CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering

Hongbin Na
Australian Artificial Intelligence Institute
University of Technology Sydney

Introduction

Challenges: Existing language models for mental health support lack access to high-quality, specialized data. Responses from current datasets, such as those from online psychological platforms, can sometimes be ineffective or inappropriate.

PsyQA Dataset

Online Psychological Platforms

User

啊,马上就要期末考试了，还要面临教资考试，压力有一点点大。
Ah, I've got finals coming up and I've got teaching interviews to face, so I'm a little stressed out.

Human

你好，加油哦，未来同行我送你一句我教师实习期师父送我的话：教师要有坚定强大的内心！...
Hello, cheers oh future peers I send you a word of advice from teacher trainee master: Teachers need to be firm and strong inside!...

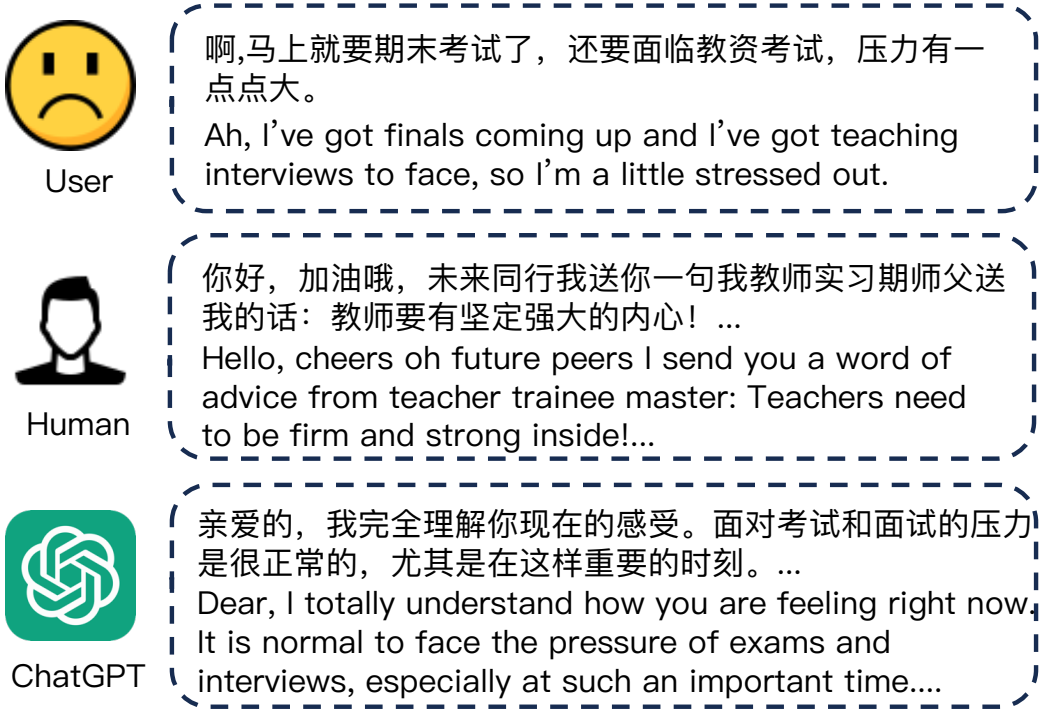
ChatGPT

亲爱的，我完全理解你现在的感受。面对考试和面试的压力是很正常的，尤其是在这样重要的时刻。...
Dear, I totally understand how you are feeling right now. It is normal to face the pressure of exams and interviews, especially at such an important time....

Introduction

Opportunity: Integrating Cognitive Behavioral Therapy (CBT) principles into language models offers a chance to improve the quality of responses. CBT's structured and evidence-based approach makes it highly suitable for this purpose.

Research Goal: We aimed to create a refined dataset tailored for CBT, which was used to train a new large language model, specifically designed to adhere more closely to CBT principles.



Methodology – Generation of CBT Responses

1. **Validation and Empathy**: Show understanding and sympathy for the patient's feelings or issues, creating a sense of safety.
2. **Identify Key Thought or Belief**: Through the problem description, identify potential cognitive distortions or core beliefs.
3. **Pose Challenge or Reflection**: Raise open-ended questions, encouraging the patient to reconsider or reflect on their initial thoughts or beliefs.
4. **Provide Strategy or Insight**: Offer practical strategies or insights to help them deal with the current situation.
5. **Encouragement and Foresight**: Motivate the individual to employ the suggested strategy, underscoring that this is merely an initial step and additional support may be warranted

CBT Prompt

请你基于下述的问题及其描述，给出一条专业的，富有同情心和具有助益性的回复。确保你的回复在保持以下认知行为疗法回答结构的基础上，尤其是识别关键思维或信念部分，流畅地将各部分内容相互连接：

1. 验证和共情：对患者的情感或问题表示理解和同情，创建安全感。
2. 识别关键思维或信念：通过问题描述，找出可能的认知扭曲或核心信仰。
3. 提出挑战或反思：提出开放性问题，鼓励患者重新考虑或反思其初始思维或信仰。
4. 提供策略或见解：提供实用策略或见解，以帮助他们处理当前情况。
5. 鼓励与前瞻：鼓励患者使用策略，强调这只是开始，并可能需要进一步的支持。

问题：[问题]
描述：[描述]
回答：

Based on the following question and its description, please provide a professional, compassionate, and helpful response. Ensure your response adheres to the structure of Cognitive Behavioral Therapy (CBT) responses, especially in identifying the key thought or belief, and seamlessly integrates each part:

1. Validation and Empathy: Show understanding and sympathy for the patient's feelings or issues, creating a sense of safety.
2. Identify Key Thought or Belief: Through the problem description, identify potential cognitive distortions or core beliefs.
3. Pose Challenge or Reflection: Raise open-ended questions, encouraging the patient to reconsider or reflect on their initial thoughts or beliefs.
4.Provide Strategy or Insight: Offer practical strategies or insights to help them deal with the current situation.
5. Encouragement and Foresight: Encourage the patient to use the strategy, emphasizing that this is just the beginning and further support may be needed.

Question: [Question]
Description: [Description]
Response:

Methodology – Generation of CBT Responses

Criteria	Statistics
No. of Question & Description	22327
No. of CBT Response	22327
Characters Per Question	21.6
Characters Per Description	168.9
Characters Per CBT Response	522.8
Percentage of Cognitive Distortions	54.4%

Table 1: CBT QA dataset statistical analysis.

	Accuracy	Recall	F1 score
Quality	0.69	0.93	0.65

Table 3: Cognitive distortion of recognition quality.

Cognitive Distortion Type	Interpretation	Samples
All-or-Nothing Thinking	The tendency to see things as extremes, either total success or total failure, ignoring the possibilities in between.	7115
Overgeneralization	It makes unreasonable inferences about the general situation based on limited experience or a single event, usually based on negative experience.	7782
Emotional Reasoning	It manifests itself by judging things based on one's emotional state without relying on objective evidence or logical reasoning.	742
Catastrophizing	It involves exaggerating the seriousness of errors or problems, seeing them as great disasters, and usually ignoring the reasonableness of the actual situation.	349
Mind Reading	It manifests itself in the false assumption that one knows what others are thinking or feeling, without relying on clear communication or evidence, and usually leads to misunderstanding and conflict.	345
Others	Contains Fortune Telling, Filtering, Attribution Error, and other types of cognitive distortions.	2094

Table 2: Cognitive distortion statistical analysis on CBT QA dataset.

Methodology – CBT-LLM

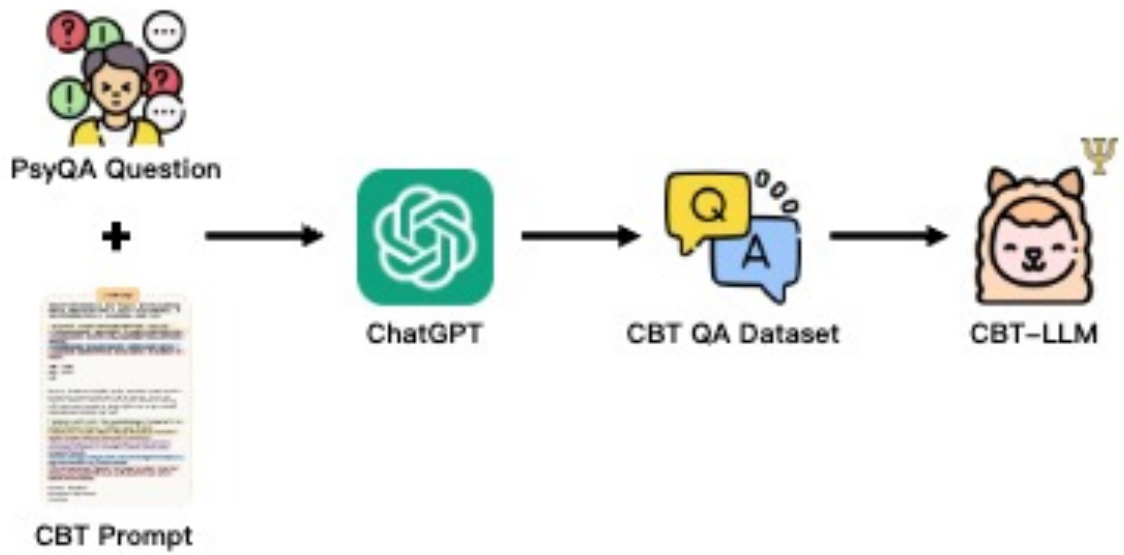


Figure 2: An overview of training CBT-LLM. It first utilizes PsyQA Questions and CBT Prompt to generate CBT answers, and then fine-tuning CBT-LLM.

Experiment

Models
- LLaMA-Chinese-7B
- Alpaca-Chinese-7B
- Qwen-7B
- Baichuan-7B

Automatic Evaluation
- BLEU
- METEOR
- CHRF
- BLEURT
- BERTSCORE

CBT-LLM Backbone	BLEU	METEOR	CHRF	BLEURT	BERTSCORE
LLama-Chinese-7B	0.2412	0.3758	0.2167	0.5091	0.7793
Alpaca-Chinese-7B	0.2607	0.3991	0.2596	0.5216	**0.7849**
Qwen-7B	0.2361	0.3726	0.2939	0.5096	0.7802
Baichuan-7B	**0.2648**	**0.4031**	**0.3839**	**0.5247**	0.7841

Table 4: Automatic evaluation results on CBT QA dataset.

Experiment

Human Evaluation Setup:

1. **Objective:** Test the model's effectiveness in psychotherapy.
2. **Method:** Evaluation of 100 randomly selected responses.
3. **Team:** Four senior psychology students and one experienced therapist.

Evaluation Criteria:

1. **Relevance (0-2):** How well the response matches the question.
2. **CBT Adherence (0-2):** How closely the response follows CBT principles.
3. **Helpfulness (0-2):** The practical value of the response in therapy.

CBT-LLM Backbone	Rele.	Stru.	Help.
Alpaca-Chinese-7B	1.732	1.508	1.408
Baichuan-7B	**1.734**	**1.644**	**1.432**

Table 5: Human evaluation by professional raters for relevance measure (Rel.), CBT structure measure (Stru.) and helpfulness measure (Help.).

Limitations

Cognitive Distortion Identification:

- Lack of specific markers to accurately identify cognitive distortions.
- Potential impact on response accuracy and effectiveness.

User Experience Concerns:

- Overwhelming users by encapsulating the entire CBT process into a single response.
- User feedback indicates potential pressure when faced with multiple simultaneous questions.

Conclusions and Future Work

Achievements:

- Integrated large language models with CBT, creating the CBT-LLM.
- Model delivers structured and relevant responses effectively.

Next Steps:

- Expand methods to include ACT and DBT.
- Develop multi-turn dialogues to mimic real therapy sessions.

Conclusions and Future Work

Thank you for listening