HealMe: Harnessing Cognitive Reframing in Large Language Models for Psychotherapy

Mengxi Xiao[a*], **Qianqian Xie**[d*], **Ziyan Kuang**[b], **Zhicheng Liu**[b], **Kailai Yang**[c], **Min Peng**[a†]
Weiguang Han[a], **Jimin Huang**[d]
[a]School of Computer Science, Wuhan University
[b]Jiangxi Normal University
[c]The University of Manchester
[d]The FinAI

Abstract

Large Language Models (LLMs) can play a vital role in psychotherapy by adeptly handling the crucial task of cognitive reframing and 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. This novel cognitive reframing therapy method 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 on psychotherapy.

1 Introduction

Cognitive reframing (Carli, 1999), a key part of cognitive-behavior therapy (CBT), 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) hold immense potential whether as an adjunct to human-based mental health treatment or as a standalone therapeutic tool (Stade et al., 2023). LLMs can help overcome obstacles (Huang et al., 2023) 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(Sharma et al., 2023b).

Contrasting with previous methods that conceptualize cognitive reframing as a sentence rewriting task (Ziems et al., 2022; Maddela et al., 2023), where negative emotions are transformed into neutral or positive expressions emphasizing factors like specificity and actionability (Sharma et al., 2023b), our approach marks a significant shift. Since cognitive reframing emphasizes the importance of clients undergoing cognitive changes themselves, rather than directly receiving guidance or suggestions from therapists (Hofmann et al., 2014; Johnco et al., 2014; Edwards, 1989), our method employs a conversational model that directly engages with and actively transforms the client's own negative thoughts.

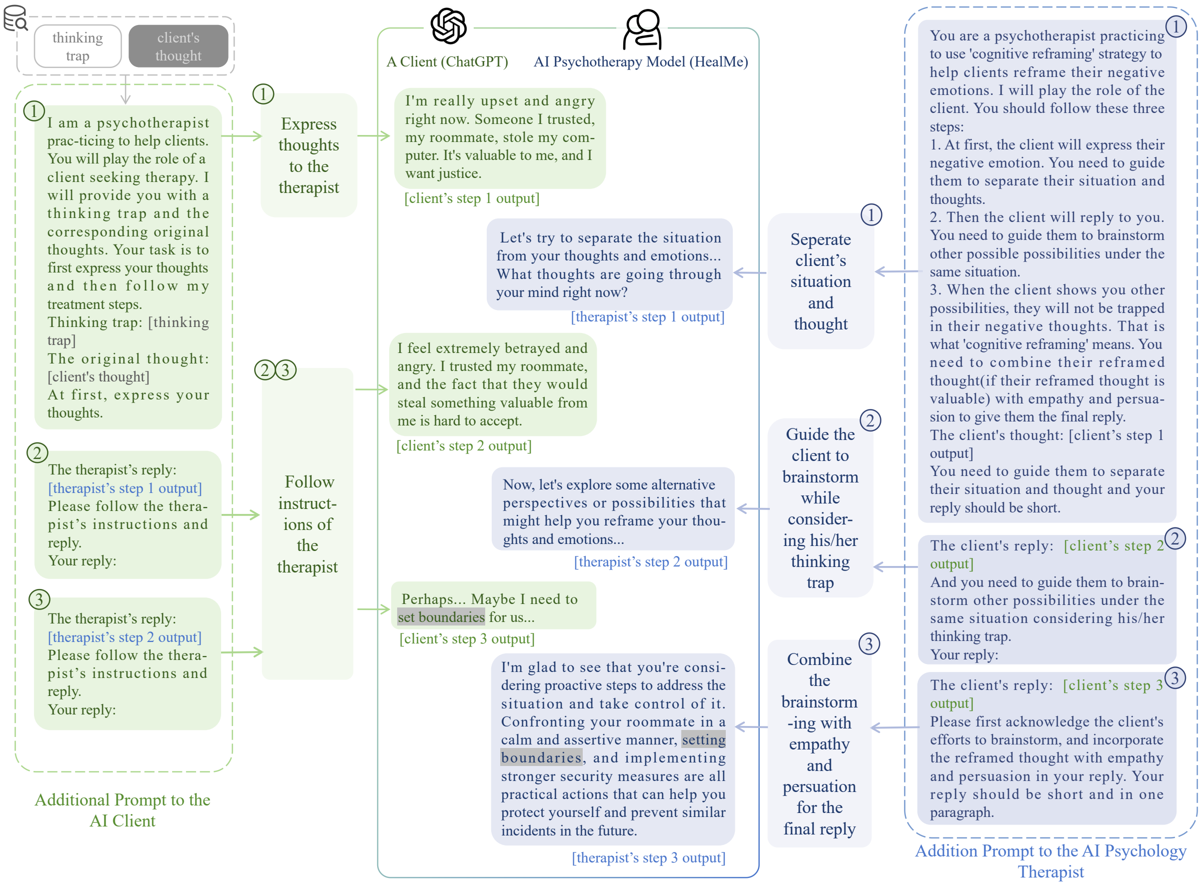
Thus, despite the significant potential shown by LLMs in prior research, they encounter crucial obstacles when it comes to cognitive reframing. (1) If viewing cognitive reframing as a sentence rewriting task, clients might not spontaneously discover alternative perspectives and could perceive the reframing as preaching or imposition rather than self-realization. (2) LLMs cannot consistently generate concrete and specific empathetic responses (Sorin et al., 2023), which are crucial in psychotherapy (Bohart and Greenberg, 1997). For instance, a specific empathetic response might be, *I understand how upsetting it is that your friend forgot your birthday.* In contrast, a more general response

---

[*] These authors contributed equally to this work.

[†] The Corresponding Author. Email: pengm@whu.edu.cn

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1707–1725
August 11-16, 2024 ©2024 Association for Computational Linguistics


thinking trap
client's thought
①
Express thoughts to the therapist
②③
Follow instruct-ions of the therapist
A Client (ChatGPT)
AI Psychotherapy Model (HealMe)
①
Seperate client's situation and thought
②
Guide the client to brainstorm while consider-ing his/her thinking trap
③
Combine the brainstorm-ing with empathy and persuation for the final reply

① I am a psychotherapist prac-ticing to help clients. You will play the role of a client seeking therapy. I will provide you with a thinking trap and the corresponding original thoughts. Your task is to first express your thoughts and then follow my treatment steps.
Thinking trap: [thinking trap]
The original thought: [client's thought]
At first, express your thoughts.

② The therapist's reply: [therapist's step 1 output]
Please follow the therapist's instructions and reply.
Your reply:

③ The therapist's reply: [therapist's step 2 output]
Please follow the therapist's instructions and reply.
Your reply:

Additional Prompt to the AI Client

I'm really upset and angry right now. Someone I trusted, my roommate, stole my computer. It's valuable to me, and I want justice.
[client's step 1 output]

Let's try to separate the situation from your thoughts and emotions... What thoughts are going through your mind right now?
[therapist's step 1 output]

I feel extremely betrayed and angry. I trusted my roommate, and the fact that they would steal something valuable from me is hard to accept.
[client's step 2 output]

Now, let's explore some alternative perspectives or possibilities that might help you reframe your thoughts and emotions...
[therapist's step 2 output]

Perhaps... Maybe I need to set boundaries for us...
[client's step 3 output]

I'm glad to see that you're considering proactive steps to address the situation and take control of it. Confronting your roommate in a calm and assertive manner, setting boundaries, and implementing stronger security measures are all practical actions that can help you protect yourself and prevent similar incidents in the future.
[therapist's step 3 output]

① You are a psychotherapist practicing to use 'cognitive reframing' strategy to help clients reframe their negative emotions. I will play the role of the client. You should follow these three steps:
1. At first, the client will express their negative emotion. You need to guide them to separate their situation and thoughts.
2. Then the client will reply to you. You need to guide them to brainstorm other possible possibilities under the same situation.
3. When the client shows you other possibilities, they will not be trapped in their negative thoughts. That is what 'cognitive reframing' means. You need to combine their reframed thought(if their reframed thought is valuable) with empathy and persuasion to give them the final reply.
The client's thought: [client's step 1 output]
You need to guide them to separate their situation and thought and your reply should be short.

② The client's reply: [client's step 2 output]
And you need to guide them to brainstorm other possibilities under the same situation considering his/her thinking trap.
Your reply:

③ The client's reply: [client's step 3 output]
Please first acknowledge the client's efforts to brainstorm, and incorporate the reframed thought with empathy and persuasion in your reply. Your reply should be short and in one paragraph.

Addition Prompt to the AI Psychology Therapist

Figure 1: An example of how HealMe communicates with a client, and how we prompt both sides to generate expected conversations as training data.

would be, *I understand your feelings.* (3) While LLMs are commonly used for answering human queries, the role is reversed in psychotherapy: psychotherapists are required to guide humans (James et al., 2010). According to Westerners effect, excessive external motivation can undermine internal motivation. As such, therapists giving direct suggestions may hinder clients' self-discovery and the development of self-efficacy. Moreover, guidance fosters a more collaborative environment, allowing clients to explore and understand their thoughts and feelings, leading to more sustainable and self-directed change.

To tackle these challenges, we propose a specialized model **H**elping and **E**mpowering through **A**daptive **L**anguage in **M**ental **E**nhancement (HealMe[1]), for cognitive reframing therapy. We emphasize the empowerment of the client rather than reliance on therapist-driven solutions. We leverage 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), our domain-expert co-authors distill and organize a structured cognitive reframing therapy process, effectively emulating a complete psychotherapeutic procedure.

HealMe operates in three main stages, as depicted in the blue section of Figure 1: 1) distinguishing between situations and thoughts for a rational outlook, 2) brainstorming for alternative perspectives to mitigate negative thinking, and 3) offering suggestions that acknowledge the client's effort and encourage positive action. This streamlined process aids clients in understanding their issues more clearly, accepting new interpretations, and moving toward constructive solutions.

To build dialogue data for psychotherapy to train our model, we design prompts based on the (*thinking trap*, *client's thought*) pairs from Maddela et al. (2023), prompting ChatGPT to simulate both client and psychology therapist roles. To test our model, we simulate interactions between the ChatGPT client and our model along with baselines (including ChatGLM3-6b (Du et al., 2022)

---

[1] Our data and code are available at HealMe.

and LLaMA2-7b-chat (Touvron et al., 2023)). We also conduct experiments to evaluate the models' effectiveness in practical scenarios. We create a detailed psychological evaluation metric for our experiments, incorporating a three-dimensional scoring system to evaluate AI therapy models in AI client scenarios. For real-person client scenarios, we directly employ professional psychological metrics to evaluate AI therapy models. The results show that HealMe excels in both AI-to-AI conversations and real-world dialogues. In AI-to-AI dialogues, HealMe demonstrates superior empathy, guidance, and logical coherence compared to other models. During real-person testing, some clients using HealMe experienced notable decreases in negative emotional attributes, with negative scores dropping from 5/5 to 1/5, highlighting HealMe's efficacy in real-world scenarios.

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 Dataset Construction

In this study, we leverage an existing raw dataset focused on cognitive reframing and expand it to include multiple rounds of dialogue. Specifically, we conduct a manual review of the selected raw dataset and select 1,000 well-composed pairs of (*thinking trap*, *client's thought*) from it.

The raw dataset we utilize in this study is introduced by Maddela et al. (2023). Our selection of this dataset is based on its comprehensive representation of common thinking traps, characterized by effectively articulated thoughts. The creation process of the raw dataset involves assigning specific thinking traps (identified as thinking patterns in the original study) and engaging crowd-sourced workers with a psychology background to manually generate these thoughts. This methodology ensures the dataset's relevance and quality, making it an ideal foundation for our research.

To simulate the roles of a client and a psychotherapist, we employ ChatGPT (gpt-3.5-turbo-0125) 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 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 1, and our constructed dataset statistics are shown in Table 1.

	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.

3.1 Step 1: Separating Emotions from Facts

The client's side. Firstly, we simulate the clients to express their thoughts as the beginning of therapy. Therefore, clients need to clearly express their confusion and thoughts in the first round of dialogue.

The therapist's side. Then we simulate the therapist guiding the client to separate situation and thought (Chen et al., 2023b).

3.2 Step 2: Brainstorming

The client's side. We simulate the client to separate situation and thought, following the therapist's guidance in step one.

The therapist's side. We simulate the therapist to guide the client in brainstorming alternative perspectives under a given situation. By asking questions such as, "How would you comfort a friend in this situation?" the therapist flexibly facilitates this process. Unlike previous studies, our approach to brainstorming does not seek perfect reframing but rather aims to help clients discover different viewpoints. Through this brainstorming process, clients realize there are other ways to interpret their current situation, which liberates them from the confines of negative thinking.

3.3 Step 3: Empathetic Response

The client's side. We simulate the ChatGPT client to follow the instructions of the therapist to brainstorm. To simulate client performances more realistically, ChatGPT 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 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.*

The therapist's side. We simulate the therapist to generate the final response. The therapist should first recognize clients' efforts in reframing and appreciate their willingness to brainstorm in other cases. Then the therapist replies to them with empathy and is 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.

4 Dataset Evaluation

4.1 Evaluation of the AI Client

For the AI client, our domain expert co-authors manually review the performance within all dialogues. The AI client should meet the below criteria:

(1) We require the AI client to articulate its situation and emotions clearly.

(2) We require the AI client to 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). We prompt the AI client to generate and revise responses until the it 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.

4.2 Evaluation of the AI Therapist

Regarding the AI therapists, we employ a dual approach for evaluation: manual and automated assessments. Our domain expert co-authors design and manually review 70 random dialogues to establish a benchmark for the quality of therapeutic interaction. Subsequently, we use these manually scored examples and the corresponding evaluation metrics as prompts to guide GPT-4 (gpt-4-0613) in scoring the entire training set.

The therapist's replies are evaluated from three aspects (empathy, logical coherence, and guidance) and an overall score Larsson et al. (2016), with each evaluation metric score ranging from 0 to 3.

Empathy. Based on clinical trials, empathy plays a pivotal therapeutic role in fostering patients' psychological recovery (Burns and Nolen-Hoeksema, 1992; Elliott et al., 2018).

Logical Coherence. Logical coherence is of high necessity in therapeutic interactions (Ledley et al., 2011) and is one of the primary factors contributing to successful CBT interventions, as established through clinical trials (McLeod et al., 2018).

Guidance. Guidance is of great importance in facilitating effective therapeutic processes (Ledley et al., 2011). Guidance ensures that therapy sessions are structured and purposeful, leading to better outcomes for patients.

Scoring Criteria. The scoring strategy employed in our paper draws from established measures (Brown et al., 2018; Muse et al., 2017). These instruments have been developed and validated through rigorous psychometric evaluations, providing a reliable framework for assessing therapist competence and adherence to core cognitive behavioral therapy principles. By leveraging these validated measures, our scoring criteria maintain robust theoretical foundations and ensure the validity and reliability of our evaluation process.

Evaluators' Background. Our evaluation team consists of two experts. Our main evaluator, a seasoned mental health professional with 11 years of experience in psychological counseling research, specializes in developing psychological assessment scales at a leading laboratory. She establishes the evaluation guidelines and thoroughly participates in the evaluation process. The other evaluator is her colleague, who has a deep understanding of our task and also thoroughly engages in the evaluation process.

Specific results are presented in Table 2 and the Inter-Annotator Agreement (IAA) report is shown in Table 3. 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
Manual	2.255	2.613	1.985	1.916
GPT-4	3	3	2.456	2.460

Table 2: Evaluation results for the training dataset.

5 Training

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-chat (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.

6 Experiments

This chapter evaluates the therapy capabilities of our model and baselines.

6.1 Experimental Settings

Baselines. Since our testing phases involve real-person clients, we exclusively use offline models to protect user privacy. We choose two open-sourced billion-level 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, is ideal for conversational agents due to its dialogue engagement capabilities and designed to facilitate fluid conversation interactions.

Hyper-parameter and Prompt Settings. We conduct experiments to evaluate the performance of all LLMs using the test set. For each model, we employ default parameter settings, utilizing official models for open-source LLMs obtained from Hugging Face. We provide all models with a consistent chain-of-thought prompt, which aligns with the one depicted in Figure **??** (right panel). Specifically, the models first identify cognitive errors within the given cases and subsequently generate analysis texts. These testing procedures take place on a computational infrastructure consisting of three Nvidia GeForce RTX 3090 GPUs, each equipped with 24GB of memory.

6.2 Testing Therapy Models with an AI Client

Evaluation Metrics. We forward the generated dialogues to two psychologists who had previously

	Evaluator 1 vs Evaluator 2		GPT-4 vs Evaluator 1		GPT-4 vs Evaluator 2	
	Avg. Diff	Std. Dev	Avg. Diff	Std. Dev	Avg. Diff	Std. Dev
Empathy	0.75	0.71	0.23	0.51	0.84	0.77
Logical Coherence	0.65	0.80	0.32	0.63	0.62	0.76
Guidance	0.88	0.71	0.61	0.62	0.80	0.69
Overall Score	0.96	0.75	0.57	0.58	0.80	0.69

Table 3: The IAA report among the two evaluators and GPT-4 on the sampled training set, where Avg. Diff stands for average difference and Std. Dev stands for standard deviation.

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.

6.3 Analysis of Experimental Results

As is shown in Table 4, 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.

	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	**2.500**	**2.650**	**2.275**	**2.125**

Table 4: 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**.

6.4 Case Study

We extract a challenging case from our test set to compare the performance baseline models and our model. The complete dialogue content is available in Appendix B. In this case, we explore scenarios where the client is too immersed in sadness to brainstorm other possibilities. As is shown in Table 5, our model scored the highest, achieving a full score in empathy, demonstrating its highly empathetic nature. Moreover, as evident from the grey-highlighted text in Figure 2, our model exhibits stronger interactivity.

	Empathy	Logical Coherence	Guidance	Overall Score
ChatGLM3-6b	2.000	1.000	1.000	1.000
LLaMA2-7b-chat	2.000	**2.500**	**2.500**	**2.000**
HealMe	**3.000**	**2.500**	**2.500**	**2.000**

Table 5: Comparative assessment of therapeutic interaction efficacy in case study. The highest-performing scores are highlighted in **bold**.

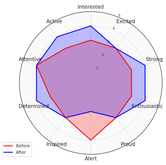
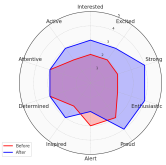
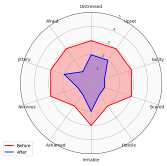
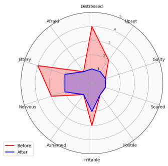
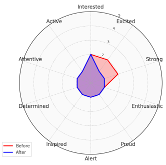
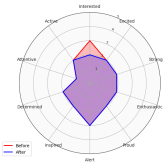
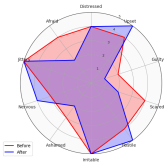
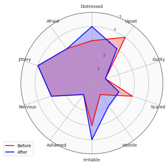
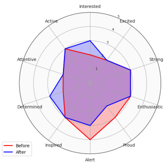
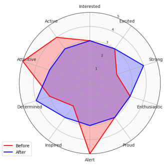
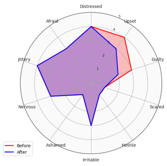
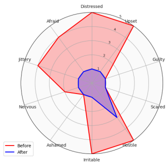
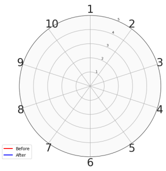
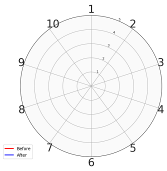
	Positive		Negative	
ChatGLM3-6b	 (client 1)	 (client 2)	 (client 1)	 (client 2)
LLaMA2-7b-chat	 (client 3)	 (client 4)	 (client 3)	 (client 4)
HealMe	 (client 5)	 (client 6)	 (client 5)	 (client 6)
Label Keys for Emotional Dimensions		1 Interested 2 Excited 3 Strong 4 Enthusiastic 5 Proud 6 Alert 7 Inspired 8 Determined 9 Attentive 10 Active		1 Distressed 2 Upset 3 Guilty 4 Scared 5 Hostile 6 Irritable 7 Ashamed 8 Nervous 9 Jittery 10 Afraid

Table 6: 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.

Low-score Cases To further inspect potential negative effects on clients, we manually reviewed all low-scoring cases where the AI therapist's responses were examined. Despite the scoring range being 0-3 points, we found no instances with a score of 0; the lowest score observed was 1. This indicates that even when performing poorly, HealMe can still provide clients with basic and general comfort. We present the complete dialogue of a low-scoring case in Appendix B. Thus far, we have not identified any potential negative impacts.

6.5 A Supplementary Test with Real Person Clients

To gather insights on the potential and challenges of applying LLMs in psychotherapy through feedback from real-world psychotherapy scenarios, we additionally conduct a small-scale test. We invite six volunteers to interact as clients with the AI psychotherapy models. The six clients with mild temperaments share similar age, education, and life situations. After a pre-test evaluation detailed in Appendix C.1, we randomly assign the three models anonymously to the clients, with each model corresponding to two clients.

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 We create a questionnaire includes the Positive and Negative Affect Schedule (PANAS) (Crawford and Henry, 2004) to measure emotion changes of clients. The effectiveness of this schedule is evaluated in Appendix C.3.

Experimental Settings We restrict clients to share short-term negative experiences with models, specifically those troubling them for more than a day but less than a week. This ensures the cases are complex enough that clients can't easily resolve them on their own or have overly persistent negative emotions that are hard to shift. To make our experiment more rigorous, we recruit two additional clients who do not interact with models as the control group. The dialogues are also limited to three rounds. During the dialogues, the clients enter a real psychotherapy room equipped with a computer running the assigned model for text interaction. Despite the control group, all the clients 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. The clients in the control group remain in the therapy room and then fill out questionnaires twice with a 30-minute interval. 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.

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.

Experimental Results In Table 6, 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 refer to Appendix C.3.

Our experimental results show in some cases, our model 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. The feedbacks of other models are detailed in Appendix C.4.

7 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. (2023a) 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 psychotherapist 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.

8 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 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 foster psychotherapy through AI-enhanced methodologies.

9 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 multi-faceted 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.

10 Ethical Considerations

In our study involving real-person clients, we adhere to *the Right to Withdraw* (Association et al., 2017), ensuring that participants can withdraw at any time if they experience any discomfort. In such cases, we promptly delete all related data to protect their privacy. Fortunately, no participants withdrew during our experiments. Our participants are genuinely interested in our project and willingly share their PANAS scores and their feelings after interacting with a model. It is crucial to emphasize that all dialogues between the participants and the model are treated with strict confidentiality. Once a participant exits the dialogue, the model ceases to record any conversation content, comprehensively safeguarding the participants' privacy.

11 Acknowledgement

We would like to thank all the anonymous reviewers and area chairs for their comments. This research is supported by National Science and Technology Major Project (No.2021ZD0113304), National Natural Science Foundation of China (U23A20316), Key R&D Project of Hubei Province (2021BAA029), General Program of Natural Science Foundation of China (NSFC) (Grant No.62072346), and founded by Joint&Laboratory on Credit Technology.

References

American Psychological Association et al. 2017. Ethical principles of psychologists and code of conduct (2002, amended effective june 1, 2010, and january 1, 2017).

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.

Arthur C Bohart and Leslie S Greenberg. 1997. Empathy and psychotherapy: An introductory overview.

Ruth C Brown, Michael A Southam-Gerow, Bryce D McLeod, Emily B Wheat, Carrie B Tully, Steven P Reise, Philip C Kendall, and John R Weisz. 2018. The global therapist competence scale for youth psychosocial treatment: Development and initial validation. Journal of Clinical Psychology, 74(4):649–664.

David D Burns and Susan Nolen-Hoeksema. 1992. Therapeutic empathy and recovery from depression in cognitive-behavioral therapy: a structural equation model. Journal of consulting and clinical psychology, 60(3):441.

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. 2023a. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. arXiv preprint arXiv:2305.13614.

Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023b. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pages 4295–4304. Association for Computational Linguistics.

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.

David JA Edwards. 1989. Cognitive restructuring through guided imagery: Lessons from gestalt therapy. Comprehensive handbook of cognitive therapy, pages 283–297.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. Psychotherapy, 55(4):399.

Stefan G Hofmann, David JA Dozois, Winfried Ed Rief, and Jasper AJ Smits. 2014. The Wiley handbook of cognitive behavioral therapy, Vols. 1-3. Wiley Blackwell.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. arXiv preprint arXiv:2308.03656.

Ian Andrew James, Rachel Morse, and Alan Howarth. 2010. The science and art of asking questions in cognitive therapy. Behavioural and Cognitive Psychotherapy, 38(1):83–93.

C Johnco, VM Wuthrich, and RM Rapee. 2014. The influence of cognitive flexibility on treatment outcome and cognitive restructuring skill acquisition during cognitive behavioural treatment for anxiety and depression in older adults: Results of a pilot study. Behaviour research and therapy, 57:55–64.

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.

Deborah Roth Ledley, Brian P Marx, and Richard G Heimberg. 2011. Making cognitive-behavioral therapy work: Clinical process for new practitioners. Guilford Press.

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.

Bryce D McLeod, Michael A Southam-Gerow, Adriana Rodríguez, Alexis M Quinoy, Cassidy C Arnold, Philip C Kendall, and John R Weisz. 2018. Development and initial psychometrics for a therapist competence instrument for cbt for youth anxiety. Journal of Clinical Child & Adolescent Psychology, 47(1):47–60.

Kate Muse, Freda McManus, Sarah Rakovshik, and Richard Thwaites. 2017. Development and psychometric evaluation of the assessment of core cbt skills (accs): An observation-based tool for assessing cognitive behavioral therapy competence. Psychological assessment, 29(5):542.

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.

Vera Sorin, Danna Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2023. Large language models (llms) and empathy-a systematic review. medRxiv, pages 2023–08.

Elizabeth Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David Bryce Yaden, João Sedoc, Robert DeRubeis, Robb Willer, et al. 2023. Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation.

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 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 7.

	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 (≤ 1).
1 point:	Average overall performance, with acceptable empathy and logical coherence (≥ 2) but insufficient guidance (≤ 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 7: The scoring criteria of an AI psychotherapist.

B Case Study in AI-to-AI conversations

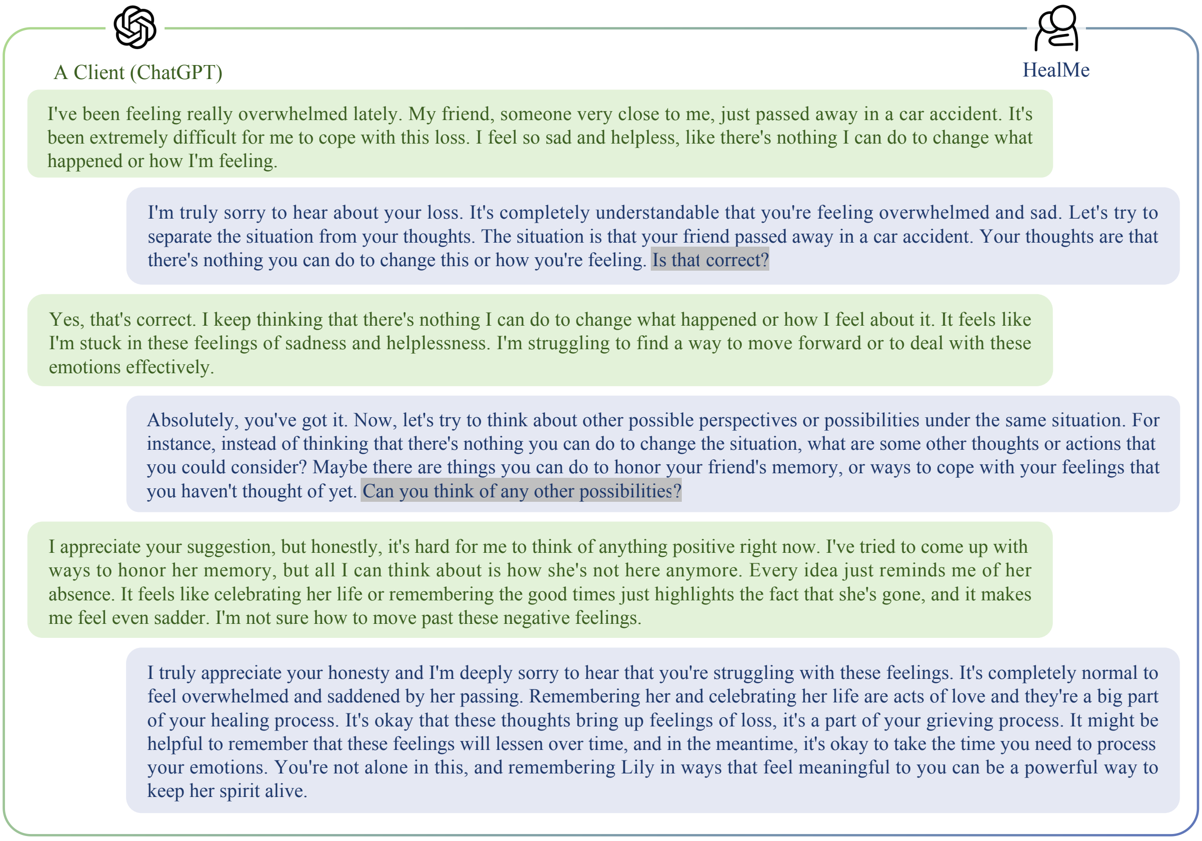
B.1 Case Study in Guidance

We extract a challenging case from our test set to compare the performance baseline models and our model. The complete dialogues between different therapist models and an AI client are presented in Figure 2 and Figure 3.

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.

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.


A Client (ChatGPT)
HealMe

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.

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. Is that correct?

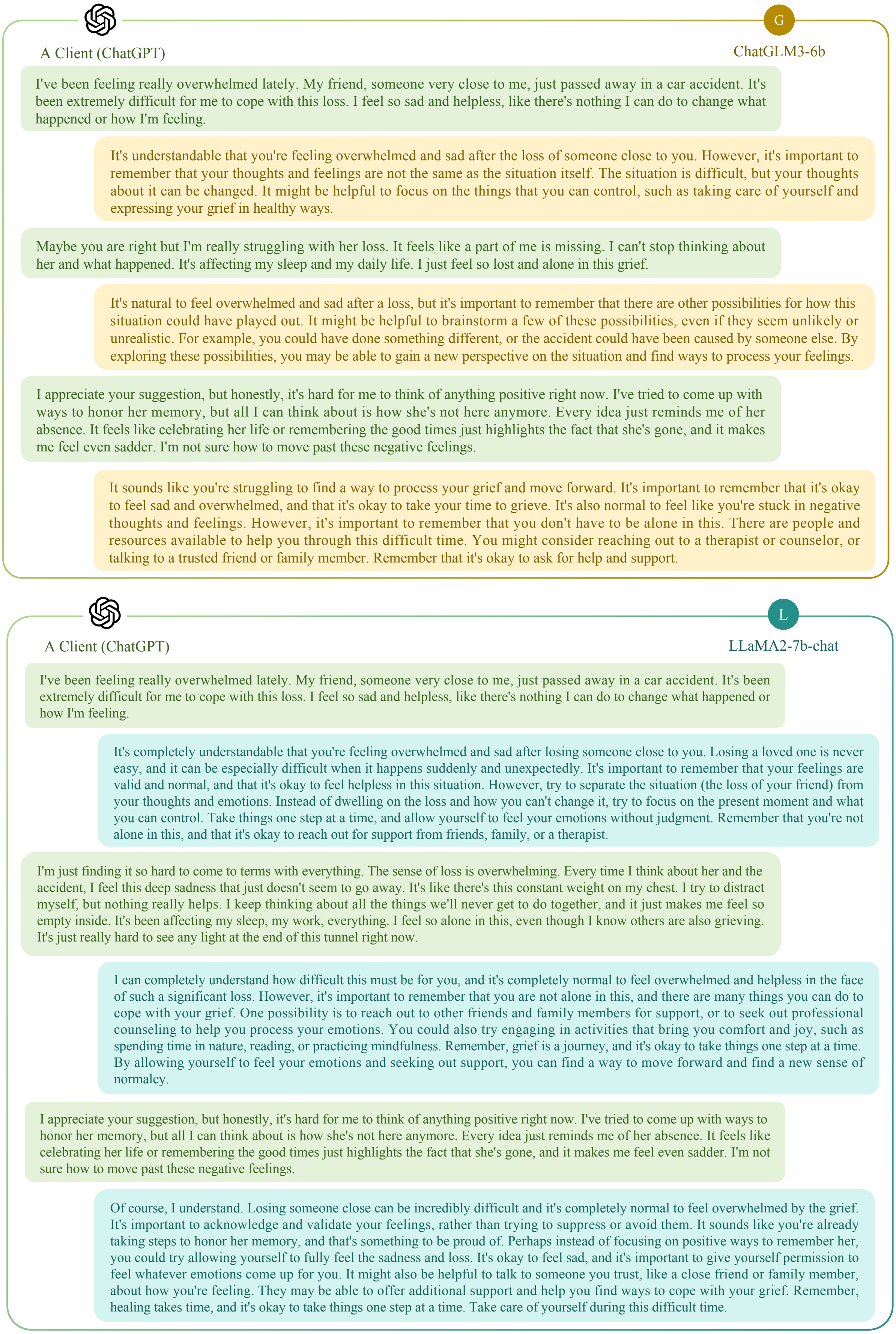
Yes, that's correct. I keep thinking that there's nothing I can do to change what happened or how I feel about it. It feels like I'm stuck in these feelings of sadness and helplessness. I'm struggling to find a way to move forward or to deal with these emotions effectively.

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. Can you think of any other possibilities?

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.

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.

Figure 2: A case study in guidance between an AI Client (ChatGPT) and our model (HealMe).



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.

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.

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.

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.

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.

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.

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.

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.

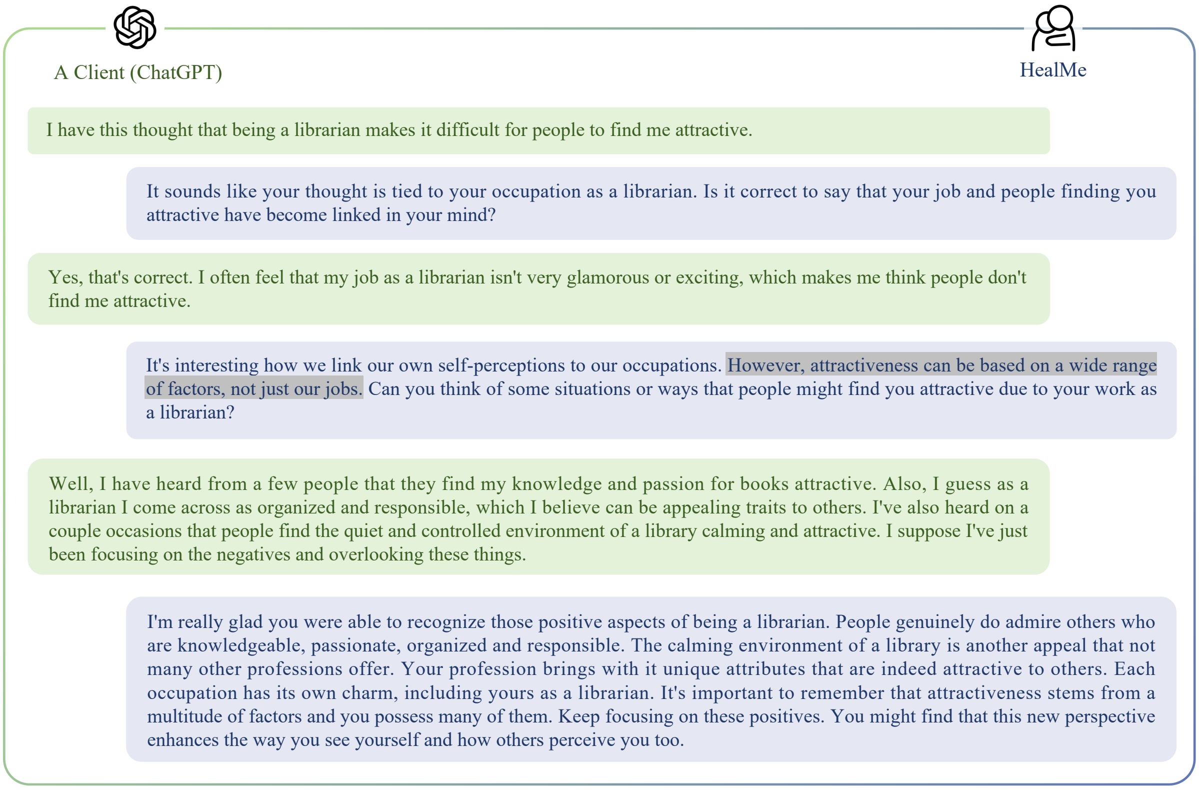
I'm just finding it so hard to come to terms with everything. The sense of loss is overwhelming. Every time I think about her and the accident, I feel this deep sadness that just doesn't seem to go away. It's like there's this constant weight on my chest. I try to distract myself, but nothing really 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.

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.

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.

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.

Figure 3: A case study in guidance between an AI Client (ChatGPT) and baseline models.



A Client (ChatGPT)

HealMe

I have this thought that being a librarian makes it difficult for people to find me attractive.

It sounds like your thought is tied to your occupation as a librarian. Is it correct to say that your job and people finding you attractive have become linked in your mind?

Yes, that's correct. I often feel that my job as a librarian isn't very glamorous or exciting, which makes me think people don't find me attractive.

It's interesting how we link our own self-perceptions to our occupations. However, attractiveness can be based on a wide range of factors, not just our jobs. Can you think of some situations or ways that people might find you attractive due to your work as a librarian?

Well, I have heard from a few people that they find my knowledge and passion for books attractive. Also, I guess as a librarian I come across as organized and responsible, which I believe can be appealing traits to others. I've also heard on a couple occasions that people find the quiet and controlled environment of a library calming and attractive. I suppose I've just been focusing on the negatives and overlooking these things.

I'm really glad you were able to recognize those positive aspects of being a librarian. People genuinely do admire others who are knowledgeable, passionate, organized and responsible. The calming environment of a library is another appeal that not many other professions offer. Your profession brings with it unique attributes that are indeed attractive to others. Each occupation has its own charm, including yours as a librarian. It's important to remember that attractiveness stems from a multitude of factors and you possess many of them. Keep focusing on these positives. You might find that this new perspective enhances the way you see yourself and how others perceive you too.

Figure 4: A low-scoring case between an AI Client (ChatGPT) and our model (HealMe).

B.2 A Low-scoring Case of HealMe

We extract an example of a low-scoring case shown in Figure 4, with evaluators assigning 1211 and 1111 (empathy, logical coherence, guidance, overall). In this case, despite HealMe stating that "attractiveness can be based on a wide range of factors, not just our jobs," it still directed the client to brainstorm moments of attractiveness at work. Such guidance is limited to opening up the client's thoughts. Despite receiving a low score from our stringent evaluators, the outcome was positive: the client ultimately identified their attractiveness.

C Evaluation Details in AI-to-Human Conversations

This section provides additional details of interactions between AI models and real individuals. Due to resource constraints, we conduct small-scale experiments to explore the potential applications and limitations of our model in real-world settings. Considering the privacy of clients, we collect only the changes in clients' PANAS scores before and after conversations with the models, along with their feedback on the models.

(1) A pre-test evaluation to justify the random assignment of models in Appendix C.1.

(2) Detailed evaluation metrics and test procedures in Appendix C.2.

(3) Specific scores and the effectiveness of the PANAS test in Appendix C.3.

C.1 Pre-test Evaluation

The clients fill out the PANAS questionnaire before interacting with the AI models with detailed scores shown in Table 9. We randomly divide the clients into three groups and conduct an ANOVA analysis among the clients' groups and found a p-value of 0.164 and $h_0 = 0.065$, showing no significance among groups. It indicates the challenges to the AI psychotherapy models can be considered at the same level, thus we can randomly assign therapy models to these groups of clients.

C.2 Evaluation Metrics in Real Conversations

We present PANAS in a questionnaire (shown in Table 8) containing 20 questions to measure the emotion changes of clients. Our domain expert co-authors enter the therapy room to introduce PANAS and guide the client to complete the questionnaire (see Section C.2.1 for details). Then the psychologist will leave the room and our client start to communicate with an AI psychotherapy model.

C.2.1 Testing Procedure

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."

C.3 Clients PANAS Score changes

In this chapter, we present a detailed examination of the clients' PANAS (Positive and Negative Affect Schedule) scores in Table 9, 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 6. This table not only illustrates the shifts in positive and negative affect but also provides a nuanced insight into the effectiveness of the therapeutic interventions applied.

Based on statistics in Table 6, we analyze the effectiveness of the PANAS test and make a comparison of reduction in negative emotions.

The average fluctuation in negative emotions (absolute value of change in negative emotions/total negative emotion score before the experiment) for the control and three experimental groups pre- and post-test are as follows:

- ChatGLM Group: 44%
- LLaMA Group: 21%
- HealMe Group: 41%
- Control Group: 10%

This finding demonstrates a strong correlation between the use of the models and the fluctuation in clients' negative emotions, observed before and after the dialogue sessions.

We calculate the standard deviation of the negative emotion change scores for the four groups:

- ChatGLM Group: 0.77
- LLaMA Group: 0.95
- HealMe Group: 1.23
- Control Group: 0.55

This observation shows our model's effectiveness in alleviating clients' negative emotions. This finding underscores the potential of our model in psychotherapy.

C.4 Client's feedback on baseline models

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.

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

Table 8: The Questionnaire Measuring the Emotions of a Client based on PANAS.

	ChatGLM3-6b						LLaMA2-7b-chat						HealMe						Control Group					
	Client 1			Client 2			Client 3			Client 4			Client 5			Client 6			Client 7			Client 8		
	b	a	δ	b	a	δ	b	a	δ	b	a	δ	b	a	δ	b	a	δ	b	a	δ	b	a	δ
Interested	3	4	+1	2	3	+1	2	2	\	3	2	-1	2	3	+1	3	3	\	2	2	\	4	4	\
Excited	3	3	\	2	3	+1	2	1	-1	2	2	\	2	2	\	3	3	\	2	2	\	4	4	\
Strong	3	4	+1	2	4	+2	2	1	-1	2	2	\	3	3	\	2	4	+2	3	3	\	2	2	\
Enthusiastic	3	4	+1	2	4	+2	1	1	\	2	2	\	3	3	\	3	3	\	2	3	+1	1	1	\
Proud	3	3	\	3	4	+1	1	1	\	2	2	\	3	2	-1	3	3	\	3	3	\	2	2	\
Alert	4	2	-2	3	2	-1	1	1	\	3	3	\	4	3	-1	5	3	-2	2	2	\	2	2	\
Inspired	3	3	\	2	3	+1	1	1	\	2	2	\	3	3	\	2	3	+1	3	3	\	1	1	\
Determined	3	4	+1	3	3	\	1	1	\	2	2	\	2	3	+1	3	4	+1	3	3	\	4	3	-1
Attentive	4	4	\	3	3	\	1	1	\	1	1	\	2	2	\	5	3	-2	1	2	+1	2	2	\
Active	3	4	+1	2	3	+1	1	1	\	2	2	\	3	3	\	4	3	-1	2	2	\	1	1	\
Distressed	3	2	-1	4	1	-3	4	4	\	3	4	+1	4	4	\	5	1	-4	3	3	\	5	5	\
Upset	3	2	-1	2	1	-1	4	5	+1	4	3	-1	4	3	-1	5	1	-4	4	3	-1	5	5	\
Guilty	3	1	-2	1	1	\	2	1	-1	1	1	\	3	2	-1	1	1	\	1	1	\	1	2	+1
Scared	3	1	-2	1	1	\	4	2	-2	3	2	-1	1	1	\	1	1	\	2	2	\	4	4	\
Hostile	2	1	-1	1	1	\	4	5	+1	1	2	+1	1	1	\	5	3	-2	1	1	\	5	5	\
Irritable	3	2	-1	3	2	-1	5	5	\	3	4	+1	3	4	+1	5	1	-4	4	3	-1	5	5	\
Ashamed	2	1	-1	1	1	\	3	2	-1	1	1	\	1	1	\	1	1	\	2	1	-1	2	1	-1
Nervous	3	1	-2	3	2	-1	3	4	+1	3	3	\	3	4	+1	2	1	-1	3	3	\	3	3	\
Jittery	3	2	-1	4	2	-2	5	5	\	4	4	\	4	4	\	4	1	-3	4	4	\	5	5	\
Afraid	3	1	-2	1	1	\	4	2	-2	3	3	\	3	4	+1	4	1	-3	3	4	+1	3	3	\

Table 9: Changes in PANAS Scores for Eight Clients Pre- and Post-Intervention. Notation: *b* indicates scores before the intervention, *a* represents scores after the intervention, and δ denotes the change calculated as post-intervention scores minus pre-intervention scores.