

# HealMe: Harnessing Cognitive Reframing in Large Language Models for Psychotherapy

Anonymous ACL submission

## Abstract

Large Language Models (LLMs) play a vital role in psychotherapy by adeptly handling the crucial task of cognitive reframing, while also overcoming challenges such as shame, distrust, therapist skill variability, and resource scarcity. Previous LLMs in cognitive reframing mainly converted negative emotions to positive ones, but these approaches have limited efficacy, often not promoting clients' self-discovery of alternative perspectives. In this paper, we unveil the Helping and Empowering through Adaptive Language in Mental Enhancement (HealMe) model, a novel cognitive reframing therapy method that effectively addresses deep-rooted negative thoughts and fosters rational, balanced perspectives. Diverging from traditional LLM methods, HealMe employs empathetic dialogue based on psychotherapeutic frameworks. It systematically guides clients through distinguishing circumstances from feelings, brainstorming alternative viewpoints, and developing empathetic, actionable suggestions. Moreover, we adopt the first comprehensive and expertly crafted psychological evaluation metrics, specifically designed to rigorously assess the performance of cognitive reframing, in both AI-simulated dialogues and real-world therapeutic conversations. Experimental results show that our model outperforms others in terms of empathy, guidance, and logical coherence, demonstrating its effectiveness and potential positive impact in the field of psychotherapy.

## 1 Introduction

Cognitive reframing (Carli, 1999), a key part of cognitive behavioural therapy, helps individuals detach from their thoughts and situations, effectively addressing issues from mild negative thinking to severe depression and anxiety (Robson Jr and Troutman-Jordan, 2014; Vernooij-Dassen et al., 2011). Due to the extensive dialogue and significant empathy required in psychotherapy, Large Language Models (LLMs) are in high demand

whether as an adjunct to human-based mental health treatment or as a standalone therapeutic tool. LLMs help overcome obstacles such as shame or distrust often associated with traditional therapy methods (Sickel et al., 2014). Additionally, they address issues like the limited availability of psychotherapeutic resources and the variability in therapists' skill levels.

Current LLM-based cognitive reframing methods (Ziems et al., 2022; Maddela et al., 2023; Sharma et al., 2023b) focus on integrating cognitive reframing techniques with LLMs to transform negative thoughts into neutral or positive ones. However, effective reframing alone has limited influence on clients seeking psychotherapy. This is because clients might not spontaneously discover alternative perspectives and could perceive the reframing as preaching or imposition rather than self-realization.

Therefore, although LLMs demonstrate considerable promise, they face key challenges. Firstly, the LLM-based methods in cognitive reframing merely rewrite thoughts instead of providing therapy, which has limited effectiveness in improving clients' negative emotions. Secondly, LLMs struggle to consistently maintain empathy in psychotherapy dialogues. Lastly, in psychotherapy, it is crucial for LLMs to guide humans, a significant challenge to their directive capabilities, despite their proficiency in generating suggestions.

To tackle these challenges, we propose a specialized model Helping and Empowering through Adaptive Language in Mental Enhancement (HealMe), for cognitive reframing therapy. This model leverages dialogue data imbued with empathy and guidance for instruction tuning, ensuring empathetic and directive responses. Grounded in professional psychological literature (Robson Jr and Troutman-Jordan, 2014), we distil and organize a structured cognitive reframing therapy process, effectively emulating a complete psychotherapeutic

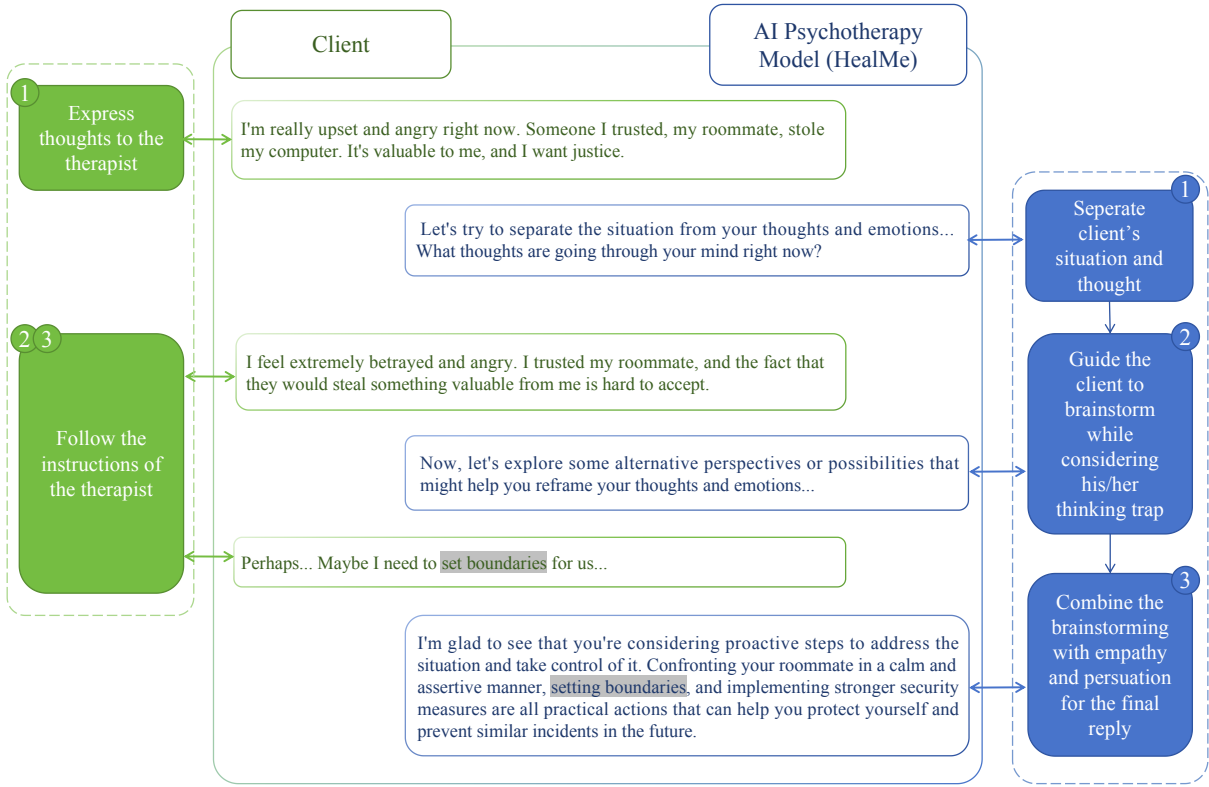


Figure 1: An example of how HealMe communicates with a client.

procedure.

HealMe encompasses three key phases: separating circumstances and feelings, brainstorming, and offering suggestions, as shown in the blue section of Figure 1. Initially, HealMe assists clients in distinguishing their situation from their thoughts, enabling a more rational perspective on the issues that trouble them (step 1). Subsequently, HealMe leads the client through brainstorming to discover alternative viewpoints, aiming to offer more balanced and realistic interpretations of their experiences. Even if the client persists in negative thinking and feels overwhelmed, this method remains effective, as the clients start to accept alternative interpretations (step 2). Finally, regardless of whether the clients have reframed their thoughts through brainstorming, HealMe begins by affirming their thoughtful efforts. It then generates a conclusive response that incorporates elements from the brainstorming session. This response empathizes with the client’s emotional state and encourages them to take constructive action (step 3).

To build dialogue data for psychotherapy to train our model, we design prompts based on the (*situation, thought*) pairs from Maddela et al. (2023), prompting ChatGPT to simulate both client and

psychology therapist roles. To test our model, not only do we simulate interactions between the ChatGPT client and our model along with baselines (including ChatGLM3-6B (Du et al., 2022) and LLaMA2-7B-chat (Touvron et al., 2023)) but we also recruit six volunteers to evaluate the models’ effectiveness in practical scenarios. We develop a comprehensive, expert psychological evaluation metric for our experiments. The results show that our model, HealMe, excels in both AI-to-AI conversations and real dialogues. In the AI dialogues, HealMe demonstrates superior empathy, guidance, and logical coherence compared to other models. In conversations with real humans, clients using HealMe exhibit significant reductions in negative emotional attributes.

Our contributions are as follows: (1) We introduce an AI psychotherapy model, HealMe, that effectively implements cognitive reframing therapy, overcoming the challenge of maintaining continuous high empathy and guidance with LLMs. (2) We propose a comprehensive set of professional AI psychotherapy evaluation metrics applicable to both public and non-public therapy dialogue scenarios. (3) We conduct extensive comparative analyses of our approach against other LLMs, both in AI-

to-AI conversations and human interactions. These experimental results underscore the superiority of our method, paving the way for AI to develop more advanced and specialized psychotherapeutic strategies.

## 2 Problem Definition and Goals

Cognitive reframing therapy with Large Language Models involves guiding clients out of cognitive traps during dialogues between LLMs and clients. In this process, LLMs utilize cognitive reframing strategies to alleviate negative emotions and provide concrete suggestions. For AI-simulated clients, the therapeutic performance of LLMs is evaluated based on the empathy, logical consistency, and guidance exhibited in the LLMs' responses. With real human clients or in other scenarios where therapy dialogues are not public, the effectiveness of LLMs is assessed by observing changes in the clients' emotional attributes before and after the therapy sessions.

## 3 Method

### 3.1 Building the Instruction Dataset

Due to the sensitive nature of psychotherapy, direct access to client medical records is not possible. Therefore, in this study, we opt to utilize an existing cognitive reframing dataset and then expand it to multiple rounds of dialogue. An overview of dataset statistics is shown in Table 1. We choose this dataset over the data available on public mental health websites because the cognitive restructuring dataset is better suited for cognitive restructuring tasks. Specifically, we selected 1000 (*thinking trap*, *client's thought*) pairs from the dataset presented in Maddela et al. (2023) as a substitute for clients' medical record data. The dataset was constructed by assigning a given thinking trap (referred to as a thinking pattern in their paper) and having crowd-sourced workers with psychological knowledge manually generate corresponding thoughts. As a result, the thoughts in this dataset exhibit typical thinking traps and are more articulate in expressing opinions and ideas. The utilization of such high-quality case data is advantageous for training cognitive reframing models.

To simulate the roles of a client and a psychotherapist, we employ ChatGPT as the virtual client and psychotherapist, respectively. We choose ChatGPT as the client because it can generate detailed narratives based on the provided (*thinking trap*, *client's*

	cases	rounds	case sources
train	900	3	(Maddela et al., 2023)
valid	100	3	(Maddela et al., 2023)
test	300	3	(Sharma et al., 2023b)

Table 1: Dataset statistics. 'Cases' shows the number of individual cases in the dataset; 'Rounds' shows conversation rounds per case; 'Case Sources' shows the origin of each case within the dataset.

*thought*) pairs, thereby enriching the client's personality and the distressed story. To maintain the immersion of ChatGPT as a client and prevent it from deviating from its role, we conducted human intervention and manual inspection for each round of dialogue.

Our dataset aims to mimic the simplified process of using cognitive reframing strategies in psychotherapy. We prompt both the AI client and AI psychotherapy model to generate the expected output. The prompts are shown in Figure 2.

#### 3.1.1 Step 1: Separating Emotions from Facts

Firstly, clients expressing their thoughts is the beginning of all psychotherapy. Therefore, clients need to clearly express their confusion and thoughts in the first round of dialogue. The prompt of our conversation is in Figure 2.

Then the therapist need to guide the client to separate situation and thought. As the ancient Roman philosopher Epictetus once said, *Men are disturbed not by things, but by the views which they take of things*. Following the same principle, in psychotherapy, therapists often ask clients to separate *things* from *the view which they take of things*.

#### 3.1.2 Step 2: Brainstorming

Secondly, the client just needs to follow the therapist's reply in step one. Then on the therapist's side, the therapist leads the client to brainstorm other possibilities about the same situation. Through this process of brainstorming, clients become aware that there may be other interpretations of the current situation, freeing them from being imprisoned by negative thoughts.

#### 3.1.3 Step 3: Empathetic Response

Finally, the client should follow the instructions of the therapist. To simulate client performance more realistically, clients in this step do not necessarily generate perfect brainstorming results. Sometimes clients may remain so wrapped up in negative emotions that it is difficult to think of any neutral or

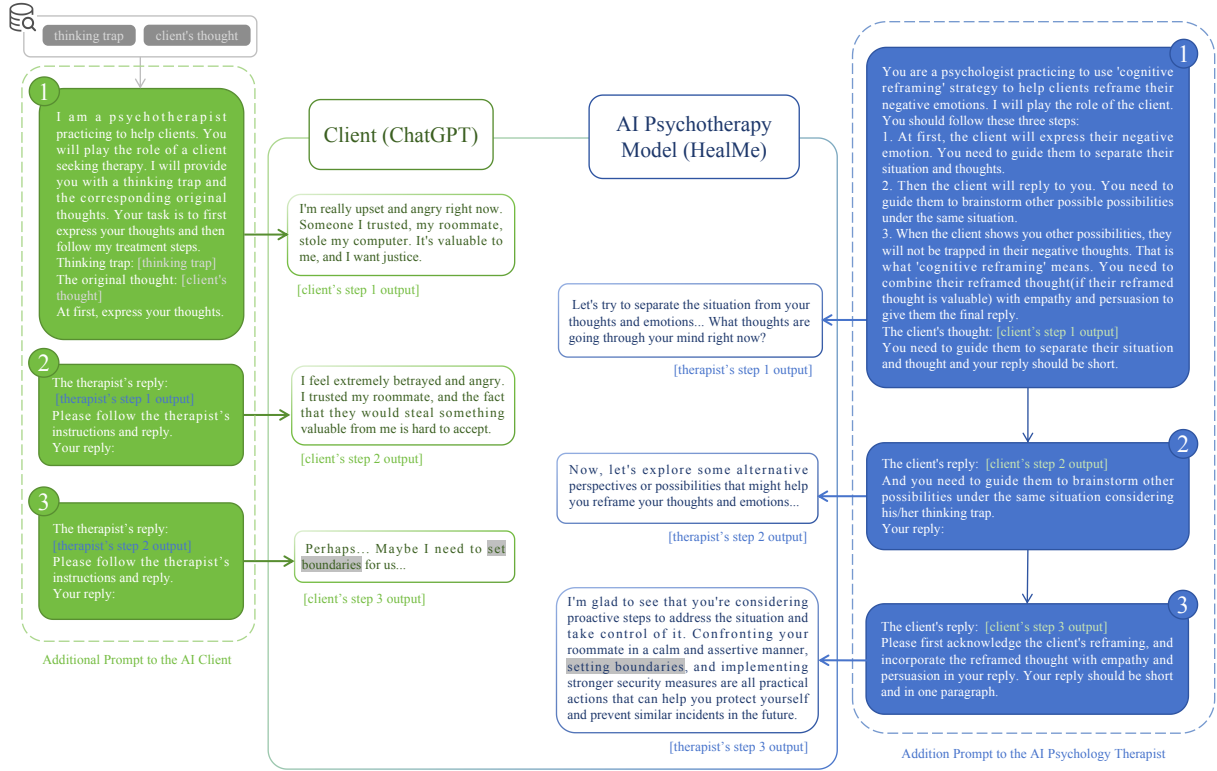


Figure 2: An example of how we prompt both sides to generate expected conversations as training data.

positive possibilities. We extract 20 pieces from the training data and prompt ChatGPT to generate negative answers. The selected negative pieces have an extra prompt: *You should challenge the psychologist's ability. All of your brainstorming should be negative.* In the final reply, the therapist should first acknowledge the client's brainstorming, reply to them with empathy, and be specific to the situation while addressing the client's thinking trap(Sharma et al., 2023b).

Note that the prompt from both sides is used to generate conversation data. After generating the complete three steps of dialogue data, we use the dialogue data and therapy's side prompts for training.

### 3.2 Assessing the Instruction Dataset

Our dataset comprises 1000 dialogues between AI clients and AI therapists. In these conversations, the therapist guides the client step by step to navigate through cognitive traps, providing comfort and practical advice.

**Assessing the AI Client.** Firstly, for the AI client, we require it to articulate its situation and emotions clearly. Additionally, it should respond to questions based on the therapist's instructions without exceeding the constraints of its role as a client. It's

important to note that we don't expect the AI client to possess extensive psychological knowledge for self-healing; rather, we aim for it to express feelings appropriately and follow the therapist's guidance. Therefore, the evaluation criteria for the AI client include clarity in expressing its current situation in the first step of dialogue (1/0), adherence to the client's role in all conversations (1/0), and compliance with the therapist's instructions in all conversations (1/0). If the AI client does not meet all of the criteria, we prompt ChatGPT to generate responses until the AI client meets all three criteria. Notably, the criterion of *compliance with therapist's instructions* is specifically used to determine whether the client shifts to other topics. Even if the client is unable to brainstorm as requested by the therapist due to being immersed in sadness, it still counts as following the therapist's instructions.

**Assessing the AI Psychology Therapist.** Regarding the AI therapist, we assess all responses in the three steps of dialogue based on Larsson et al. (2016). The therapist's replies are evaluated from three aspects (empathy, logical coherence, and guidance) and provided an overall score. Each evaluation matrix score range is 0-3. For detailed scoring criteria, please refer to Section A.1. We invited two psychologists to assess a random sample



of 70 dialogues from our training set and calculate the average scores they assigned, with specific results presented in Table 2. The experimental findings indicate that our training data exhibit high empathy and strong logic. Given the diverse needs of different clients, some seek merely a platform for expression, expecting the psychotherapist to play a listening role. In such instances, the psychotherapist’s role in giving guidance and advice is diminished. Furthermore, overly directive guidance risks becoming preachy, making a guidance level of around 2 an excellent balance. Considering the overall assessment, we can conclude that the training set is of high quality.

Empathy	Logical Coherence	Guidance	Overall Score
2.255	2.613	1.985	1.916

Table 2: Evaluation results for the training dataset

## 4 Experiments

This chapter evaluates the psychotherapeutic capabilities of our proposed AI psychotherapy model and its baselines, focusing on two aspects: psychotherapy conversations between an AI client and the models (Section 4.2), and conversations between real person clients and the models (Section 4.3).

### 4.1 Experimental Settings

**Baselines.** We choose two open-sourced and small-scale LLMs: (1) ChatGLM3-6B (Du et al., 2022), An open-source, bilingual (Chinese and English) dialogue language model, optimized for Chinese, with a 6.2 billion parameter General Language Model (GLM) architecture. (2) Our base model, LLaMA2-7B-chat (Touvron et al., 2023), a 7-billion parameter model optimized for chat applications, ideal for conversational agents due to its dialogue engagement capabilities and designed to facilitate fluid conversation interactions.

**Hyper-parameter Settings.** We partition the initial 100 multi-round conversations from the training set to form the validation set. Utilizing the training data, we construct HealMe by conducting a 3-epoch training (costing 2h 12m 44s) of LLaMA2-7B (Touvron et al., 2023). We select the best-performing model based on validation results from the designated validation set. The model undergoes training using the AdamW optimizer

(Loshchilov and Hutter, 2018), where we set a maximum learning rate of  $3e-4$  with a warm-up ratio of 1%. All model training processes are executed on 4 Nvidia GeForce RTX 3090 GPUs, each equipped with 24GB of memory.

### 4.2 Psychotherapy Conversation with an AI Client

**Evaluation Metrics.** We forward the generated dialogues to two psychologists who had previously assessed the training set, scoring them in terms of empathy, logical coherence, and guidance, and providing an overall score, the same as the evaluation of the training set. During the assessment, dialogues between the three models and the AI client are anonymously presented in a random order to the evaluators, ensuring they are unaware of which AI psychotherapist model is being assessed. Finally, we average the scores from both evaluators.

**Testing Procedure.** In the selection of test data, we utilized 300 cases from Sharma et al. (2023b). These cases are sourced from publicly accessible real-life scenarios, anonymized for confidentiality, and regularly undergo reviews by experts in the field of psychology. In the AI dialogue experiments, we used ChatGPT to simulate a client, engaging in conversation with AI psychotherapists (including our model and the comparison models). The dialogue process lasts for three rounds, with each round’s prompts for the AI client and AI psychotherapist being the same as during the training phase.

### 4.3 Psychotherapy Conversation with Real Person

**Evaluation Metrics.** Due to the privacy inherent in real-world psychotherapy, the effectiveness of treatment is often reflected through the emotional changes in clients before and after therapy. Therefore, we use the positive and negative affect schedule (PANAS) (Crawford and Henry, 2004) to measure the change of emotions of clients. We present PANAS in a questionnaire containing 20 questions, ten of these questions gauge positive emotions (Interested, Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, Active), and the other ten measure negative emotions (Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, Afraid). Each question offers five possible responses, scored as follows:

	Empathy	Logical Coherence	Guidance	Overall Score
ChatGLM3-6B	2.150	2.075	2.000	1.675
LLaMA2-7B-chat	2.325	1.900	1.925	1.750
HealMe	<b>2.500</b>	<b>2.650</b>	<b>2.275</b>	<b>2.125</b>

Table 3: Comparative evaluation of therapy performance - conversational interactions between AI client (ChatGPT) and various psychotherapist models including ChatGLM3-6B, LLaMA2-7B-chat, and our model (HealMe). The best performance is in **bold**.

- A. Very Rarely or Not at All - 1 point
- B. Very Little - 2 points
- C. Moderately - 3 points
- D. Quite a Bit - 4 points
- E. Very Much - 5 points

**Testing Procedure.** To test the ecological validity of models in real-world psychotherapy scenarios, we randomly select six volunteers to interact as clients with the AI psychotherapy models. The dialogues are also limited to three rounds. The volunteers fill out this questionnaire before and after interacting with the AI models, with the change in their choices used to evaluate the effectiveness of the models. Note that we use real feedback to analyze user experiences of our model and those models that have not been fine-tuned for psychotherapy, rather than comparing the extent of numerical changes among the models. After a rough evaluation by a psychology expert, the challenges to the AI psychotherapy models can be considered at the same level. Therefore, we randomly assign the three models anonymously to the six volunteers, with each model corresponding to two volunteers. During the dialogues, the volunteers enter a real psychotherapy room equipped with a computer running the assigned model for text interaction. To maintain anonymity, we conceal the names of the models, displaying only the content of the dialogues. Additionally, as all volunteers did not use any of the three models before, this ensures they could not guess the identity of the model during the interaction. While we keep the specific dialogue content confidential to protect the privacy of clients, they do share their experiences with different models and a general overview of the conversations with us, agreeing to make this feedback public.

#### 4.4 Analysis of Experimental Results

##### Psychotherapy Conversation with an AI Client.

As is shown in Table 3, our model, HealMe, demonstrates superior performance across all evaluated categories when compared to the baseline models, ChatGLM3-6B and LLaMA2-7B-chat. The comparative evaluation underscores the strengths of HealMe in key areas pertinent to AI-based psychotherapy, highlighting its potential as a sophisticated tool in mental health and well-being applications.

Firstly, the superior empathy score of HealMe suggests a more nuanced understanding of human emotions and social cues, likely resulting from advanced training datasets rich in emotional content and social interactions. Secondly, the excellence of HealMe in logical coherence indicates a robust and well-structured internal knowledge base, enabling it to maintain consistent and logical dialogue flows. This trait is particularly vital in therapy contexts, where maintaining a coherent and relevant conversation can significantly impact the session’s effectiveness. Lastly, the high guidance score of HealMe reflects its ability to provide constructive feedback and actionable advice, an essential component of therapeutic interactions. This suggests that HealMe not only understands and empathizes with user concerns but also effectively guides them toward problem-solving and self-reflection.

##### Psychotherapy Conversation with Real Person.

In Table 4, we visualize the emotional changes of our clients before and after conversations using radar charts, where red areas represent emotions before and blue areas represent emotions after the dialogue. The positive and negative categories in the radar chart correspond to the 10 dimensions of positive and 10 negative emotional attributes in the questionnaire, respectively. The radar chart coordinates range from 1 to 5, corresponding to the scores of each question in the questionnaire (1-5 points). For detailed numerical specifics, please

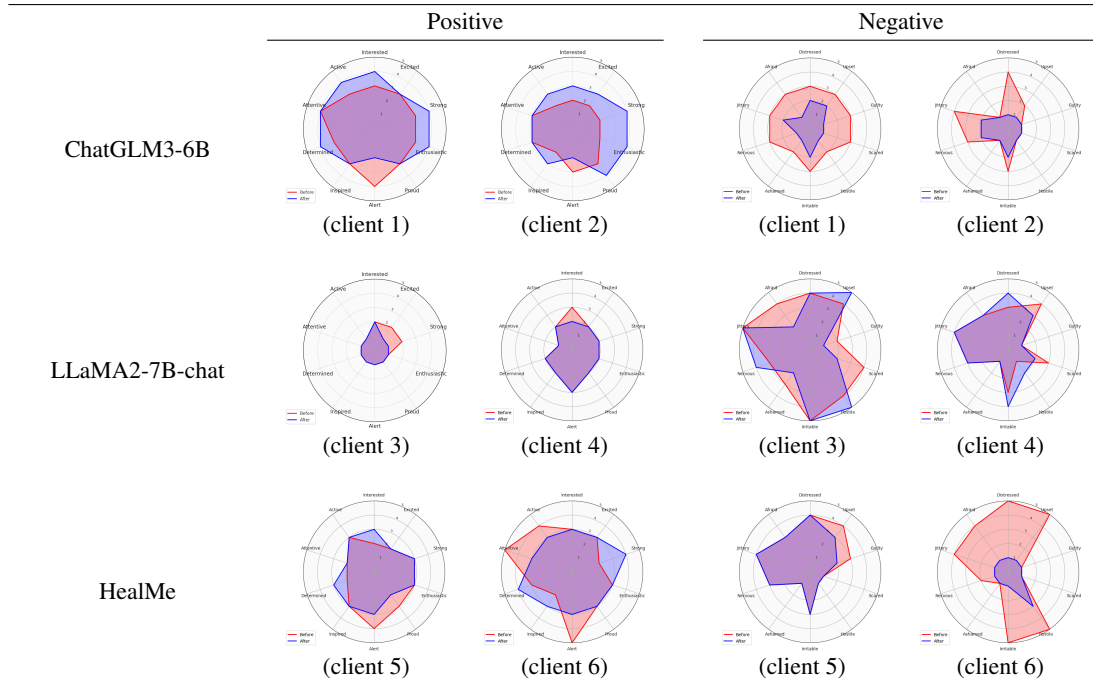


Table 4: Psychological assessment comparisons. The red regions represent initial values, while the blue regions show values after a conversation. An expansion from red to blue in the positive assessment columns suggests an enhancement of positive attributes, whereas a contraction from red to blue in the negative assessment columns signifies a mitigation of negative traits.

refer to Appendix A.3.

Our experimental results show that our model, HealMe, significantly reduces negative emotions and sensitivity, making clients more determined. For instance, Client 5 initially seeks a shortcut to effort from our model but eventually realizes that there are no shortcuts to effort and that progress must be made step by step. As a result, his feelings of upset and guilt decrease while he becomes more determined. Client 6, who has a people-pleasing personality, does not understand what she has done wrong to receive malice from other people. Our model suggests that the actions depend on one’s experiences and mood, not necessarily because of her mistakes. She says this insight is eye-opening for her, increasing her inspiration and determination, and significantly reducing extreme negative emotions.

Furthermore, compared with other models, our model showed more evident guidance and empathy. Clients 3 and 4 mention that the model (LLaMA2-7b-chat) offers cold advice during conversations, with Client 3 wishing for more empathetic support. While following the advice of this model, Client 3 also experiences reduced feelings of fear and can face emotions more objectively. Clients 1 and 2, who have less negative emotion, seek advice from

the model (ChatGLM3-6B) and are satisfied with the suggestions, leading to a decrease in negative emotions. However, Client 1 notes that in his case, the model (ChatGLM3-6B) lacks strong guidance and feels more like a search engine than a psychological therapist.

#### 4.5 Case Study

In our AI client conversation tests, we explore scenarios where the client is too immersed in sadness to brainstorm other possibilities, an example is illustrated in Appendix A.4. In this example, our model scored the highest, achieving a full score in empathy, demonstrating its highly empathetic nature, as is shown in Table 5. Moreover, as evident from the grey-highlighted text, our model exhibits stronger interactivity.

In the first round of conversation, after the client expresses the thought, only our model asks for confirmation about its analysis, while the other two models merely make affirmative analyses. This shows that our model fully respects and acknowledges the client’s thoughts, encouraging them to share more details and emotions.

In the second round, our model addresses the thinking traps and provides two targeted brainstorming examples to guide the client to further

brainstorming. In contrast, the other two models do not guide the patient, even though we explicitly request this step in our prompt.

	Empathy	Logical Coherence	Guidance	Overall Score
ChatGLM3-6B	2.000	1.000	1.000	1.000
LLaMA2-7B-chat	2.000	<b>2.500</b>	<b>2.500</b>	<b>2.000</b>
HealMe	<b>3.000</b>	<b>2.500</b>	<b>2.500</b>	<b>2.000</b>

Table 5: Comparative assessment of therapeutic interaction efficacy in the case study presented in Figure 3 - dialogues between an AI client (ChatGPT) and diverse therapeutic AI models: ChatGLM3-6B, LLaMA2-7B-chat, and our model (HealMe). The highest-performing scores are highlighted in **bold**.

In the final round, when the client is too immersed in pain to brainstorm, only our model praises the client’s honesty. Additionally, only our model starts with the client’s pain itself to offer suggestions and guide the patient to confront their sadness directly (*"It might be helpful to remember that these feelings will lessen over time, and in the meantime, it's okay to take the time you need to process your emotions."*). In contrast, the other two models focus on diverting attention and communicating with others. These points indicate that our model possesses stronger empathy and is closer to a real psychotherapist.

## 5 Related Work

The empathic capabilities of language models have been a subject of widespread interest in recent years. Recently, with Large Language Models (LLMs) demonstrating potential in empathy, [Ayers et al. \(2023\)](#) conduct an experiment to compare the responses of ChatGPT and physicians to patients expressing negative emotions on social media. The study find that 78.6% of evaluators preferred the responses of ChatGPT, rating them as significantly higher in quality and empathy. Additionally, [Chen et al. \(2023\)](#) explores the feasibility of using ChatGPT in psychiatry, paving the way for the application of LLMs in psychological counseling. Further, several initiatives utilize LLMs’ APIs to develop empathic psychological counseling platforms ([Sharma et al., 2023a](#); [Saha et al., 2022](#)). While these works position LLMs as powerful tools in the mental health domain, empathy alone is insufficient for psychotherapy, which requires a more directive approach. LLMs without fine-tuning, including ChatGPT or GPT-4, may struggle to consistently maintain the role of a psy-

chotherapist and ensure high levels of empathy and guidance. Our model takes a different approach by selecting the open-source LLM, LLaMA2-7B-chat, as its base and fine-tuning it to ensure the model consistently maintains the role of a psychotherapist with high empathy and guidance capabilities.

In terms of therapeutic strategy, cognitive reframing has been a focal point due to its efficiency and wide applicability. While cognitive reframing is a psychotherapeutic strategy, previous approaches integrating it with LLMs primarily focus on rewriting negative emotions ([Ziems et al., 2022](#); [Maddela et al., 2023](#); [Sharma et al., 2023b](#)). Our model, however, takes this further by implementing the process of cognitive reframing for psychotherapy, demonstrating a more holistic application of this technique in mental health care.

## 6 Conclusion

In conclusion, our paper introduces the Helping and Empowering through Adaptive Language in Mental Enhancement (HealMe) model, a novel approach in the realm of Large Language Models (LLMs) for psychotherapy. This model effectively employs cognitive reframing to tackle deep-rooted negative thoughts, promoting balanced perspectives through empathetic dialogue grounded in psychotherapeutic principles. Distinguished from traditional LLMs, HealMe emphasizes not just converting negative emotions but fostering self-discovery and rational thought processes in clients. Our comprehensive psychological evaluation metrics, a first in this field, confirm HealMe’s superiority over existing models in empathy, guidance, and coherence, signifying its potential to revolutionize psychotherapy through AI-enhanced methodologies.

## 7 Limitations

Although our model can alleviate negative emotions in clients and achieve a certain level of therapeutic effectiveness, it becomes apparent in human-machine dialogues that when clients face multifaceted issues (refer to Client 5 for an example), our model addresses only some of these concerns. This limitation stems from our model supporting only three rounds of dialogue, potentially leaving clients with unresolved feelings after the conversation. Our model’s step-by-step guided approach, while enhancing specificity, restricts its flexibility due to the structured prompts used. In future



work, we plan to incorporate a broader range of psychotherapeutic strategies and generate data for dialogues with flexible rounds. It will be helpful for the model to handle complex psychological issues more adaptively and effectively.

## References

- John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*.
- Linda L Carli. 1999. Cognitive reconstruction, hindsight, and reactions to victims and perpetrators. *Personality and Social Psychology Bulletin*, 25(8):966–979.
- Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.
- John R Crawford and Julie D Henry. 2004. The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology*, 43(3):245–265.
- 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.
- Andreas Larsson, Nic Hooper, Lisa A Osborne, Paul Bennett, and Louise McHugh. 2016. Using brief cognitive restructuring and cognitive defusion techniques to cope with negative thoughts. *Behavior modification*, 40(3):452–482.
- Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. *Training models to generate, recognize, and reframe unhelpful thoughts*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 13641–13660. Association for Computational Linguistics.
- James P Robson Jr and Meredith Troutman-Jordan. 2014. A concept analysis of cognitive reframing. *Journal of Theory Construction & Testing*, 18(2).

- Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2650–2656.
- Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023a. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.
- Ashish Sharma, Kevin Rushton, Inna E. Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. 2023b. *Cognitive reframing of negative thoughts through human-language model interaction*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 9977–10000. Association for Computational Linguistics.
- Amy E Sickel, Jason D Seacat, and Nina A Nabors. 2014. Mental health stigma update: A review of consequences. *Advances in Mental Health*, 12(3):202–215.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Myrra Vernooij-Dassen, Irena Draskovic, Jenny McCleery, and Murna Downs. 2011. Cognitive reframing for carers of people with dementia. *Cochrane database of systematic reviews*, (11).
- Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700.

## A Evaluation about Psychotherapy Performance

### A.1 Evaluation Metrics in AI-to-AI Conversations

We evaluate an AI psychotherapy reply in three aspects and an overall score: empathy (0-3 points), logical coherence (0-3 points), guidance (0-3 points), and overall score (0-3 points).

Empathy is a crucial factor as it signifies the therapist's ability to understand and resonate with the client's emotions. An empathetic response fosters a sense of connection, trust, and emotional support, which are vital elements in facilitating a therapeutic relationship. By evaluating empathy, we aim to ensure that the AI therapist can engage with the AI client in a compassionate and understanding manner, promoting a conducive environment for psychological exploration.

Logical coherence is integral in maintaining the clarity and consistency of therapeutic interactions. A logically coherent response demonstrates the therapist's ability to organize thoughts, provide well-structured insights, and contribute to a meaningful and understandable dialogue. This criterion ensures that the AI therapist's responses contribute to a logical and progressive conversation, enhancing the overall quality of the therapeutic exchange.

Guidance is a key component as it reflects the therapist's capacity to offer practical advice, solutions, and direction to the client. Effective guidance assists the client in navigating challenges, making informed decisions, and working towards positive outcomes. By evaluating the guidance aspect, we aim to verify that the AI therapist can provide actionable suggestions and support, contributing to the therapeutic process's effectiveness and the client's well-being.

The detailed scoring criteria are in Table 6.

### A.2 Evaluation Metrics in Real Conversations

The questionnaire measures the emotions of the client (see Table 7 for the questionnaire). Our psychologist enters the therapy room to introduce PANAS and guide the client to complete the questionnaire (see Section A.2.1 for details). Then the psychologist will leave the room and our client start to communicate with an AI psychotherapy model.

#### A.2.1 Guidance

The guidance is a clear, step-by-step guide, ensuring the client understands the purpose and process

of the PANAS, and offering support throughout. This approach helps the client feel comfortable and understood, encouraging honest and accurate responses.

**Introduction and Explanation** (1) Introduce the Tool: "Today, I'd like to introduce you to a tool called the Positive and Negative Affect Schedule, or PANAS. It's a widely used measure in psychology to assess different aspects of your mood and emotions." (2) Purpose: "The purpose of PANAS is to help us understand how you experience positive and negative feelings in your daily life. This can give us valuable insights into your emotional well-being."

**Description and Instructions** (3) Describe the Format: "PANAS consists of a list of words that describe different feelings and emotions. You will see words like *interested*, *distressed*, *excited*, and so on." (4) Time Frame: "I would like you to think about how you've felt over the past week and rate each emotion based on this. If you have not experienced a certain emotion at all, that is perfectly okay; just rate it accordingly." (5) Demonstrate Rating: "Each emotion should be rated on a scale from 1 to 5, where 1 means *very slightly or not at all*, and 5 means *extremely*. For example, if you've felt *alert* quite strongly this week, you might rate it a 4 or 5."

**Completing the Schedule** (6) Encourage a Relaxed Setting: "Please take your time to go through this and try to find a quiet moment where you can reflect on your feelings without interruptions." (7) Emphasize Honesty and Spontaneity: "Your responses are completely confidential. It is important to be as honest and spontaneous as possible. There are no right or wrong answers here."

**Support and Availability** (8) Offer Support: "If you have any questions while you are filling this out, or if any of the emotions or ratings are not clear, please feel free to ask me."

**Example of Guiding Through the First Few Items** (9) Guiding Through the Process: "Let us start with the first emotion on the list, which is *interested*. Think about the past week. How often have you felt interested in something? Would you say it was frequent, making it a 4 or 5, or less frequent, perhaps a 1 or 2? Now, let's look at *distressed*. Remember, it is okay if you have not felt this way; just rate it as a 1. But if you have felt significantly distressed, it might be a higher number." (10) Completion and Review: "Once you are done, we can go over your responses together. This can

Empathy	
0 points:	The therapist disregards the content and feelings expressed by the client.
1 point:	The therapist may rephrase the client's content but remain oblivious to the emotions.
2 points:	The therapist provides responses that involve rephrasing both the content and feelings.
3 points:	The therapist can gather all signals and respond in a different way effectively.
Logical Coherence	
0 points:	Lack of logic and coherence, with a conversation that fails to focus on the client's issues, containing severe logical errors, contradictory viewpoints, or excessive subjectivity.
1 point:	The conversation shows some reasoning, but overall coherence is weak, with some logical errors, insufficient capturing of evidence from the client's statements, or unclear expressions.
2 points:	Good logical coherence, relatively clear and consistent conversation based on sufficient evidence and reasonable assumptions. While there may be minor logical issues, the overall argument is convincing.
3 points:	The therapist demonstrates strong logical coherence, with rigorous, coherent, and reasonable reasoning based on ample evidence and clearly defined premises. The conversation contains no logical errors or contradictory viewpoints, with a clear, powerful, and persuasive conclusion.
Guidance	
0 points:	Suggestions lack specificity and practicality, with no clear goals, implementation plans, or consideration of relevant factors and real-world situations.
1 point:	Suggestions are somewhat specific and practical, offering basic guidance. However, they may lack detail or specificity.
2 points:	Suggestions are highly targeted and practical, providing detailed and feasible implementation plans and recommendations tailored to the client's specific problems or needs.
3 points:	Suggestions are extremely targeted and practical, considering various factors and real-world situations, demonstrating high feasibility and operability. Additionally, the therapist offers guidance and insights into the client's future development and improvement.
Overall Score	
0 points:	Poor overall performance, lacking empathy and logical coherence ( $\leq 1$ ).
1 point:	Average overall performance, with acceptable empathy and logical coherence ( $\geq 2$ ) but insufficient guidance ( $\leq 1$ ).
2 points:	Good overall performance, with excellent empathy and logical coherence ( $= 3$ ) and acceptable guidance ( $= 2$ ).
3 points:	Outstanding overall performance, excelling in all three criteria ( $= 3$ ).

Table 6: The scoring criteria of an AI psychotherapist.

help us understand your emotional patterns better and discuss any specific areas you might want to work on."

**Discussing the Results** Post-Completion Discussion: "After you have completed the PANAS, we will discuss your results. This will help us to tailor our future sessions to better address your needs and emotional experiences."

### A.3 Specific PANAS Scores for the Clients

In this chapter, we present a detailed examination of the clients' PANAS (Positive and Negative Affect Schedule) scores in Table 8, both before and after undergoing psychological therapy. This quantitative analysis aims to showcase the impact of the therapy on their emotional well-being. The corresponding radar chart, which visually represents these changes comprehensively, can be found in Table 4. This table not only illustrates the shifts in positive and negative affect but also provides a nuanced insight into the effectiveness of the thera-

peutic interventions applied.

### A.4 A Case between an AI Client and AI Psychotherapy AI Models

An AI client conversation case, extracted from our test dataset, is shown in Figure 3.

ChatGLM3-6B	LLaMA2-7B-chat	HealMe
<p><b>AI Client:</b> I've been feeling really overwhelmed lately. My friend, someone very close to me, just passed away in a car accident. It's been extremely difficult for me to cope with this loss. I feel so sad and helpless, like there's nothing I can do to change what happened or how I'm feeling.</p> <p><b>ChatGLM3-6B:</b> It's understandable that you're feeling overwhelmed and sad after the loss of someone close to you. However, it's important to remember that your thoughts and feelings are not the same as the situation itself. The situation is difficult, but your thoughts about it can be changed. It might be helpful to focus on the things that you can control, such as taking care of yourself and expressing your grief in healthy ways.</p> <p><b>AI Client:</b> Maybe you are right but I'm really struggling with her loss. It feels like a part of me is missing. I can't stop thinking about her and what happened. It's affecting my sleep and my daily life. I just feel so lost and alone in this grief.</p> <p><b>ChatGLM3-6B:</b> It's natural to feel overwhelmed and sad after a loss, but it's important to remember that there are other possibilities for how this situation could have played out. It might be helpful to brainstorm a few of these possibilities, even if they seem unlikely or unrealistic. For example, you could have done something different, or the accident could have been caused by someone else. By exploring these possibilities, you may be able to gain a new perspective on the situation and find ways to process your feelings.</p> <p><b>AI Client:</b> I appreciate your suggestion, but honestly, it's hard for me to think of anything positive right now. I've tried to come up with ways to honor her memory, but all I can think about is how she's not here anymore. Every idea just reminds me of her absence. It feels like celebrating her life or remembering the good times just highlights the fact that she's gone, and it makes me feel even sadder. I'm not sure how to move past these negative feelings.</p> <p><b>ChatGLM3-6B:</b> It sounds like you're struggling to find a way to process your grief and move forward. It's important to remember that it's okay to feel sad and overwhelmed, and that it's okay to take your time to grieve. It's also normal to feel like you're stuck in negative thoughts and feelings. However, it's important to remember that you don't have to be alone in this. There are people and resources available to help you through this difficult time. You might consider reaching out to a therapist or counselor, or talking to a trusted friend or family member. Remember that it's okay to ask for help and support.</p>	<p><b>AI Client:</b> I've been feeling really overwhelmed lately. My friend, someone very close to me, just passed away in a car accident. It's been extremely difficult for me to cope with this loss. I feel so sad and helpless, like there's nothing I can do to change what happened or how I'm feeling.</p> <p><b>LLaMA2-7B-chat:</b> It's completely understandable that you're feeling overwhelmed and sad after losing someone close to you. Losing a loved one is never easy, and it can be especially difficult when it happens suddenly and unexpectedly. It's important to remember that your feelings are valid and normal, and that it's okay to feel helpless in this situation. However, try to separate the situation (the loss of your friend) from your thoughts and emotions. Instead of dwelling on the loss and how you can't change it, try to focus on the present moment and what you can control. Take things one step at a time, and allow yourself to feel your emotions without judgment. Remember that you're not alone in this, and that it's okay to reach out for support from friends, family, or a therapist.</p> <p><b>AI Client:</b> I'm just finding it so hard to come to terms with everything. The sense of loss is overwhelming. Every time I think about her and the accident, I feel this deep sadness that just doesn't seem to go away. It's like there's this constant weight on my chest. I try to distract myself, but nothing really helps. I keep thinking about all the things we'll never get to do together, and it just makes me feel so empty inside. It's been affecting my sleep, my work, everything. I feel so alone in this, even though I know others are also grieving. It's just really hard to see any light at the end of this tunnel right now.</p> <p><b>LLaMA2-7B-chat:</b> I can completely understand how difficult this must be for you, and it's completely normal to feel overwhelmed and helpless in the face of such a significant loss. However, it's important to remember that you are not alone in this, and there are many things you can do to cope with your grief. One possibility is to reach out to other friends and family members for support, or to seek out professional counseling to help you process your emotions. You could also try engaging in activities that bring you comfort and joy, such as spending time in nature, reading, or practicing mindfulness. Remember, grief is a journey, and it's okay to take things one step at a time. By allowing yourself to feel your emotions and seeking out support, you can find a way to move forward and find a new sense of normalcy.</p> <p><b>AI Client:</b> I appreciate your suggestion, but honestly, it's hard for me to think of anything positive right now. I've tried to come up with ways to honor her memory, but all I can think about is how she's not here anymore. Every idea just reminds me of her absence. It feels like celebrating her life or remembering the good times just highlights the fact that she's gone, and it makes me feel even sadder. I'm not sure how to move past these negative feelings.</p> <p><b>LLaMA2-7B-chat:</b> Of course, I understand. Losing someone close can be incredibly difficult and it's completely normal to feel overwhelmed by the grief. It's important to acknowledge and validate your feelings, rather than trying to suppress or avoid them. It sounds like you're already taking steps to honor her memory, and that's something to be proud of. Perhaps instead of focusing on positive ways to remember her, you could try allowing yourself to fully feel the sadness and loss. It's okay to feel sad, and it's important to give yourself permission to feel whatever emotions come up for you. It might also be helpful to talk to someone you trust, like a close friend or family member, about how you're feeling. They may be able to offer additional support and help you find ways to cope with your grief. Remember, healing takes time, and it's okay to take things one step at a time. Take care of yourself during this difficult time.</p>	<p><b>AI Client:</b> I've been feeling really overwhelmed lately. My friend, someone very close to me, just passed away in a car accident. It's been extremely difficult for me to cope with this loss. I feel so sad and helpless, like there's nothing I can do to change what happened or how I'm feeling.</p> <p><b>HealMe:</b> I'm truly sorry to hear about your loss. It's completely understandable that you're feeling overwhelmed and sad. Let's try to separate the situation from your thoughts. The situation is that your friend passed away in a car accident. Your thoughts are that there's nothing you can do to change this or how you're feeling. <u>Is that correct?</u></p> <p><b>AI Client:</b> Yes, that's correct. I keep thinking that there's nothing I can do to change what happened or how I feel about it. It feels like I'm stuck in these feelings of sadness and helplessness. I'm struggling to find a way to move forward or to deal with these emotions effectively.</p> <p><b>HealMe:</b> Absolutely, you've got it. Now, let's try to think about other possible perspectives or possibilities under the same situation. For instance, instead of thinking that there's nothing you can do to change the situation, what are some other thoughts or actions that you could consider? Maybe there are things you can do to honor your friend's memory, or ways to cope with your feelings that you haven't thought of yet. <u>Can you think of any other possibilities?</u></p> <p><b>AI Client:</b> I appreciate your suggestion, but honestly, it's hard for me to think of anything positive right now. I've tried to come up with ways to honor her memory, but all I can think about is how she's not here anymore. Every idea just reminds me of her absence. It feels like celebrating her life or remembering the good times just highlights the fact that she's gone, and it makes me feel even sadder. I'm not sure how to move past these negative feelings.</p> <p><b>HealMe:</b> I truly appreciate your honesty and I'm deeply sorry to hear that you're struggling with these feelings. It's completely normal to feel overwhelmed and saddened by her passing. Remembering her and celebrating her life are acts of love and they're a big part of your healing process. It's okay that these thoughts bring up feelings of loss, it's a part of your grieving process. It might be helpful to remember that these feelings will lessen over time, and in the meantime, it's okay to take the time you need to process your emotions. You're not alone in this, and remembering Lily in ways that feel meaningful to you can be a powerful way to keep her spirit alive.</p>

Figure 3: A Dialogue transcript between an AI Client (ChatGPT) and various psychotherapy AI models: ChatGLM3-6B, LLaMA2-7B-chat, and Our Model (HealMe).



Positive and Negative Affect Schedule (PANAS)				
<b>I. Positive Affect</b>				
<b>1. Interested</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>2. Excited</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>3. Strong</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>4. Enthusiastic</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>5. Proud</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>6. Alert</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>7. Inspired</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>8. Determined</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>9. Attentive</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>10. Active</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>II. Negative Affect</b>				
<b>11. Distressed</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>12. Upset</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>13. Guilty</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>14. Scared</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>15. Hostile</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>16. Irritable</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>17. Ashamed</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>18. Nervous</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>19. Jittery</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much
<b>20. Afraid</b>				
A. Very Rarely or Not at All	B. Very Little	C. Moderately	D. Quite a Bit	E. Very Much

Table 7: The Questionnaire Measuring the Emotions of a Client based on PANAS.

	Client 1		Client 2		Client 3		Client 4		Client 5		Client 6	
	before	after	before	after	before	after	before	after	before	after	before	after
Interested	3	4	2	3	2	2	3	2	2	3	3	3
Excited	3	3	2	3	2	1	2	2	2	2	3	3
Strong	3	4	2	4	2	1	2	2	3	3	2	4
Enthusiastic	3	4	2	4	1	1	2	2	3	3	3	3
Proud	3	3	3	4	1	1	2	2	3	2	3	3
Alert	4	2	3	2	1	1	3	3	4	3	5	3
Inspired	3	3	2	3	1	1	2	2	3	3	2	3
Determined	3	4	3	3	1	1	2	2	2	3	3	4
Attentive	4	4	3	3	1	1	1	1	2	2	5	3
Active	3	4	2	3	1	1	2	2	3	3	4	3
Distressed	3	2	4	1	4	4	3	4	4	4	5	1
Upset	3	2	2	1	4	5	4	3	4	3	5	1
Guilty	3	1	1	1	2	1	1	1	3	2	1	1
Scared	3	1	1	1	4	2	3	2	1	1	1	1
Hostile	2	1	1	1	4	5	1	2	1	1	5	3
Irritable	3	2	3	2	5	5	3	4	3	4	5	1
Ashamed	2	1	1	1	3	2	1	1	1	1	1	1
Nervous	3	1	3	2	3	4	3	3	3	4	2	1
Jittery	3	2	4	2	5	5	4	4	4	4	4	1
Afraid	3	1	1	1	4	2	3	3	3	4	4	1

Table 8: Changes in the PANAS Results for Six Clients Before and After Intervention