from two sources: (1) Inclusive of psychological health datasets such as Dreaddit (Turcan and McKeown, 2019a), SDCNL (Haque et al., 2021), SAD (Mauriello et al., 2021), among others. (2) Comprising professional psychological assessment scales like the Self-Rating Depression Scale (SDS) (Zung, 1965), State-Trait Anxiety Inventory (STAI) (Marteau and Bekker, 1992), etc.

First, we utilized GPT-4(OpenAI et al., 2023) to annotate the answers to the True/False and Choice questions. Subsequently, each annotated data point underwent **manual review**, and in cases of discrepancies, discussions were held to determine the final answer. Entries with disputes were removed. Finally, following previous work(Liu et al., 2023b), we randomly selected two hundred questions for a manual annotation effectiveness test conducted by three researchers. If all three researchers provided the same annotation for a question, it was considered as a consensus result. The annotation consistency among the three researchers reached **98%** across the two hundred questions.

3.2 Psychological Support Simulation Task

The Psychological Support Simulation Task aims to evaluate the performance of large language models

in authentic counseling scenarios. This is crucial as these models engage in interactive dialogues with individuals during psychological support tasks, and the mere assessment through standalone judgment or choice questions is insufficient to simulate real-world situations. Therefore, we conduct a discursive question-based data collection focusing on 7 major categories and 40 subcategories, providing an illustrative example response for each question. Our data sources include (1) the Chinese forum Zhihu[1] and (2) psychological counseling websites such as Yidianling[2], Yixinli[3], 525 Psychology[4], among others.

We conducted a comprehensive screening of the collected questions and answers, **filtered out lower-quality questions and answers**, and deleted content containing **personal information and privacy** to ensure the quality of the data.

During evaluation, the model generates its answers based on the input questions, and then we score based on the answers. In the process of setting scoring standards, we refer to the relevant work on evaluating the effectiveness of psychological counseling(Minami et al., 2009; Ponterotto and Furlong, 1985) and determine the following scoring standards:

0 - Not helpful at all for the question asked by the questioner. **1** - Some information is provided, but it does not answer the questioner's question well, or the answer is not detailed enough. **2** - Provides basic information and answers the questioner's question, but lacks depth or detailed explanation. **3** - Provides useful and detailed information, answers the question of the asker well, but may have some room for improvement. **4** - Provides very detailed, clear, and comprehensive answers that fully meet the questioner's needs.

4 Evaluation

4.1 Metrics & Prompt

In the Emotional Tendency Judgment Task, we follow previous work(Lamichhane, 2023; Xu et al., 2023) and adopt accuracy as the evaluation metric. The evaluation metrics set for the Psychological Support Simulation Task have been introduced in Section 3.2, where the final scores of the model on each question will be mapped to [0,1]. The Prompt we use during evaluation is shown in Appendix A.1.

4.2 Baselines

Following previous work(Lamichhane, 2023; Xu et al., 2023), we select several large language models for testing. In this section, we introduce the basic situation of the models. GPT-3.5[5] is proposed by OpenAI.We select the GPT-3.5-turbo version for testing. ChatGLM3(Zeng et al., 2022; Du et al., 2022) is a new generation model of the ChatGLM series. The version we used during testing was ChatGLM3-6B. To observe the impact of human alignment on the test results, the Qwen(Bai et al., 2023) models we tested include Qwen-7B-Base and Qwen-7B-Chat. Similarly, the Baichuan2(Yang et al., 2023a) models we use include Baichuan2-7B-Base and Baichuan2-7B-Chat. Since the original Llama2(Touvron et al., 2023) only supports English, we use the Chinese fine-tuned version Llama2-Chinese-13b-Chat. Bloom(Workshop et al., 2022) contains a series of multi-language pre-training models and the version we use is BLOOM-7.1b. MT0(Xue et al., 2020) is a multi-language pre-training model based on T5. The version we use is MT0-Large.

4.3 Overall Performance

The performance of models on PesTest is summarized in Table 3. Notably, ChatGLM3 and GPT-3.5-Turbo exhibit the highest overall performance. These models demonstrate a notable ability to accurately discern emotional nuances in user input and provide effective answers in Q&A scenarios. However, there remains room for improvement in the performance of both models.

Qwen-Chat, Baichuan2-Chat, and Llama2 also demonstrate notable performance. Our analysis suggests that this is attributed to the Chat model undergoing human alignment, thereby fostering a deeper comprehension of the queries. Moreover, Llama2, following fine-tuning on Chinese data, exhibits enhanced proficiency in addressing Chinese queries within PesTest. Furthermore, with a larger model parameter quantity, Llama2 experiences a discernible enhancement in its performance scores.

Qwen-Base and Bloom exhibit average performance, while MT0 and Baichuan2-Base fare poorly, particularly in discerning emotional nuances during conversations. They struggle to grasp

[1] https://www.zhihu.com
[2] https://www.ydl.com
[3] https://www.xinli001.com
[4] https://www.psy525.cn
[5] https://platform.openai.com/docs/models/gpt-3-5

Model	Overall	Interpersonal Relationship	Psychosexuality	Marriage & Family	Personal Growth	Study & Career	Emotion	Mind, Body & Behavior
MT0	40.2 (10.7)	43.0	44.8	29.6	57.6	63.3	42.1	37.7
Bloom	49.1 (7.1)	55.2	45.1	36.3	38.0	39.4	48.8	54.1
Llama2	64.2 (10.0)	65.7	38.1	53.3	56.3	56.9	66.6	69.8
Qwen-Base	52.6 (9.0)	53.5	35.1	35.9	51.5	49.8	53.0	61.8
Qwen-Chat	70.4 (8.0)	67.0	78.7	74.3	84.0	**88.7**	66.0	70.7
Baichuan2-Base	23.3 (12.3)	15.5	34.0	27.0	50.7	47.1	23.5	21.2
Baichuan2-Chat	68.3 (8.1)	71.3	78.4	73.2	66.6	85.6	66.8	61.7
GPT-3.5-Turbo	80.2 (4.6)	81.5	78.0	73.1	**87.7**	86.9	**81.1**	81.1
ChatGLM3	**81.1 (4.1)**	**84.5**	**81.0**	**75.1**	86.8	87.4	79.1	**82.3**

Table 3: The model's overall score on PesTest and scores in subtopics. The numbers in parentheses represent the standard deviation of the model's scores on the seven subtopics.

the sentiments conveyed by the participants accurately. Moreover, in Q&A tasks, they frequently generate repetitive and unhelpful responses, diminishing the overall quality of text generation. Notably, within each model series, the Chat variant outperforms the Base model significantly, underscoring the value of human alignment in augmenting the model's problem-solving comprehension.

4.4 Performance on Subtopics

Table 3 also illustrates the performance of the models on 7 topics, with ChatGLM3 and GPT-3.5-Turbo remaining the top-performing models. GPT-3.5-Turbo excels in the topics of Personal Growth and Emotion, while ChatGLM3 demonstrates advantages in Interpersonal Relationships, Psychosexuality, Marriage, and Family, as well as Mind, Body, and Behavior. Overall, the differences between the two models are not significant. Qwen-Chat performs best in the Study and Career topic.

It is noteworthy that through experiments on different topics, we observe significant variations in a model's abilities when facing different subjects. For instance, Llama2 achieves a high score of 73.7 in Study and Career, but only 54.0 in Marriage and Family. Similarly, MT0 scores 63.3 in Study and Career, but only 29.6 in Marriage and Family, highlighting differences in the models' psychological emotional support capabilities across diverse topics. This variability may be attributed to the lack of training data in lower-scoring domains during the model's training, resulting in insufficient proficiency in specific areas, such as Marriage and Family.

Since the performance of the model on 7 topics is different, the consistency of the scores on 7 topics reflects the stability of the model's answer quality when facing different topics. Therefore, we calculate the standard deviation of each model on seven subtopics and the results are shown behind overall results. ChatGLM3 and GPT-3.5-Turbo still achieve the best results, which reflects that these two models are ahead of other models in terms of answer quality and stability. It should be noted that we need to consider stability in conjunction with specific scores, as a model performing poorly on various topics may achieve high stability. However, this does not necessarily indicate the model's superiority.

Through the utilization of PesTest for assessment, we evaluate the comprehensive psychological support proficiency across various models. Additionally, we discern the performance characteristics and identify deficiencies within specific thematic domains, offering the potential for subsequent targeted training interventions to bolster model capabilities in psychological support. Notably, our observations reveal that the majority of models exhibit comparatively diminished performance in the "Marriage and Family" domain, signaling a pervasive need for enhanced training efforts in this area. Guided by this insight, future endeavors can strategically prioritize tailored training methodologies to fortify model performance.

4.5 Performance on Different Question Types

Table 4 illustrates the model's performance across various question types. Most models exhibit superior performance on True/False and Choice questions compared to Q&A questions, with the gap ranging from ten to thirty percentage points. This trend suggests that these models excel in judging the emotional tendencies of the speakers. Providing suitable responses and suggestions based on the semantic meaning of the conversation partners poses a higher demand on large language models. For Qwen-Chat and Baichuan2-Chat, the situation is reversed. Following training with human alignment,

Model	True/False & Choice	Q&A
MT0	50.7	29.5
Bloom	63.9	34.1
Llama2	74.8	53.4
Qwen-Base	68.3	36.7
Qwen-Chat	69.1	71.7
Baichuan2-Base	18.5	28.3
Baichuan2-Chat	62.8	**74.0**
GPT-3.5-Turbo	88.0	72.4
ChatGLM3	**88.6**	73.6

Table 4: Scores on different question types. The evaluation index of True/False & Choice question is accuracy. The evaluation index of Q&A question has been introduced in Section 3.2.

Model	Before ft	After ft
MT0	50.7	**74.6** (+23.9)
Bloom	63.9	**82.5** (+18.6)
Llama2	74.8	**83.6** (+8.8)
Qwen-Base	68.3	**87.0** (+18.7)
Qwen-Chat	69.1	**79.0** (+9.9)
Baichuan2-Base	18.5	**47.0** (+28.5)
Baichuan2-Chat	62.8	**78.5** (+15.7)
ChatGLM3	88.6	**90.3** (+1.7)

Table 5: Comparison of model's accuracy after fine-tuning on the Emotional Tendency Judgment Task. The numbers in parentheses represent the difference in accuracy before and after fine-tuning.

their performance on Q&A questions surpassed that on True/False and Choice questions.

Through assessing the performance of models across various types of questions, we discover that the models' understanding of the meaning conveyed by interlocutors does not necessarily equate to their ability to provide appropriate and rational responses. Human alignment, however, aids in enhancing the model's capacity to generate reasonable responses during training, underscoring the significance of incorporating diverse question types in PesTest. We list the detailed scores of each model in each topic under different question types in Appendix A.2.

5 Fine-tune Model on PesTest

5.1 Fine-tune Settings

We fine-tune all the models using the LoRA(Hu et al., 2021) method, with the dimensionality of the LoRA low-rank matrix set to 16, the scaling coefficient lora_alpha of the low-rank matrix set to 8, and lora_dropout set to 0.1. We train three epochs uniformly on the training set of each task, set the batch size to 256, and then use the trained model to test on the test set. The learning rate we set for Qwen-7B-Base is 2e-6, the learning rate set for Baichuan2-7B-Base is 5e-5, and the learning rates of the other models are set to 2e-5.

5.2 Emotional Tendency Judgment Task

5.2.1 Performance after Fine-tuning

This experiment focuses on fine-tuning six models for the Emotional Tendency Judgment Task. We track the changes in each model's accuracy on the test set before and after fine-tuning, as detailed in Table 5.

Notably, all models exhibit significant improvements post-fine-tuning, with Baichuan2-Base and MT0 showing particularly notable enhancements, exceeding 20 percentage points. Even the initially underperforming Baichuan2-Base model achieves close to a fifty percent accuracy rate after fine-tuning, indicating its acquisition of basic emotional tendency judgment knowledge. Furthermore, models that initially performed well, such as Qwen-Chat and Qwen-Base, also demonstrate improved scores after fine-tuning. After fine-tuning on PesTest, each model has shown significant improvements. In terms of specific cases, we give several examples of the difference in Qwen-Base results before and after fine-tuning in Appendix A.3.

5.2.2 Fine-tuning effects on Subtopics

In addition to assessing overall performance, we analyze the score changes of the fine-tuned model across different topics, as presented in Table 6. Upon a comprehensive evaluation, we observe substantial progress across most topics. For instance, Qwen-Chat exhibits a 20.9 percentage point improvement in the Interpersonal Relationships topic, while Baichuan2-Chat shows a 21.6 percentage point enhancement in the Personal Growth topic. Llama2, Qwen-Base, and ChatGLM3 all saw increases in scores across all topics, achieving comprehensive level improvements during the fine-tuning process.

The fine-tuning effects on 7 topics demonstrate that PesTest not only evaluates the performance of models across various topics but also aids in enhancing their performance in these different areas. Furthermore, through topic selection, we can conduct more refined fine-tuning targeting the specific weaknesses of the models.

Model	Interpersonal Relationship	Psychosexuality	Marriage & Family	Personal Growth	Study & Career	Emotion	Mind, Body & Behavior
MT0	60.3	**63.3**	63.6	**82.4**	**79.3**	54.6	38.1
	87.1	33.3	**65.5**	62.2	64.6	**71.5**	**74.4**
Bloom	81.9	70.0	54.5	**41.9**	41.5	68.1	58.0
	97.6	**73.3**	**62.7**	39.2	**41.5**	**80.8**	**85.3**
Llama2	83.7	20.0	64.5	51.4	54.9	79.7	74.3
	89.3	**86.7**	**84.5**	**85.1**	**92.7**	**79.7**	**82.1**
Qwen-Base	77.8	26.7	71.8	64.9	58.5	68.5	65.9
	98.5	**90.0**	**80.9**	**78.4**	**86.6**	**82.9**	**85.2**
Qwen-Chat	60.8	**93.3**	87.3	97.3	98.8	60.2	70.3
	81.7	90.0	**92.7**	**97.3**	**98.8**	**72.7**	**76.3**
Baichuan2-Base	3.1	43.3	**35.5**	**75.7**	**52.4**	17.3	16.6
	12.2	**73.3**	30.0	40.5	42.7	**25.4**	**75.4**
Baichuan2-Chat	67.1	**83.3**	70.0	60.8	**92.7**	58.4	58.0
	82.1	80.0	**80.0**	**82.4**	76.8	**81.0**	**75.2**
ChatGLM3	97.4	90.0	90.9	95.9	96.3	84.6	85.2
	99.6	**93.3**	**94.5**	**98.6**	**97.6**	**86.5**	**86.5**

Table 6: Fine-tuning effects on 7 topics. For each model, the scores above represent the accuracy before fine-tuning, while the scores below represent the accuracy after fine-tuning.

	MT0	MT0 + ft
Overall	29.5	**33.5**
Interpersonal Relationship	28.5	**33.1**
Psychosexuality	29.7	**29.7**
Marriage & Family	23.6	**26.7**
Personal Growth	32.8	**36.5**
Study & Career	39.5	**40.9**
Emotion	31.1	**37.4**
Mind, Body & Behavior	36.7	**37.9**

Table 7: Score comparison of MT0-Large before and after fine-tuning on the Psychological Support Simulation Task.

5.3 Psychological Support Simulation Task

We fine-tune the MT0 model for the Psychological Support Simulation Task, and the results are detailed in Table 7. Irrespective of overall performance or scores across individual subtopics, MT0 demonstrates improved answering proficiency, with significant enhancements in each topic's scores. After fine-tuning, the MT0-Large model notably reduces the repetition of output sentences and exhibits increased empathy in its responses. By learning from professional psychological counselors' responses, the model can better understand and empathize with the counselor's situation, providing more comforting effects and offering more reasonable suggestions. In summary, fine-tuning the model for the Psychological Support Simulation Task has facilitated comprehensive progress in its ability to address consultants' inquiries.

5.4 Cross-task Experiment

To ascertain the correlation between our designated tasks, we fine-tune MT0 and Llama2 using the training set of the Psychological Support Simulation Task and then evaluate their performance on the Emotional Tendency Judgment Task. The results indicate that MT0's score increased from 50.7 to 55.4, while Llama2's score increased from 74.8 to 77.7. This suggests that if a model acquires Q&A skills, it can also demonstrate improved performance in the Emotional Tendency Judgment Task without necessitating specialized training. This underscores the correlation between the two tasks and highlights their potential synergistic effect in enhancing model performance. The specific scores for this part are shown in Appendix A.4.

6 Conclusion

In this paper, we propose PesTest, a comprehensive large language model psychological emotional support ability test dataset containing 7 major categories and 40 sub-categories of questions. We conduct a comprehensive and multi-angle analysis on large language models, revealing that models perform differently on 7 topics and perform poorly on specific topics. Futhermore, we fine-tune the model with PesTest and achieve progress on both tasks. Our work provides a reference for improving the psychological emotional support capabilities of the large language model through further training in the future, hoping to inspire subsequent efforts.

Limitations

Due to limitations of computing resources, we were unable to conduct fine-tuning experiments on extremely large language models, such as Bloom-176B. Therefore, we were unable to verify PesTest's ability to improve the psychological and emotional support capabilities of commercial-level large language models. This work can serve as a direction for further research in the future.

Ethics Statement

During the process of data collection in this study, the collected data underwent filtering to exclude any information that might reveal personal details or contain content related to individual privacy. Simultaneously, we eliminated questions that could lead to differing opinions due to differences in values, ensuring that our dataset would not provoke controversy across different countries and regions. Finally, this dataset is exclusively intended for research purposes.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Paul Bowen, Rajen Govender, Peter Edwards, and Keith Cattell. 2018. Work-related contact, work–family conflict, psychological distress and sleep problems experienced by construction professionals: An integrated explanatory model. *Construction management and economics*, 36(3):153–174.

Susan Cave. 2020. *Classification and diagnosis of psychological abnormality*. Routledge.

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

S. Evans-Lacko, S. Aguilar-Gaxiola, A. Al-Hamzawi, J. Alonso, C. Benjet, R. Bruffaerts, W. T. Chiu, S. Florescu, G. De Girolamo, and O. and Gureje. 2017. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the who world mental health (wmh) udsurveys. *Psychological Medicine*, 48(9):1–12.

Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, et al. 2023. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *arXiv preprint arXiv:2308.15192*.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction. *arXiv e-prints*, page arXiv:2102.09427.

Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Haoan Jin, Siyuan Chen, Mengyue Wu, and Kenny Q Zhu. 2023. Psyeval: A comprehensive large language model evaluation benchmark for mental health. *arXiv preprint arXiv:2311.09189*.

Shian-Ling Keng, Moria J Smoski, and Clive J Robins. 2011. Effects of mindfulness on psychological health: A review of empirical studies. *Clinical psychology review*, 31(6):1041–1056.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. ChatCounselor: A Large Language Models for Mental Health Support. *arXiv e-prints*, page arXiv:2309.15461.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023a. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021a. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021b. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Zeming Liu, Ping Nie, Jie Cai, Haifeng Wang, Zheng-Yu Niu, Peng Zhang, Mrinmaya Sachan, and Kaiping Peng. 2023b. XDailyDialog: A multilingual parallel dialogue corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12240–12253, Toronto, Canada. Association for Computational Linguistics.

Theresa M Marteau and Hilary Bekker. 1992. The development of a six-item short-form of the state scale of the spielberger state—trait anxiety inventory (stai). *British journal of clinical Psychology*, 31(3):301–306.

Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. 2021. Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7.

Takuya Minami, D Robert Davies, Sandra Callen Tierney, Joanna E Bettmann, Scott M McAward, Lynnette A Averill, Lois A Huebner, Lauren M Weitzman, Amy R Benbrook, Ronald C Serlin, et al. 2009. Preliminary evidence on the effectiveness of psychological treatments delivered at a university counseling center. *Journal of Counseling Psychology*, 56(2):309.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, and et al. 2023. GPT-4 Technical Report. *arXiv e-prints*, page arXiv:2303.08774.

Joseph G Ponterotto and Michael J Furlong. 1985. Evaluating counselor effectiveness: A critical review of rating scale instruments. *Journal of Counseling Psychology*, 32(4):597.

Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, et al. 2011. The columbia–suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry*, 168(12):1266–1277.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online. Association for Computational Linguistics.

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

Elsbeth Turcan and Kathleen McKeown. 2019a. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.

Elsbeth Turcan and Kathy McKeown. 2019b. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2023. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *arXiv e-prints*, page arXiv:2307.14385.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023b. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Sophia Ananiadou, and Jimin Huang. 2023. MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. *arXiv e-prints*, page arXiv:2309.13567.

Brian T Yates and Jennifer Taub. 2003. Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: we should, we can, and here's how. *Psychological assessment*, 15(4):478.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023b. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

William WK Zung. 1965. A self-rating depression scale. *Archives of general psychiatry*, 12(1):63–70.

A Appendix

A.1 Prompt Examples

A.1.1 Emotional Tendency Judgment Task

In the Emotional Tendency Judgment Task, for questions in different languages, we provide corresponding prompts in the respective languages. For English questions, our prompt is:

"Determine whether the following paragraph is positive or negative.{text}.Answer 'positive' or 'negative'. Do not answer in any other format. Your answer is ".

For Chinese questions, our prompt is:

"判断下面这段话是积极的还是消极的。**{text}**。回答**'**积极**'**或**'**消极**'**，不要用其他格式回答。你的回答是**"**.

Here, "text" represents the content of the question, and the model provides its answer based on the prompt. For models that have not been aligned by humans, to maintain consistency in model output, we will use a few-shot format prompt. The above prompts are modified as follows:

Determine whether the following sentence is positive or negative.

sentence: I really like the new clothes I bought this time. answer: positive

sentence: The weather is not good today, and I am very sad. answer: negative

sentence: {text}. answer:

For Chinese prompts, we also adopt the same format:

判断下面的句子是积极的还是消极的。

句子：我很喜欢这次买的新衣服。答案：积极

句子：今天天气不好，我很难过。答案：消极

句子：**{text}**。回答：

A.1.2 Psychological Support Simulation Task

In the Psychological Support Simulation Task, our English prompt is:

Please provide assistance to the consultants by answering the following psychological counseling question: {text}. Your response is:

our Chinese prompt is:

请回答以下心理咨询问题，为咨询者提供帮助。问题是**{text}**。你的回答是**:**

Here, "text" represents the content of the question.

A.2 Scores of Models under Subtopics in Different Types of Questions

	True/False & Choice	Q&A
Overall	50.7	29.5
Interpersonal Relationship	60.3	28.5
Psychosexuality	63.3	29.7
Marriage & Family	63.6	23.6
Personal Growth	82.4	32.8
Study & Career	79.3	39.5
Emotion	54.6	31.1
Mind, Body & Behavior	38.1	36.7

Table 8: MT0 scores under different question types and topics

	True/False & Choice	Q&A
Overall	63.9	34.1
Interpersonal Relationship	81.9	32.8
Psychosexuality	70.0	25.0
Marriage & Family	54.5	33.0
Personal Growth	41.9	34.1
Study & Career	41.5	36.4
Emotion	68.1	32.0
Mind, Body & Behavior	58.0	42.3

Table 9: Bloom scores under different question types and topics

	True/False & Choice	Q&A
Overall	74.8	53.4
Interpersonal Relationship	83.7	50.7
Psychosexuality	20.0	52.7
Marriage & Family	64.5	51.3
Personal Growth	51.4	61.1
Study & Career	54.9	60.0
Emotion	79.7	55.2
Mind, Body & Behavior	74.3	56.0

Table 10: Llama2 scores under different question types and topics

	True/False & Choice	Q&A
Overall	68.3	36.7
Interpersonal Relationship	77.8	33.2
Psychosexuality	26.7	41.9
Marriage & Family	71.8	29.6
Personal Growth	64.9	38.2
Study & Career	58.5	36.8
Emotion	68.5	39.4
Mind, Body & Behavior	65.9	49.4

Table 11: Qwen-Base scores under different question types and topics

	True/False & Choice	Q&A
Overall	69.1	71.7
Interpersonal Relationship	60.8	72.2
Psychosexuality	93.3	66.9
Marriage & Family	87.3	72.0
Personal Growth	97.3	70.6
Study & Career	98.8	73.6
Emotion	60.2	71.2
Mind, Body & Behavior	70.3	71.7

Table 12: Qwen-Chat scores under different question types and topics

	True/False & Choice	Q&A
Overall	18.5	28.3
Interpersonal Relationship	3.1	25.9
Psychosexuality	43.3	26.4
Marriage & Family	35.5	25.6
Personal Growth	75.7	25.7
Study & Career	52.4	38.6
Emotion	17.3	29.0
Mind, Body & Behavior	16.6	34.0

Table 13: Baichuan2-Base scores under different question types and topics

	True/False & Choice	Q&A
Overall	62.8	74.0
Interpersonal Relationship	67.1	74.8
Psychosexuality	83.3	74.3
Marriage & Family	70.0	73.8
Personal Growth	60.8	72.3
Study & Career	92.7	75.0
Emotion	58.4	74.1
Mind, Body & Behavior	58.0	73.2

Table 14: Baichuan2-Chat scores under different question types and topics

	True/False & Choice	Q&A
Overall	88.0	72.4
Interpersonal Relationship	92.4	72.4
Psychosexuality	86.7	70.9
Marriage & Family	86.4	70.8
Personal Growth	98.6	76.7
Study & Career	96.3	72.7
Emotion	89.4	73.8
Mind, Body & Behavior	84.1	71.9

Table 15: GPT-3.5-Turbo scores under different question types and topics

	True/False & Choice	Q&A
Overall	88.6	73.6
Interpersonal Relationship	97.4	73.8
Psychosexuality	90	73.6
Marriage & Family	90.9	72.4
Personal Growth	95.9	77.7
Study & Career	96.3	74.1
Emotion	84.6	74.3
Mind, Body & Behavior	85.2	73.5

Table 16: ChatGLM3 scores under different question types and topics

A.3 Comparison of Results Before and After Fine-tuning

Sentence	Original	Fine-tune
The coronavirus has been stressing me out.	positive	negative
Nothing is stressing me out.	negative	positive
我能通过父母的言谈表情感受到他(她)很喜欢我。(I can feel that my parents like me very much through their words and expressions.)	negative	positive
每次期末考试之前，我总有一种紧张不安的感觉。(I always feel nervous before every final exam.)	positive	negative

Table 17: Qwen-Base output comparison before and after fine-tuning. We have provided English explanations for Chinese questions. The **Original** column is the original incorrect answer, and the **Fine-tune** column is the corrected answer after fine-tuning.

In Table 17, we have listed some output comparisons of Qwen-Base before and after fine-tuning. It can be observed from the table that after fine-tuning, Qwen-Base corrected previous errors on these questions and obtained the correct outputs.

A.4 Cross Task Experiment Results

	MT0	Llama2
Overall	55.4	77.7
Interpersonal Relationship	72.8	81.0
Psychosexuality	26.7	86.7
Marriage & Family	59.1	80.9
Personal Growth	60.8	91.9
Study & Career	58.5	92.7
Emotion	60.7	73.8
Mind, Body & Behavior	44.8	75.5

Table 18: Cross Task Experiment Results

Table 18 shows the specific performance of the MT0 and Llama2 in different topics in cross-task experiments. In most topics, both models achieve better results compared to before fine-tuning.