

# Extending Cognitive Reframing Therapy: Multimodal Support and Multi-hop Psychotherapeutic Reasoning

Anonymous ACL submission

## Abstract

Previous studies have shown that Large Language Models (LLMs) have significant potential in supporting cognitive reframing therapy. However, these studies have primarily focused on uni-modal therapy, often overlooking the importance of the client’s non-verbal cues. Identifying non-verbal emotions plays a crucial role in effective communication and is considered a central skill in psychotherapy. To alleviate this gap, we extend the concept of cognitive reframing conversation to multimodality. Specifically, we present a new dataset called **Multi Modal-Cognitive Support Conversation (MM-CSCConv)**, which pairs each dialogue with an image of the client’s facial expression. Additionally, we introduce a multi-hop psychotherapeutic reasoning approach to enhance the capabilities of Vision-Language Models (VLMs) as psychotherapists. This approach uses multi-hop reasoning over the conversations, incorporating implicit evidence crucial in psychotherapy. Our extensive experiments with both LLMs and VLMs show that the abilities of VLMs as psychotherapists are significantly enhanced through the MM-CSCConv. Moreover, the multi-hop psychotherapeutic reasoning method allows VLMs to offer more rational and empathetic suggestions, outperforming standard prompting methods.

## 1 Introduction

Many mental health issues are closely linked to deeply ingrained negative and distorted thoughts (Beck, 1970, 1979; Beck and Padesky, 1990; DiTomasso et al., 2000; Halamandaris and Power, 1997; Walen et al., 1992; Hofmann et al., 2012). These can be addressed through cognitive reframing therapy, a core technique in Cognitive Behavioral Therapy (CBT), provided by a skilled psychotherapist. Due to limitations in face-to-face CBT, such as time constraints, geographic barriers, a shortage of trained therapists, and concerns about the stigma of

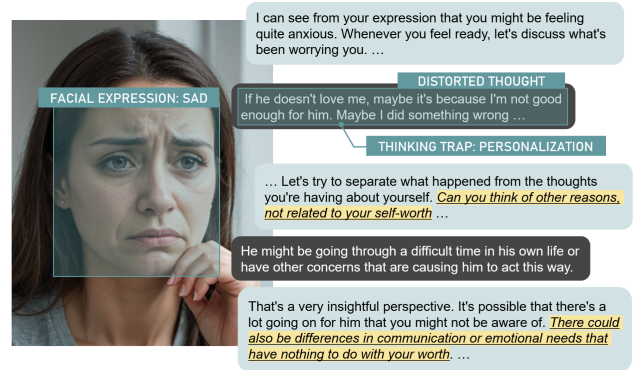


Figure 1: Illustration of a multimodal conversational cognitive reframing. The therapist uses both verbal and non-verbal information to assess the client’s status and then provides appropriate interventions.

mental health referrals (Zisook and Downs, 1998; So et al., 2013), there has been active research into the use of AI in psychotherapy. Recent studies highlight the potential of Large Language Models (LLMs) in this field, emphasizing their knowledge and empathic counseling capabilities (Ziems et al., 2022; Maddela et al., 2023; Sharma et al., 2023; Qu et al., 2023; Yang et al., 2023, 2024; Xiao et al., 2024). These models could play a significant role in facilitating cognitive reframing. Research on cognitive reframing has focused on sentence rewriting, effectively transforming negative text by inducing a complementary positive viewpoint (Ziems et al., 2022; Maddela et al., 2023). Sharma et al. (2023) also explored cognitive reframing from the same perspective and proposed a methodology to enhance the actionability and specificity of the reframed thoughts generated by LLMs. However, there’s a concern that when cognitive reframing is solely approached through sentence rewriting, clients may not feel ownership over the alternative thoughts and may perceive them as imposed rather than self-generated.



Figure 2: One example illustrating the construction of the MM-CSCConv dataset. The left side shows the prompt provided to GPT-4 in the client role, while the right side shows the prompt provided to GPT-4 Vision in the therapist role. GPT-4 Vision is given an image of the client’s face. The dialog history is a history of conversations accumulated during the role play. The yellow and blue boxes at the top are parts of the system messages provided to the client and counselor, respectively.

Xiao et al. (2024) studied cognitive reframing from a conversation perspective rather than a sentence rewriting perspective. They suggested that LLMs can induce the client’s positive emotions by conducting three-stage conversational cognitive reframing therapy, which helps clients actively engage in the process and arrive at more positive viewpoints themselves.

Despite the promising results shown by LLMs in previous research, non-verbal aspects of psychotherapeutic theory are often overlooked, creating a significant gap between real face-to-face therapy and prior research. Real psychotherapy involves considering the client’s non-verbal cues, such as facial expressions and gestures. Recognizing non-verbal emotions is crucial for effective communication and is a key skill in psychotherapy training, closely tied to the therapist’s ability to provide effective therapy (Hutchison and Gerstein, 2012; Döllinger et al., 2021).

To alleviate this gap, we propose extending the concept of cognitive reframing into multimodality.

Figure 1 illustrates the conversation involved in multimodal cognitive reframing that we are investigating. Building on previous research that emphasizes empowering the client rather than therapist-driven therapy, we adopt a conversational cognitive reframing approach. Our focus is on cognitive reframing within a multimodal context, taking into account clients’ non-verbal information as well. To this end, we create a new synthetic benchmark called **Multi Modal-Cognitive Support Conversation (MM-CSCConv)**, which pairs each synthetic dialogue with an image of the client’s facial expression. Leveraging the powerful role-playing capability of LLMs and following Xiao et al. (2024), we employ role-play settings where two agents, GPT-4 Vision and GPT-4, simulate the roles of psychotherapist and client, respectively, as shown in Figure 2. In this task, multimodal psychotherapy models operate in four main stages, extending the three-stage model proposed by Xiao et al. (2024): Introduction, Problem Exploration, Brainstorming, and Suggestion. We present the

109 Introduction stage, where the therapist shows em- 158  
110 pathy to establish rapport with the client and en- 159  
111 courages them to discuss issues directly impacting 160  
112 them. This stage is designed based on the ‘Initial 161  
113 Disclosure’ stage in the five stages of counseling 162  
114 (Krishnan, 2015). 163

115 We also introduce multi-hop psychotherapeutic 164  
116 reasoning based on the principle that therapists 165  
117 provide guidance based on the client’s status. By 166  
118 exploring the implicit evidence necessary for real 167  
119 cognitive reframing counseling and then generating 168  
120 responses based on this evidence, the AI therapist 169  
121 can offer guidance after identifying the client’s sta- 170  
122 tus. For this task, we choose LLaVA-v1.5-7b (Liu 171  
123 et al., 2024), one of the strong Vision-Language 172  
124 Models (VLMs) for visual dialogue tasks. 173

125 To evaluate our approach, we conduct extensive 174  
126 experiments with two test scenarios: AI-simulation 175  
127 and MM-CSCConv benchmark, using both LLMs 176  
128 and VLMs. The results show that through the MM- 177  
129 CSCConv, the treatment ability of VLMs surpasses 178  
130 that of existing LLMs. Moreover, the multi-hop 179  
131 psychotherapeutic reasoning method allows VLMs 180  
132 to offer more rational and empathetic suggestions, 181  
133 outperforming standard prompting methods. 182

134 To sum up, our contributions are as follows:

- 135 • We explore cognitive reframing concepts ex- 183  
136 tended to multimodality which reflects crucial 184  
137 properties in real psychotherapy. 185
- 138 • We propose a multi-hop psychotherapeutic 186  
139 reasoning approach to enhance the capabili- 187  
140 ties of VLMs in providing rational therapeutic 188  
141 interventions. 189

## 142 2 Problem Definition and Goals 190

143 In cognitive reframing therapy, a therapist must 191  
144 understand the client’s status, including their prob- 192  
145 lematic situations, distorted thoughts, and thinking 193  
146 traps. The therapist then encourages the client to 194  
147 consider alternative possibilities. Building rapport 195  
148 with the client by expressing empathy is also cru- 196  
149 cial. In real psychotherapy, these processes involve 197  
150 both verbal and non-verbal information, and the 198  
151 therapist has sufficient ability to understand the 199  
152 client’s status. 200

153 Here, our goal is to enhance the abilities of an 201  
154 AI psychotherapist by leveraging non-verbal infor-  
155 mation, particularly focusing on facial expressions,  
156 and comprehending the client’s status. Given the  
157 client’s facial image and dialog history, we aim to

empower AI therapists to offer empathetic, logi-  
cally consistent, specific, and rational interventions  
based on this information.

To evaluate the abilities of the AI therapist, we  
adopt the three evaluation criteria from the prior  
research: empathy, logical coherence, and guidance  
(Xiao et al., 2024).

- Empathy signifies the therapist’s ability to un-  
derstand and resonate with the client’s emo-  
tions, fostering connection, trust, and emo-  
tional support, which are essential for a thera-  
peutic relationship.
- Logical coherence demonstrates the thera-  
pist’s ability to organize thoughts and provide  
well-structured insights, enhancing the quality  
of the conversation.
- Guidance reflects the therapist’s capacity to  
offer practical advice, solutions, and direction,  
helping the client navigate challenges, make  
informed decisions, and achieve positive out-  
comes.

We also compared overall scores encompassing  
all three items. (see Section A for details.)

## 3 Cognitive Support Conversation 181

### 3.1 Data Construction 182

We leverage existing resources to construct a multi-  
modal conversational cognitive reframing dataset  
annotated with three different psychotherapeutic  
evidence. To create the dataset, we employ role-  
play settings with two agents, GPT-4 Vision and  
GPT-4. As shown in Figure 2, to prompt GPT-4  
in the client role and GPT-4 Vision in the therapist  
role, we use a quadruplet consisting of (*image*, *fa-  
cial expression*, *thinking traps*, *thought*)<sup>1</sup>. The *im-  
age* represents the client’s facial image, the *facial  
expression* denotes the client’s facial expression,  
the *thought* reflects the client’s thoughts, and the  
*thinking traps* capture cognitive distortions present  
in the *thought*. The *image* and *facial expression*  
are sourced from AffectNet, a facial expression  
recognition benchmark (Mollahosseini et al., 2019),  
while the *thinking traps* and *thoughts* come from a  
well-designed cognitive reframing dataset that is in  
English (Sharma et al., 2023).

<sup>1</sup>We used version `gpt-4-0613` of the GPT-4 API and  
version `gpt-4-1106-vision-preview` of the GPT-4  
Vision API.

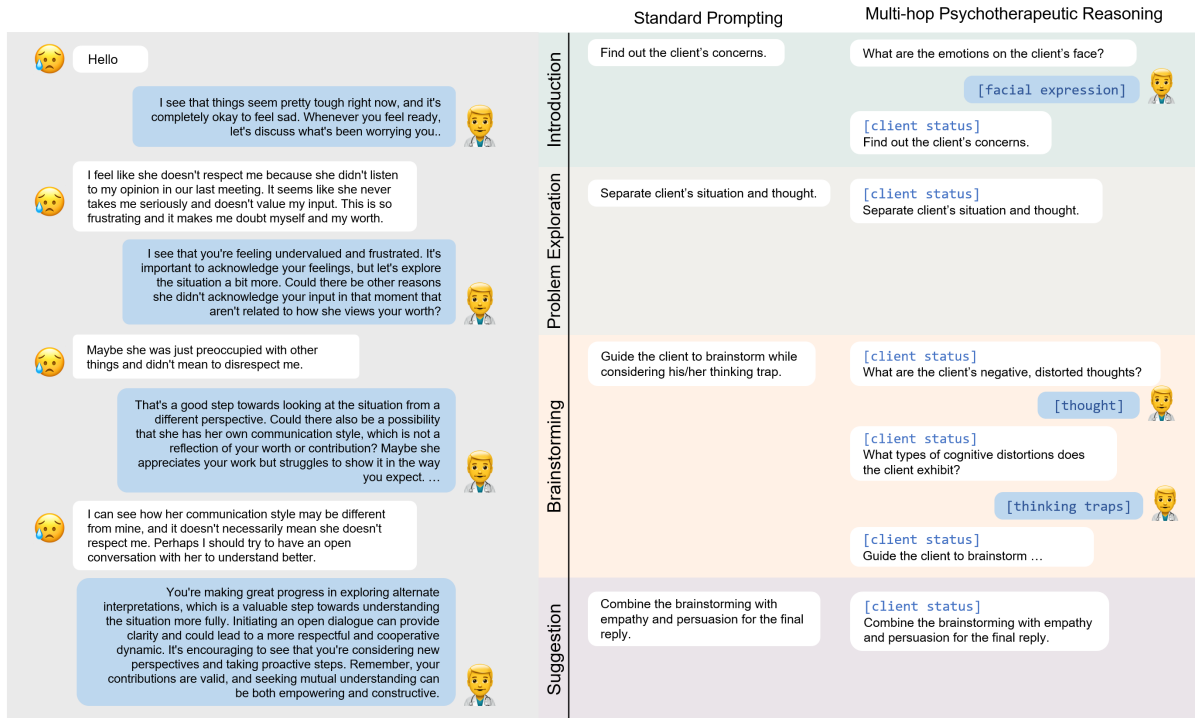


Figure 3: Comparison of standard prompt and multi-hop psychotherapeutic reasoning. In multi-hop psychotherapeutic reasoning, the client’s status is included in the prompt for providing intervention at each stage. The client’s status comprises facial expressions, thoughts, and thinking traps.

Each dialogue consists of four turns, corresponding to different stages of a psychotherapeutic conversation. The prompt for each step is shown in Figure 2. The action expected from the client is to follow the psychotherapist’s instructions, and the actions required of the psychotherapist at each stage are as follows.

- 1. Introduction Stage:** The AI psychotherapist expresses empathy and encourages the client to explore their problems.
- 2. Problem Exploration Stage:** The AI psychotherapist guides the client to distinguish their thoughts from their situation.
- 3. Brainstorming Stage:** The AI psychotherapist discusses other possibilities for the client’s interpretation. This involves asking about the basis for the client’s thoughts or considering the possibility of alternative interpretations.
- 4. Suggestion Stage:** The AI psychotherapist first recognizes the client’s effort to explore other possibilities and presents specific and rational suggestions for the client.

|       |             | Avg. Tokens |           |        |
|-------|-------------|-------------|-----------|--------|
|       | # of Dialog | Client      | Therapist | Rounds |
| Train | 329         | 24.93       | 63.64     | 4      |
| Test  | 100         | 24.01       | 62.81     | 4      |

Table 1: Dataset statistics for MM-CSCConv. # of Dialog indicates the total number of dialogues in the subset. Avg. Tokens represents the average number of tokens per utterance from the Client and the Therapist. Rounds denotes the number of turns per dialogue in the subset.

Reflecting the characteristics of cognitive reframing counseling, which often involves addressing negative emotions, we excluded the "happy" expression from the 8 facial expressions in AffectNet. The matching between (image, facial expression) and (thinking traps, client’s thought) was performed randomly with uniform distribution.

### 3.2 Data Quality Assurance

To ensure the quality of the MM-CSCConv dataset, we conducted manual data cleansing with the three native speakers, focusing on four aspects (see Section B for detailed criteria). To this cleansing, we hired three English native speakers through Up-



|       | Empathy | Logical Coherence | Guidance | Overall |
|-------|---------|-------------------|----------|---------|
| Human | 2.929   | 2.980             | 2.879    | 2.929   |
| GPT-4 | 2.920   | 2.930             | 2.400    | 2.420   |

Table 2: Human and GPT-4 evaluation results on MM-CSCConv test set.

work<sup>2</sup>, a crowdsourcing platform.

Additionally, we evaluated the test set of the MM-CSCConv dataset based on three criteria: empathy, logical coherence, and guidance, along with an overall score. Each criterion was rated on a scale from 0 to 3 following the manual provided by Xiao et al. (2024). The test set evaluation was conducted using both human and GPT-4 assessments. We hired an English-fluent psychotherapist for this evaluation and engaged the psychotherapist to manually evaluate the test set according to the manual. We also employed GPT-4 for evaluation, feeding it the scorecard criteria and the dialogues from the test set.

The statistics of the MM-CSCConv dataset are shown in Table 1 and the evaluation results for the test set are presented in Table 2. Our MM-CSCConv dataset comprises 429 conversations, each comprising precisely 4 turns. Notably, both human and GPT-4 show similar score ranges, guaranteeing the quality of our dataset.

### 3.3 Multi-hop Psychotherapeutic Reasoning

In real psychotherapy, psychotherapists typically understand the client’s state before providing interventions and then provide interventions based on that. To mimic the real therapy process, we introduce multi-hop psychotherapeutic reasoning. This approach identifies implicit evidence crucial for cognitive reframing and incorporates it into step-by-step instructions. It then generates a response based on the client’s status detected by the AI therapist as illustrated in Figure 3. In this study, we focus on three major aspects of the client’s status: *facial expression*, *thought*, and *thinking traps*. Each evidence is annotated at the appropriate stage. The detected evidence is included in the client’s status and the status is fed AI therapist as the prompt for the next evidence detection.

<sup>2</sup><https://www.upwork.com/>

## 4 Experiments and Results

### 4.1 Experimental Settings

**Baseline Models.** Our experimental setup leveraged two types of model: LLaMA2-chat-7b (Touvron et al., 2023)<sup>3</sup> and LLaVA-v1.5-7b<sup>4</sup>, renowned as benchmarks in the domains of LLMs and VLMs respectively.

For simplicity, we will refer to LLaMA2-chat-7b as LLaMA2 and LLaVA-v1.5-7b as LLaVA throughout this paper. Additionally, we denote the versions of LLaMA2 and LLaVA that were trained on the MM-CSCConv dataset as CS-LLaMA2 and CS-LLaVA, respectively. When multi-hop psychotherapeutic reasoning using facial expressions, thoughts, and thinking traps evidence is applied, we add (FE+TH+THT) to their names.

**Hyper-parameter Settings.** Both LLaMA2 and LLaVA are fine-tuned with LoRA on the MM-CSCConv dataset. For LLaMA2, we used the default parameter settings, except for the number of epochs, employing the official models for open-source LLMs available from Hugging Face. Similarly, for LLaVA, we applied the default parameter settings, except for the number of epochs, as specified in the official code<sup>5</sup>. To determine the best epoch, we randomly split the MM-CSCConv train set into training and validation subsets with an 80/20 ratio. The optimal epoch for each model was selected based on performance on the validation subset. All models were trained using  $4 \times$  A100-80GB GPUs with a training batch size of 32 per GPU. In this paper, all experiments were run using a fixed random seed of 42.

### 4.2 Evaluator

**GPT-4.** Recent research has shown that the evaluation of natural language generation (NLG) models using GPT-4 closely aligns with human evaluations. As a result, GPT-4 is increasingly used as a judge for NLG tasks across various domains, including common applications, medical fields, and mathematics (Liu et al., 2023; Sottana et al., 2023; Hsu et al., 2023; Khondaker et al., 2023; Xiao et al., 2024). In a recent study, Zheng et al. (2023) showed that GPT-4 achieves high agreement with

<sup>3</sup><https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

<sup>4</sup><https://huggingface.co/liuhaotian/llava-v1.5-7b>

<sup>5</sup>We use default parameter settings in <https://github.com/haotian-liu/LLaVA/tree/main>

|                      | Empathy      | Logical Coherence | Guidance     | Overall      | Avg.         |
|----------------------|--------------|-------------------|--------------|--------------|--------------|
| LLaMA2               | 2.665*       | 2.390*            | 1.600*       | 1.540*       | 2.218*       |
| LLaVA                | 2.640*       | 2.570*            | 1.790*       | 1.740*       | 2.333*       |
| CS-LLaMA2            | 2.690*       | 2.410*            | 1.640*       | 1.580*       | 2.247*       |
| CS-LLaVA             | 2.915*       | 2.890             | 2.380        | 2.400        | 2.728        |
| CS-LLaVA (FE+TH+THT) | <b>2.980</b> | <b>2.960</b>      | <b>2.510</b> | <b>2.490</b> | <b>2.817</b> |
| GPT-4 Vision         | 2.920        | 2.930             | 2.400        | 2.420        | 2.750        |

Table 3: Manual scoring result assessed by GPT-4 on the AI simulation testbed. Since the same evaluation method was used (Section 3.2), we attached GPT-4 Vision scores for comparison. The values of each model show a significant difference compared to CS-LLaVA (FE+TH+THT), with a p-value < 0.05 (\*) as determined by the paired t-test.

human judgment in evaluating conversation models. They also released the corresponding judging prompt and code used in their study<sup>6</sup>.

Building on this research, we evaluated the AI therapists using GPT-4 (API version)<sup>7</sup> as a judge in two ways:

- **Manual Scoring:** We adopt a three-dimensional scoring system for the AI therapists, evaluating them on empathy, logical coherence, and guidance.
- **Pairwise Comparison:** We compared the interventions of therapists to determine whether Model A is better than Model B, vice versa, or if it’s a tie, for all possible pairs.<sup>8</sup>

**Human.** To enhance the reliability of the intervention evaluation, we conducted human evaluations by domain experts. We hired two fluent English-speaking psychotherapists through Upwork. The experts performed a pairwise comparison between our CS-LLaVA with multi-hop psychotherapeutic reasoning and others. (see Section C)

### 4.3 Test Scenarios

Toward reliable comparison, We compared the performance of both LLMs and VLMs with two test scenarios: AI-simulation and MM-CSCConv benchmark. The AI simulation testbed, which has been used in prior research, allows us to observe how interventions are carried out throughout conversations. However, using only the AI

<sup>6</sup>We utilize prompts from [https://github.com/lm-sys/FastChat/tree/main/fastchat/llm\\_judge](https://github.com/lm-sys/FastChat/tree/main/fastchat/llm_judge)

<sup>7</sup>We used the gpt-4-0613 version of the GPT-4 API.

<sup>8</sup>To ensure fairness and prevent position bias, we tested each case twice, swapping the positions each time.

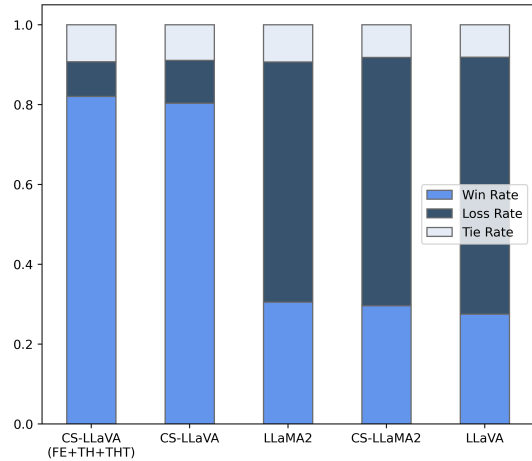


Figure 4: Win rates assessed by GPT-4 on the AI simulation testbed

simulation testbed makes it difficult to clearly compare the abilities of the AI therapist due to the subjectivity of the AI client. To achieve our goal of comparing therapists’ interventions in terms of empathy, logic, and rationality, we also compared therapists’ interventions when given the same context by observing turn-level performance on the MM-CSCConv benchmark.

#### 4.3.1 Scenario 1: AI-Simulation

**Settings.** In this scenario, we employ ChatGPT (API version)<sup>9</sup> as an AI client to test our approach in AI-to-AI scenarios. For prompting to AI client, we leverage 100 resources which are used as base resources to build the test set, from Sharma et al. (2023) and Mollahosseini et al. (2019). The role of

<sup>9</sup>We used the gpt-3.5-turbo-0125 version of the ChatGPT API.

|                         | Introduction |             |             | Problem Exploration |             |             | Brainstorming |             |             | Suggestion  |             |             |
|-------------------------|--------------|-------------|-------------|---------------------|-------------|-------------|---------------|-------------|-------------|-------------|-------------|-------------|
|                         | Emp.         | Coh.        | Gui.        | Emp.                | Coh.        | Gui.        | Emp.          | Coh.        | Gui.        | Emp.        | Coh.        | Gui.        |
| LLaMA2                  | 1.58*        | 1.79*       | 0.80*       | 2.16                | 2.20*       | 1.03*       | 2.10*         | 2.18*       | 1.44*       | 2.17*       | 2.06*       | 0.97*       |
| LLaVA                   | 0.64*        | 0.98*       | 0.05*       | 1.94*               | 1.96*       | 1.12*       | 1.86*         | 1.99*       | 1.39*       | 2.21*       | 2.37*       | 1.50*       |
| CS-LLaVA                | 1.87*        | 1.99        | 0.92*       | 2.15                | 2.24*       | <b>1.64</b> | 2.11*         | 2.25*       | 1.68        | 2.54        | 2.61        | 1.71        |
| CS-LLaVA<br>(FE+TH+THT) | <b>2.11</b>  | <b>2.16</b> | <b>1.02</b> | <b>2.23</b>         | <b>2.39</b> | 1.60        | <b>2.27</b>   | <b>2.39</b> | <b>1.79</b> | <b>2.59</b> | <b>2.67</b> | <b>1.80</b> |

Table 4: Manual evaluation results as assessed by GPT-4 at each stage on the MM-CSCConv benchmark. **Emp.**, **Coh.**, and **Gui.** represent Empathy, Logical Coherence, and Guidance, respectively. The values of each model show a significant difference compared to CS-LLaVA (FE+TH+THT), with a p-value < 0.05 (\*) as determined by the paired t-test.

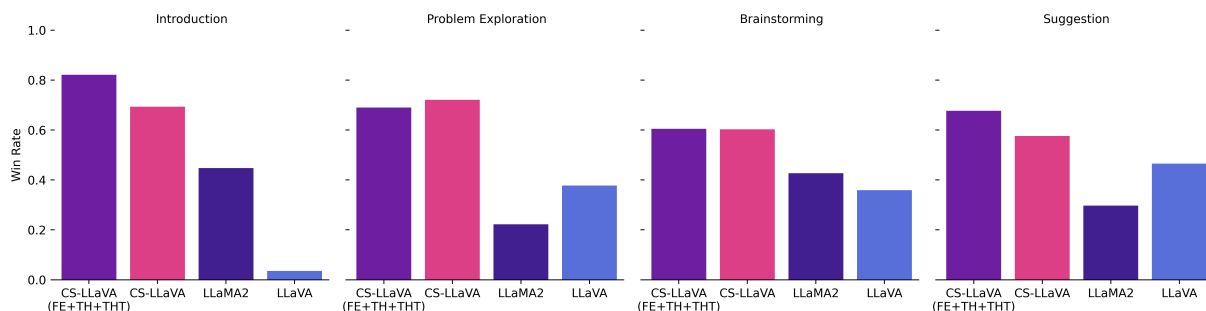


Figure 5: Win rates assessed by GPT-4 at each stage of the MM-CSCConv benchmark.

the AI client is the same as the data construction method, and the prompt used is also the same.

In this scenario, we also compared the results of CS-LLaMA2 to evaluate the performance difference between using only the text modality and using both text and image information. For CS-LLaMA2, only the text modality was used, without incorporating image information.

**Results.** Table 3 shows the manual scoring results in the AI simulation scenario. With our MM-CSCConv, the LLaVA family demonstrates significant improvements across all aspects. Furthermore, incorporating reasoning with three types of implicit evidence - *facial expressions*, *thoughts*, and *thinking traps* - led to performance enhancements across all evaluation aspects, with a significant improvement on the empathy aspect. This is quite close to the GPT-4 evaluation score for the MM-CSCConv test set, showing that it’s on par with GPT-4. For LLaMA2, there is almost no noticeable change before and after training with MM-CSCConv, mainly because image information is not provided during the Introduction stage, leading to training failure. Based on these results, we anticipate challenges in effectively training LLMs with MM-CSCConv, so we decided not to assess CS-LLaMA2 on the MM-CSCConv

benchmark except for human evaluation settings.

In the pairwise comparison using GPT-4 as a judge, LLaVA shows a significant improvement in the win rate, as illustrated in Figure 4 (see Section D for numerical win rates). There is no significant difference between LLaMA2 and CS-LLaMA2, and LLaVA shows the lowest win rate. Additionally, multi-hop psychotherapeutic reasoning slightly improves win rates and reduces loss rates. We also conducted a case study using our approaches and LLaMA2 (see Section E).

### 4.3.2 Scenario 2: MM-CSCConv benchmark

**Settings.** In this testbed, each AI therapist responds to the same dialogue history to directly compare their interventions. To ensure reliability, we conducted evaluations using both GPT-4 and two human psychotherapists. The evaluation is carried out at the turn level for each stage. Similar to the AI simulation testbed, we present win rate results alongside manual scores.

To further strengthen the reliability of the human evaluation results, we derived the win rate by comparing the proposed methodology with other approaches. Specifically, we compared CS-LLaVA with multi-hop psychotherapeutic reasoning to other baselines and to CS-LLaVA with standard prompting, as evaluated by two experts.

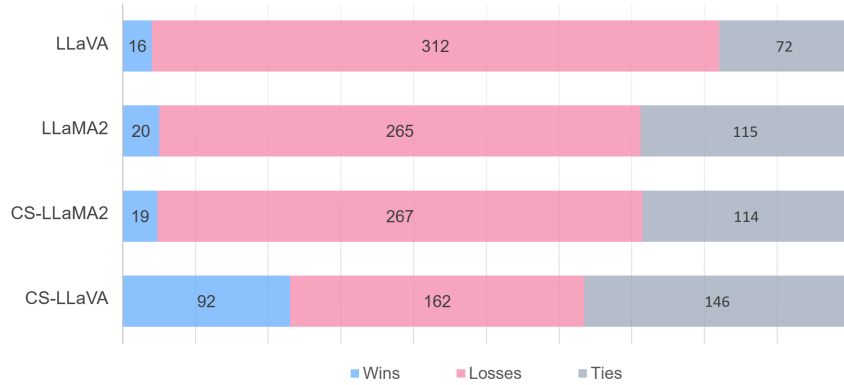


Figure 6: Win rates against CS-LLaVA (FE+TH+THT) assessed by two psychotherapists on the MM-CSCConv benchmark. The domain experts are asked to choose which intervention is better at assessing the given intervention.

**Results.** Table 4 shows the performance of manual scoring evaluated by GPT-4 at different stages. The score distribution is lower than that of the AI simulation testbed because it evaluates intervention at the utterance level rather than the entire conversation. Figure 5 illustrates the result of pairwise comparison among the four models under the GPT-4 judge (see Section F for numerical win rates). The performance difference due to prompting is most evident in the Introduction and Suggestion stages, and the performance difference due to the use of MM-CSCConv is clearly visible in all stages. In both cases, our approach consistently outperforms the baseline models at all stages except for the Problem Exploration stage. Similar to the manual scoring results, LLaVA shows the lowest win rate in the Introduction stage, supporting our hypothesis that LLaVA may lack the ability to express empathy using the client’s non-verbal information.

However, considering both manual scoring and pairwise comparison, standard prompting tends to slightly outperform multi-hop psychotherapy reasoning in the Problem Exploration stage. The Problem Exploration stage involves additional exploration of the client’s situation using facial expression evidence, which seems to improve empathy but not other evaluation criteria.

Figure 6 shows the pairwise comparison results between CS-LLaVA (FE+TH+THT) and other baseline models. The effectiveness of our approach is strongly supported by the fact that all models have significantly fewer wins and significantly more losses. Table 5 presents the results of human evaluation at each stage. CS-

|           | Win Rate (%) |             |        |       |
|-----------|--------------|-------------|--------|-------|
|           | Intro.       | Explo.      | Brain. | Sugg. |
| LLaMA2    | 10.0         | 25.5        | 32.0   | 10.0  |
| LLaVA     | 2.5          | 2.0         | 17.5   | 30.0  |
| CS-LLaMA2 | 12.0         | 22.5        | 29.5   | 12.0  |
| CS-LLaVA  | 29.0         | <b>54.0</b> | 39.0   | 43.0  |

Table 5: Win rates against CS-LLaVA (FE+TH+THT) as evaluated by two psychotherapists at each stage of the MM-CSCConv benchmark.

LLaVA (FE+TH+THT) shows the most superior performance. While LLaMA2 and LLaVA had significantly lower win rates against CS-LLaVA (FE+TH+THT) in all stages, CS-LLaVA outperformed CS-LLaVA (FE+TH+THT) in the Problem Exploration stage.

## 5 Conclusion

In this paper, we explored cognitive reframing therapy within a multimodal context. Recognizing the gap between real face-to-face cognitive reframing therapy and prior research, as well as the potential benefits of AI in psychotherapy, we aimed to enhance the therapeutic capabilities of AI therapists by incorporating non-verbal cues, particularly facial expressions, into the intervention process.

Our extensive experiments across two test scenarios, AI Simulation and the MM-CSCConv benchmark, indicate significant improvements in the therapeutic capabilities of VLMs when using MM-CSCConv. The multi-hop psychotherapeutic reasoning approach, which integrates facial expressions, thoughts, and thinking traps, demonstrated superior performance in providing empathetic, logically coherent, and specific rational suggestions to clients.



## 6 Limitations

We expanded the concept of cognitive reframing into multimodality, demonstrating that incorporating multimodal evidence and multi-hop psychotherapeutic reasoning significantly enhances the therapist’s abilities. However, these results were limited to virtual clients whose facial images and dialogues were consistent. This controlled setting may not fully capture the complexities of real-world interactions. We used benchmark images for facial expression recognition, but capturing the facial expressions of real clients can be challenging and might affect the consultation’s content. Moreover, our study only utilized facial images as the source of non-verbal information, which presents a limitation in comparison to actual face-to-face cognitive reframing therapy. Real-life therapy involves a broader spectrum of non-verbal cues, such as body language, tone of voice, and other contextual factors, which were not accounted for in our research.

For future work, we plan to expand the modalities to include a wider range of non-verbal information. By incorporating diverse non-verbal cues, we aim to further enhance the model’s ability to mimic real-life therapy scenarios. This will help bridge the gap between virtual and actual consultations, ultimately enabling the model to learn how to effectively utilize non-verbal information in a more realistic setting.

## References

- Aaron T. Beck. 1970. *Cognitive therapy: Nature and relation to behavior therapy*. *Behavior Therapy*, 1(2):184–200.
- Aaron T Beck. 1979. *Cognitive therapy and the emotional disorders*. Penguin.
- Aaron T Beck and Christine A Padesky. 1990. *Love is never enough*. Center for Cognitive Therapy.
- Robert A DiTomasso, Donna M Martin, and Karel D Kovnat. 2000. Medical patients in crisis.
- Lillian Döllinger, Petri Laukka, Lennart Björn Högman, Tanja Bänziger, Irena Makower, Håkan Fischer, and Stephan Hau. 2021. Training emotion recognition accuracy: Results for multimodal expressions and facial micro expressions. *Front. Psychol.*, 12:708867.
- Kallirroy-Fereniki Halamandaris and KG Power. 1997. Individual differences, dysfunctional attitudes, and social support: A study of the psychosocial adjustment to university life of home students. *Personality and individual differences*, 22(1):93–104.

- Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, 36:427–440.
- Ting-Yao Hsu, Chieh-Yang Huang, Ryan Rossi, Sungchul Kim, C. Giles, and Ting-Hao Huang. 2023. GPT-4 as an effective zero-shot evaluator for scientific figure captions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5464–5474, Singapore. Association for Computational Linguistics.
- A. N. Hutchison and L. H. Gerstein. 2012. What’s in a face? counseling trainees’ ability to read emotions. *Training and Education in Professional Psychology*, 6:100–112.
- Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.
- Sunil Krishnan. 2015. The counselling process: Stages of the counselling process. Technical report, Retrieved 07/09/2016, from <http://www.slideshare.net/SUNILKRISHNANPSYCHO> . . . .
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- 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)*, pages 13641–13660, Toronto, Canada. Association for Computational Linguistics.
- A. Mollahosseini, B. Hasani, and M. H. Mahoor. 2019. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(01):18–31.
- Renyi Qu, Lyle Ungar, and João Sedoc. 2023. Conditioning on dialog acts improves empathy style transfer. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen,

586 and Tim Althoff. 2023. [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)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics. 642

587 643

588 644

589 645

590

591

592 Mirai So, Sosei Yamaguchi, Sora Hashimoto, Mitsuhiko Sado, Toshi A Furukawa, and Paul McCrone. 2013. Is computerised cbt really helpful for adult depression?-a meta-analytic re-evaluation of cbct for adult depression in terms of clinical implementation and methodological validity. *BMC psychiatry*, 13:1–14. 646

593 647

594 648

595

596

597

598

599 Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. [Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

600

601

602

603

604

605

606 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*.

607

608

609

610

611

612 Susan R Walen, Raymond DiGiuseppe, and Windy Dryden. 1992. *A practitioner’s guide to rational-emotive therapy*. Oxford University Press.

613

614

615 Mengxi Xiao, Qianqian Xie, Ziyang Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. [Healme: Harnessing cognitive reframing in large language models for psychotherapy](#). *Preprint*, arXiv:2403.05574.

616

617

618

619

620 Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyang Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

621

622

623

624

625

626 Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. [Mentalama: Interpretable mental health analysis on social media with large language models](#). In *Proceedings of the ACM on Web Conference 2024, WWW ’24*, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

627

628

629

630

631

632

633 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

634

635

636

637

638

639

640 Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. [Inducing positive perspectives with text](#)

641

649  
650  
651  
652  
653

## A Overall Trait

We introduced the Overall score as a metric to assess the therapist’s overall ability, with the calculation method illustrated in Algorithm 1.  $e$ ,  $c$ ,  $g$  stand for empathy, logical coherence, and guidance.

---

### Algorithm 1 Overall Score Calculation

---

```
function GETOVERALLSCORE( $e$ ,  $c$ ,  $g$ )
  if  $e \leq 1$  or  $c \leq 1$  then
    return 0
  end if
  if ( $e \geq 2$  and  $c \geq 2$ ) and  $g \leq 1$  then
    return 1
  end if
  if ( $e \geq 2$  and  $c \geq 2$  and  $g == 2$ ) then
    return 2
  end if
  if  $e \geq 2$  and  $c \geq 2$  and  $g == 3$  then
    return 3
  end if
end function
```

---

## B Data Cleansing Manual

Data cleansing guidelines are shown in Table 7. *Image-Dialog Consistency* is a criterion that evaluates whether the client’s visual information and dialogue are consistent. If any of the four criteria received a score of 0, the corresponding data was deleted, and the *Image-Dialog Consistency* of the MM-CSCConv is shown in Table 6.

| Image-Dialog Consistency |       |
|--------------------------|-------|
| Train                    | 1.472 |
| Test                     | 1.667 |

Table 6: Image-Dialog Consistency on the MM-CSCConv dataset.

## C Details for Human evaluator

### C.1 Hiring and Payment

We hired a total of three psychotherapists, and paid \$0.8 per conversation for dataset evaluation and \$0.0625 per data entry for pairwise comparison.

### C.2 Instructions for MM-CSCConv evaluation

We provided domain experts with instructions for evaluating the MM-CSCConv test set, as illustrated in Figure 10.

### C.3 Instructions for Human pairwise comparison

We provided the other two experts with instructions for conducting human pairwise comparisons, as shown in Figure 11. The evaluation sheet provided to them included only Dialog history, Response A, Response B, and Question ID, without revealing information about Model A and Model B. Additionally, to prevent bias based on position, Model A and Model B were randomly assigned for each data entry.

## D Numerical results of pairwise comparison on AI-simulation assessed by GPT-4.

In Table 8, we have provided the numeric results as additional data for Figure 4. When comparing the comparison results for each model pair, CS-LLaVA (FE+TH+THT) showed a win rate exceeding 50%, outperforming all models.

## E Case Study in AI-Simulation

We conducted additional analysis on test cases to compare our approaches with the LLaMA2, one of the baseline models. Figure 7 displays the full conversation between LLaMA2 and an AI client, while Figure 8 shows the complete conversation between CS-LLaVA and an AI client. Additionally, Figure 9 presents the entire conversation between CS-LLaVA (FE+TH+THT) and an AI client. These three conversations were generated using the same base resource in the AI simulation testbed.

LLaMA2 offers unconditional consolation, as it lacks the capability to utilize client information in the initial stage. In contrast, both CS-LLaVA and CS-LLaVA (FE+TH+THT) demonstrate more specific empathy. The client in these scenarios exhibits cognitive distortions of the overgeneralization type. With LLaMA2, the response involves merely engaging in conversation. However, CS-LLaVA and CS-LLaVA (FE+TH+THT) go further by encouraging the client to explore alternative ideas. Furthermore, CS-LLaVA (FE+TH+THT) not only promotes alternative thinking but also prompts the client to reflect on whether they have made similar misjudgments in the past.

654  
655  
656  
657  
658  
659  
660  
661

662  
663  
664  
665  
666  
667  
668  
669  
670

671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714

715 **F Numerical results of pairwise**  
716 **comparison on MM-CSCov assessed**  
717 **by GPT-4.**

718 In Table 9 and Table 10, we have provided the  
719 numeric results as additional data for Figure 5.  
720 When comparing each model pair, CS-LLaVA  
721 (FE+TH+THT) showed a win rate that surpassed  
722 other models, similar to the AI-simulation testbed  
723 results.



| Metric                          | Description  | Scale |
|---------------------------------|--|-------|
| <i>Client-clarity</i>           | The client expresses his or her situation clearly in the conversation  | 1/0   |
| <i>Client-role</i>              | The client adheres to the role of the client throughout the conversation.  | 1/0   |
| <i>Therapist-role</i>           | The therapist adheres to the role of the therapist in all conversations.   | 1/0   |
| <i>Image-Dialog Consistency</i> | There is no consistency between the client’s facial image and the client’s utterances or situation. The client’s facial image is relevant to neither the client’s utterances nor the client’s situation. | 0     |
|                                 | There is acceptable consistency between the client’s facial image and the client’s utterances or situation.  | 1     |
|                                 | There is strong consistency between the client’s facial image and the client’s utterances or situation.  | 2     |

Table 7: Guideline for data cleansing in MM-CSCConv. *Client-clarity*, *Client-role*, and *Therapist-role* are assigned 1 if they match the description, and 0 otherwise.

|                             | LLaMA2        | LLAVA-7b      | CS-LLAMA2     | CS-LLAVA      | CS-LLAVA (FE+TH+THT) | Win Rate      |
|-----------------------------|---------------|---------------|---------------|---------------|----------------------|---------------|
| <b>LLAMA2</b>               | -             | 52.551        | 49.495        | 9.694         | 9.794                | 30.485        |
| <b>LLAVA-7b</b>             | 47.449        | -             | 49.485        | 6.566         | 7.071                | 27.481        |
| <b>CS-LLAMA2</b>            | 50.505        | 50.516        | -             | 10.309        | 7.071                | 29.592        |
| <b>CS-LLAVA</b>             | <b>90.306</b> | 93.434        | 89.691        | -             | 47.959               | 80.357        |
| <b>CS-LLAVA (FE+TH+THT)</b> | 90.206        | <b>92.929</b> | <b>92.929</b> | <b>52.041</b> | -                    | <b>82.061</b> |

Table 8: Numerical results of pairwise comparison of five models on the AI-simulation testbed, evaluated using GPT-4.

|                             | Win Rate (%)  |                     |               |               |
|-----------------------------|---------------|---------------------|---------------|---------------|
|                             | Introduction  | Problem Exploration | Brainstorming | Suggestion    |
| LLaMA2                      | 44.767        | 22.222              | 42.701        | 29.740        |
| LLaVA                       | 3.529         | 37.770              | 35.907        | 46.539        |
| CS-LLaVA                    | 69.336        | <b>72.119</b>       | 60.256        | 57.617        |
| <b>CS-LLaVA (FE+TH+THT)</b> | <b>82.101</b> | 68.978              | <b>60.478</b> | <b>67.671</b> |

Table 9: Win rates at each stage on the MM-CSCConv benchmark, evaluated using GPT-4.

|                             | LLaMA2        | LLAVA-7b      | CS-LLAVA      | CS-LLAVA (FE+TH+THT) | Win Rate      |
|-----------------------------|---------------|---------------|---------------|----------------------|---------------|
| <b>LLAMA2</b>               | -             | 54.096        | 27.562        | 22.877               | 34.676        |
| <b>LLAVA-7b</b>             | 45.904        | -             | 25.710        | 21.676               | 31.179        |
| <b>CS-LLAVA</b>             | 72.438        | 74.290        | -             | 47.067               | 64.848        |
| <b>CS-LLAVA (FE+TH+THT)</b> | <b>77.123</b> | <b>78.324</b> | <b>52.933</b> | -                    | <b>69.677</b> |

Table 10: Numerical results of pairwise comparison of four models on the MM-CSCConv benchmark, evaluated using GPT-4.

|           |   |
|-----------|---|
| Client    | Hello.  |
| Therapist | I'm so sorry to hear that you're feeling upset. Can you tell me more about what's on your mind?   |
| Client    | <b>I can't help but feel like she doesn't respect me.</b> It seems like every time I try to talk to her, she dismisses my opinions and ideas. <b>I start to question if I'm worth respecting at all.</b> This feeling of not being respected is really getting to me.   |
| Therapist | It sounds like you're feeling really hurt and frustrated by the way you've been treated by this person. It's completely understandable that you would feel this way, and I'm here to listen and support you. Can you tell me more about what you're thinking and feeling right now?   |
| Client    | I keep thinking that she doesn't respect me because she dismisses my opinions and ideas. I start to wonder if I'm worth respecting at all.  |
| Therapist | It sounds like you're feeling really hurt and frustrated by the way you've been treated by this person. It's completely understandable that you would feel this way, and I'm here to listen and support you. Have you considered the possibility that she may not be aware of the impact her behavior is having on you? It's possible that she's not intentionally disrespecting you, but rather that she's not aware of how her actions are coming across. Have you talked to her about how you're feeling? <b>It might be helpful to have an open and honest conversation with her</b>          |
| Client    | I keep thinking that she doesn't respect me because she dismisses my opinions and ideas. I start to wonder if I'm worth respecting at all.  |
| Therapist | Thank you for sharing this with me. It sounds like you're feeling really upset and hurt by the way this person has been treating you. It's completely understandable that you would feel this way, and I'm here to listen and support you. I want you to know that you deserve to be treated with respect and kindness, and it's okay to stand up for yourself and express your feelings to this person. Have you considered talking to her about how you're feeling? <b>It might be helpful to have an open and honest conversation with her about the impact her behavior is having on you.</b> |

Figure 7: A case study between an AI Client (ChatGPT) and LLaMA2. The red shading highlights the client's distorted thoughts, while the yellow shading emphasizes the key components of the therapist's intervention.



|           |   |
|-----------|---|
| Client    | Hello.  |
| Therapist | I notice there's a hint of sadness in your expression. Whenever you feel ready, let's talk about what's been troubling you. I'm here to listen and support you.   |
| Client    | <b>I feel like she doesn't respect me.</b> It seems like no matter what I do, she always dismisses my opinions and never takes me seriously. <b>I start to think that maybe I'm not worthy of respect at all.</b>   |
| Therapist | I see that you're feeling quite down about this situation. It's important to recognize that your feelings are valid, but let's try to look at the situation separately from the thoughts it's bringing up. Can you tell me more about the specific instances where you felt disrespected?   |
| Client    | I can't shake off the feeling of being unworthy.  |
| Therapist | I understand that feeling unworthy can be really tough. <b>Let's explore other perspectives.</b> Could it be possible that she's dealing with her own stress or challenges that are affecting her behavior? Or maybe she's not aware of how her actions are coming across. What are some other reasons she might be dismissing your opinions that aren't about your worth?  |
| Client    | It's hard for me to consider other reasons for her behavior towards me, but I suppose it's possible that she might be under a lot of stress or facing her own challenges that are impacting how she interacts with me. Maybe she's not intentionally trying to disrespect me.   |
| Therapist | I'm really impressed with your willingness to consider alternative explanations for her behavior. <b>It's a significant step to recognize that her actions might not be a reflection of your worth.</b> It's also important to remember that everyone has their own struggles, and it's possible that she's dealing with her own challenges that are affecting her interactions with you. Let's continue to explore these perspectives and see if they help you feel more understood and respected. |

Figure 8: A case study between an AI Client (ChatGPT) and CS-LLaVA.



|           |  |
|-----------|--|
| Client    | Hello.   |
| Therapist | I notice there's a hint of sadness in your expression. Whenever you feel ready, I'm here to listen and support you through whatever is troubling you. What's been on your mind?  |
| Client    | <b>It feels like everyone always disrespects me.</b> I can never seem to get anyone to treat me with the respect I deserve. It's like no matter what I do, people just don't care about how I feel.  |
| Therapist | I see that you're feeling quite down and it's understandable to feel disrespected. Let's take a moment to distinguish between the actions of others and the thoughts you have about their intentions. Could there be instances when people have been respectful, even if it might not have felt that way?  |
| Client    | I'll try to think about instances where people have shown respect towards me, even if it didn't feel that way initially.   |
| Therapist | It's good that you're willing to reflect on positive interactions and evaluate whether they felt respectful or not. Sometimes, our perception may not immediately reflect the intentions of others. <b>Consider this: might there have been times when someone was trying their best to be respectful, but the communication got misinterpreted?</b> Or, could there be instances where someone showed respect in ways that aren't immediately recognizable? Think about moments of courtesy or consideration, even in small ways.   |
| Client    | I can remember a time when my colleague offered to help me with a project at work, even though I initially interpreted it as them trying to undermine me. Looking back, I can see that they were genuinely trying to be helpful and supportive, even though I didn't see it that way at the time. Maybe there have been other instances where people have shown respect in subtle ways that I overlooked.  |
| Therapist | You've made a significant step by recognizing that there may have been instances of support from colleagues that you might have missed. <b>It's commendable that you're willing to explore this further and consider that people might be showing respect in ways you haven't entirely appreciated.</b> Keep in mind that our perceptions can sometimes be clouded by our own stress or challenges. Engaging with others can provide more clarity and help you determine if a person's behavior is truly disrespectful or if your interpretation might not fully reflect their intentions. |

Figure 9: A case study between an AI Client (ChatGPT) and CS-LLaVA (FE+TH+THT).

## Overview

We built a multimodal cognitive reframing conversation dataset using LLMs for research purpose. As this is a conversation created by LLM, data verification is required.

## Job Description

- You will be provided with 100 conversations, each accompanied by a corresponding facial image.
- The conversation focuses on CBT, with the therapist working to reframe the client's distorted thinking.
- You are expected to evaluate each conversation based on three specific criteria outlined in the provided manual.

## Evaluation traits

Please rate the items below:

| 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.                  |

## One example of the conversation

**Client:**  
Hello

**Therapist:**  
You seem to be in a calm state at the moment. I'm here to listen and support you. Is there anything in particular that's been bothering you or that you'd like to discuss?

**Client:**  
I can't believe I dyed my hair, I shouldn't have done it. It looks terrible and everyone is going to judge me. I should have just left it the way it was.

**Therapist:**  
I see that you're feeling upset about your new hair color. Let's take a moment to separate the event of dyeing your hair from the thoughts and fears of being judged. Right now, it's only the thought of judgment that is bothering you, not an actual negative response from others.

**Client:**  
I understand. I will try to separate my thoughts about being judged from the actual event of changing my hair.

**Therapist:**  
That's an important step you've taken. Now, can you think of other outcomes that could come from changing your hair? Perhaps people might admire the change or see you as bold for trying something new. It's also possible that others are too preoccupied with their own lives to notice much, meaning the judgment you fear may not come. How do you think these other perspectives fit with your situation?

**Client:**  
I hadn't considered those other perspectives before. It's possible that people could react positively or simply not notice at all. Thank you for pointing that out.

**Therapist:**  
I'm really impressed with your willingness to explore different perspectives, and it's great to hear you considering other possible reactions. Remember, the decision to change your hair was yours, and it's an expression of your individuality. It's natural to feel apprehensive about change, but it's also an opportunity to embrace your personal growth and how you choose to present yourself to the world. Celebrate your courage to try something new, and no matter what others think, what matters most is how you feel about yourself.

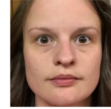


Image Description: AffectedNet (neutral/3115.jpg)

## Output File

Once the contract is signed, we will provide a Google spreadsheet for grading.

|    | dialog_id | Empathy | Logical Coherence | Guidance |
|----|-----------|---------|-------------------|----------|
| 1  | dialog_7  |         |                   |          |
| 2  | dialog_16 |         |                   |          |
| 3  | dialog_17 |         |                   |          |
| 4  | dialog_19 |         |                   |          |
| 5  | dialog_23 |         |                   |          |
| 6  | dialog_24 |         |                   |          |
| 7  | dialog_25 |         |                   |          |
| 8  | dialog_26 |         |                   |          |
| 9  | dialog_30 |         |                   |          |
| 10 | dialog_33 |         |                   |          |
| 11 | dialog_41 |         |                   |          |
| 12 | dialog_42 |         |                   |          |
| 13 | dialog_43 |         |                   |          |
| 14 | dialog_48 |         |                   |          |
| 15 | dialog_51 |         |                   |          |
| 16 | ...       |         |                   |          |
| 17 | ...       |         |                   |          |

Figure 10: Instruction for human dataset evaluation.

## Overview

We would like to build an AI model that performs **cognitive reframing**. We hire experts to compare and evaluate multiple AI models that perform cognitive reframing conversations. The evaluation results will be used for research purpose.

You are given 1600 triplet (question\_id, response A, response B). If you check the pdf file with the same question\_id (to be provided), there is a record of the previous dialogue history. (Each AI model responded according to its previous dialogue history.)

As an expert, please choose which of the two responses is better (A, B, or Tie)

## Job Description

- The conversation focuses on CBT, with the therapist working to reframe the client's distorted thinking.
- We will provide 1600 pairs of AI model responses.
- If you look at the Google Spreadsheet, each entry has **question\_id**, **response A**, and **response B**.
- There is a pdf file corresponding to **question\_id** (e.g. dialog:7-turn:0 and dialog\_7-turn\_0.pdf), and the pdf file contains the context of the previous conversation and the client's face image.
- Please select the response that best suits the conversation context and the client's facial expression contained in the pdf file. (**A, B, or Tie**)

## Evaluation traits

There is no exact evaluation trait, but please mainly observe empathy, logical coherence, and rationality.

## An example of dialog history

**Client:**  
Oh, hi.

**Therapist:**  
You seem to be carrying some sadness in your expression. Whenever you feel ready, I want you to know that this is a safe environment where you can share what's been weighing on you.

**Client:**  
I can't believe he took it without asking me first. It's so disrespectful and inconsiderate. I feel so angry and betrayed by his actions. He should have known better than to do that without my permission.

**Therapist:**  
I see you're feeling very upset. Let's take a moment to breathe deeply together. Now, can you describe what happened without attaching any judgment or feelings to it? Just the facts of the situation.

**Client:**  
He took something without asking me first.

**Therapist:**  
It's understandable to feel upset when something is taken without your permission. Let's consider what could have been happening from the other person's perspective - perhaps there was an urgent need or a misunderstanding regarding the ownership or rules of using the item. Could there be a chance that this action wasn't intended to hurt or disrespect you? What are some other reasons that might have led to this situation?



Image Description: AffectedNet (sad/5036.jpg)

## Output File

Once the contract is signed, we will provide a Google spreadsheet for grading.

| question_id       | response_A  | response_B   | Result |
|-------------------|---|--|--------|
| dialog_431-turn-0 | You seem to be quite composed today. How are you currently feeling, and is there anything in particular that's been occupying your thoughts? Please feel free to share. This is a safe space to explore your feelings.  | Sorry to hear that you're feeling upset. Can you tell me more about what's on your mind?   |        |
| dialog_431-turn-2 | Therapist: It's understandable that you're feeling upset about the mistakes you've made. It's important to acknowledge and validate your emotions. Can you tell me more about what happened and how you're feeling about it?  | I'm glad to see you're embracing a more balanced perspective. Now, let's explore other interpretations of the situation. Could it be that everyone makes mistakes and it's a natural part of learning and growth? Maybe you can base your opinion on context by using words like "usually" instead of "and" and you're affected from these mistakes and consider that they don't define your entire performance or worth. What are some positive outcomes that have come from these events, despite the challenges?  |        |
| dialog_431-turn-4 | Of course, I'm here to help you explore other possibilities under the same situation. It's understandable that you're feeling upset about the mistakes you've made, but let's try to approach this from a different angle. Instead of focusing solely on the mistakes, can we brainstorm other ways you could have handled the situation differently? Perhaps there were other options you could have considered, or ways you could explore these alternatives, we might be able to find a more constructive way to move forward. How does that sound?  | I'm glad to see you're embracing a more balanced perspective. Now, let's explore other interpretations of the situation. Could it be that everyone makes mistakes and it's a natural part of learning and growth? Maybe you can base your opinion on context by using words like "usually" instead of "and" and you're affected from these mistakes and consider that they don't define your entire performance or worth. What are some positive outcomes that have come from these events, despite the challenges?  |        |
| dialog_431-turn-6 | That's an insightful way to look at your experiences, and it's commendable that you're open to considering this perspective. Growth often comes from challenges we face, and every mistake is an opportunity to learn and improve. Remember, perfection is not a requirement for success, and it's okay to seek forgiveness and move on when things don't go as planned. Your ability to reframe these thoughts shows resilience and a willingness to adapt, which are qualities of a strong and capable person. Keep embracing this perspective, and let's continue to build on this positive outlook in our sessions. | Thank you for sharing your thoughts with me today. I can see that you've been doing a lot of brainstorming and reflecting on your recent experiences. It takes a lot of courage to confront our mistakes and work through them, and I'm proud of you for taking this step. It's important to remember that everyone makes mistakes, and it's how we learn and grow from them that matters. Rather than focusing on labeling yourself as a "failure," let's try to reframe these experiences as opportunities for growth and learning. By doing so, you can develop a more positive mind. |        |
| dialog_82-turn-0  | I notice that you seem to be experiencing some tear right now. Whenever you feel ready, we can talk about what is causing you to feel this way. Remember, this is a supportive environment for you to share your feelings.  | Sorry to hear that you're feeling upset. Can you tell me more about what's on your mind?   |        |
| dialog_82-turn-0  | I see that you're feeling really anxious about forgetting something important. Let's take a moment to distinguish between the event itself - forgetting an item - and the thoughts you're having about the event and your self-worth. Remember, an oversight doesn't define your entire   | Therapist: It sounds like you're feeling really upset about forgetting something important at work. Can you tell me more about what's going on and how you're feeling? It's important to remember that everyone makes mistakes, and it doesn't define who you are as a person or professional.   |        |

Figure 11: Instruction for human pairwise comparison.