

Multimodal Cognitive Reframing Therapy via Multi-hop Psychotherapeutic Reasoning

Subin Kim^{*1}, Hoonrae Kim^{*1}, Heejin Do¹, and Gary Geunbae Lee^{1,2}

¹Graduate School of Artificial Intelligence, POSTECH, South Korea

²Department of Computer Science and Engineering, POSTECH, South Korea
{kimsubin, hoonrae, heejindo, gblee}@postech.ac.kr

Abstract

Previous research has revealed the potential of large language models (LLMs) to support cognitive reframing therapy; however, their focus was primarily on text-based methods, often overlooking the importance of non-verbal evidence crucial in real-life therapy. To alleviate this gap, we extend the textual cognitive reframing to multimodality, incorporating visual clues. Specifically, we present a new dataset called **Multi Modal-Cognitive Support Conversation (M2CoSC)**, which pairs each GPT-4-generated dialogue with an image that reflects the virtual client's facial expressions. To better mirror real psychotherapy, where facial expressions lead to interpreting implicit emotional evidence, we propose a multi-hop psychotherapeutic reasoning approach that explicitly identifies and incorporates subtle evidence. Our comprehensive experiments with both LLMs and vision-language models (VLMs) demonstrate that the VLMs' performance as psychotherapists is significantly improved with the M2CoSC dataset. Furthermore, the multi-hop psychotherapeutic reasoning method enables VLMs to provide more thoughtful and empathetic suggestions, outperforming standard prompting methods.

1 Introduction

As a crucial part of cognitive behavioral therapy (CBT), *cognitive reframing* addresses lots of mental health issues rooted in deeply ingrained negative and distorted thought patterns (Beck, 1970; Powles, 1974; Beck, 1987, 1988; Walen et al., 1992; Halamandaris and Power, 1997; DiTomasso et al., 2000; Hofmann et al., 2012). Recently, studies attempted to utilize large language models (LLMs) in this task, highlighting their growing potential in the field of psychotherapy (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). Conventionally, cognitive reframing has been explored



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

with text-based sentence rewriting methods aimed at shifting negative viewpoints to positive ones (Ziems et al., 2022; Maddela et al., 2023; Sharma et al., 2023). Concerned that sentence-based cognitive reframing can feel unnaturally imposed, Xiao et al. (2024) suggest a three-stage conversational approach with LLMs encouraging clients to engage more actively and form self-positive viewpoints.

Despite the promising results of LLMs in previous systems, non-verbal aspects of psychotherapeutic theory are often overlooked, creating a significant gap between real face-to-face therapy and the systems. In actual psychotherapy situations, recognizing non-verbal emotions is essential for effective communication and is a critical skill closely linked to the therapist's ability to provide effective therapy (Hutchison and Gerstein, 2012; Döllinger et al., 2021).

In this study, we propose to extend the concept of cognitive reframing into multimodality, integrating visual clues into the therapy process. Our aim is to provide a natural and effective cognitive reframing

¹To comply with the terms of the AffectNet license, all images presented in this paper are synthesized using DALL-E 3 (Betker et al., 2023) and not sourced from the AffectNet dataset.

^{*}Equal Contribution

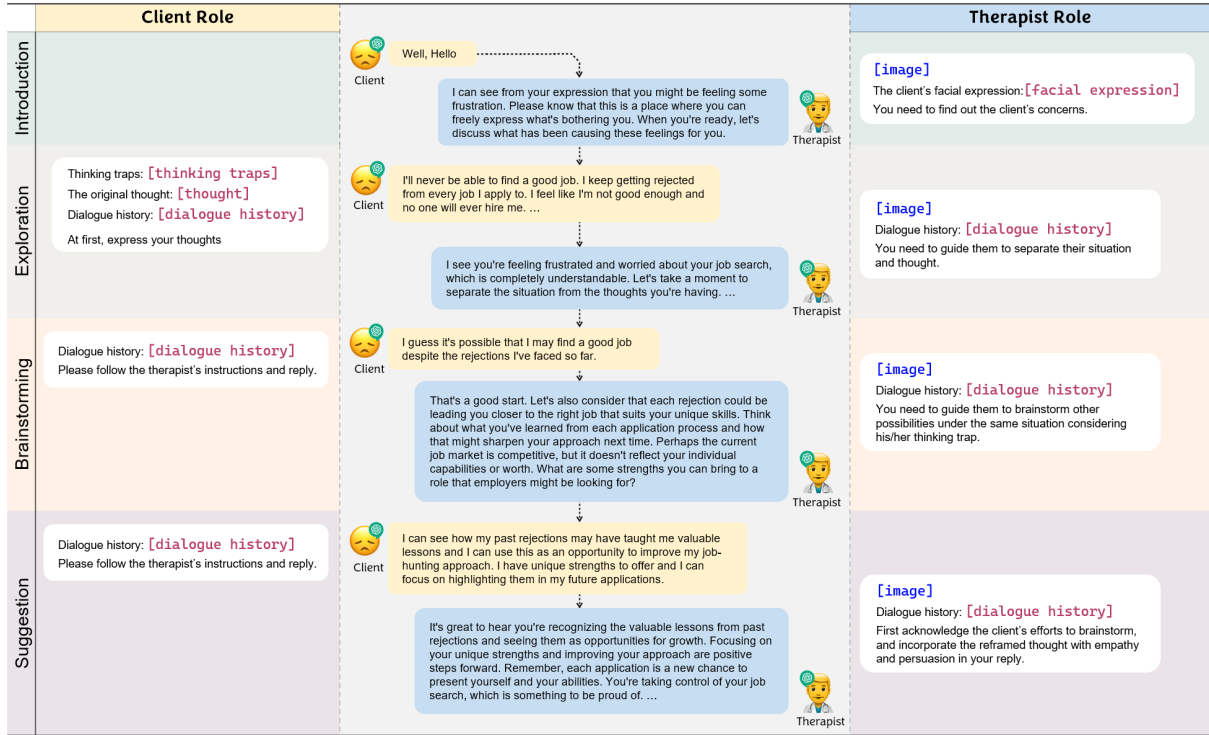


Figure 2: An example illustrating the construction of the M2CoSC dataset. Left: the prompt given to GPT-4 in the **client** role; Right: the prompt given to GPT-4 Vision in the **therapist** role. GPT-4 Vision is given a client’s face image. [dialogue history] denotes a history of conversations accumulated during role play.

framework within a multimodal context, incorporating clients’ non-verbal clues in the conversation (Figure 1). To this end, we create a novel synthetic benchmark dataset, **Multi Modal-Cognitive Support Conversation (M2CoSC)**, which pairs each synthetic dialogue with an image of the client’s facial expression.

To create M2CoSC, we utilize the powerful role-playing capabilities of LLMs. We leverage GPT-4 Vision and GPT-4 (Achiam et al., 2023) to take on the roles of psychotherapist and client, respectively, simulating therapy sessions as shown in Figure 2. Inspired by counseling theory’s (Krishnan, 2015; Psylog, 2024; Claibourne Counseling, 2024) *Initial Disclosure* stage among the five stages, we add the Introduction phase into our multimodal psychotherapy framework, extending the three-stage model proposed by Xiao et al. (2024): Introduction, Problem Exploration, Brainstorming, and Suggestion. In addition, to provide explicit guidance based on the client’s state, we introduce a multi-hop psychotherapeutic reasoning method. By exploring the implicit evidence necessary for the therapy and then generating responses based on this evidence, the AI therapist can offer adequate guidance after identifying the client’s state.

We evaluate our approach by conducting extensive experiments with two test scenarios, dialogue-level evaluation and stage-level evaluation, using both LLMs and Vision-Language Models (VLMs). The results show that training with our M2CoSC significantly enhances the counseling capabilities of VLMs, surpassing those of existing LLMs. Moreover, the multi-hop psychotherapeutic reasoning method allows VLMs to offer more rational and empathetic suggestions, outperforming standard prompting methods. Human evaluations further suggest that capturing the client’s facial expressions in the system, as practical therapists do, can remarkably assist counseling.

To sum up, our contributions are as follows: (1) We propose multimodal cognitive reframing therapy using non-verbal information, creating the M2CoSC dataset that pairs dialogues with client facial expressions. (2) We establish a baseline for the M2CoSC dataset and propose a multi-hop psychotherapeutic reasoning approach to improve the capabilities of VLMs in delivering rational therapeutic interventions.

This work is a first step toward bringing multimodal cognitive reframing into AI-enhanced psychotherapy. By introducing a benchmark dataset

and a structured reasoning approach, we hope to inspire future research on leveraging non-verbal cues for more effective therapeutic conversations.

2 Problem Definition and Goals

In cognitive reframing therapy, a therapist must understand the client's states, which include their problematic situations, distorted thoughts, and thinking traps. The therapist then encourages the client to consider alternative possibilities. Building rapport with the client by expressing empathy is also crucial (Horvath and Luborsky, 1993; Lambert and Barley, 2001). In the real-world psychotherapy procedure, these stages involve both verbal and non-verbal information, where the therapist has sufficient ability to understand the client's states.

Here, our goal is to enhance the abilities of an AI psychotherapist by leveraging non-verbal information, guiding it to focus on facial expressions and to comprehend the client's states. Given the client's facial image and dialogue history, we aim to provide empathetic responses while maintaining a consistent focus on the client's issues throughout the procedure, offering rational interventions free from logical errors or contradictions.

To achieve our goals, we created the M2CoSC dataset founded on three key values—empathy, logical coherence, and guidance—which serve as evaluation criteria in a prior study (Xiao et al., 2024).

- Empathy reflects the therapist's ability to understand and connect with the client's emotions, assisting in building trust, connection, and emotional support, all critical to a strong therapeutic relationship.
- Logical coherence denotes the therapist's ability to organize thoughts and provide well-structured insights, enhancing the quality of the conversation.
- Guidance indicates the therapist's capacity to offer practical advice, solutions, and direction, aiding the client to navigate challenges, make informed decisions, and achieve positive outcomes.

We also utilize overall scores encompassing all three items. (see Appendix A for details.)

3 M2 Cognitive Support Conversation

3.1 Dataset Construction

Actual therapy conversations are rarely accessible due to the sensitive nature of mental health therapy; thus, we have created a synthetic dataset that can be shared with the research community². To construct multimodal conversational cognitive reframing dataset, we utilize two publicly available sources: the Facial Expression Recognition (FER) dataset called AffectNet (Mollahosseini et al., 2019), and the cognitive reframing dataset from Sharma et al. (2023). To address potential privacy concerns associated with using images of real people from AffectNet, we obtained agreement for all research participants, ensuring full compliance with AffectNet's policies.

For construction, we set up role-play scenarios with two agents: GPT-4 in the client role and GPT-4 Vision in the therapist role.

As shown in Figure 2, we prompt GPT-4 in the client role and GPT-4 Vision in the therapist role using a set of four inputs: *image*, *facial expression*, *thinking traps*, and *thought*.³ The *image* 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*.

For facial expressions and images, We employ AffectNet, containing publicly accessible images from the internet collected for research under non-commercial use. For thinking traps and thoughts, we utilize the well-designed open-sourced dataset from Sharma et al. (2023), which was collected following ethical guidelines, including informed consent and participant privacy safeguards. To the best of our knowledge, this work is the first to combine multiple datasets to create a multimodal conversation specifically designed for the mental health domain.

Each dialogue consists of four turns, which correspond to different stages of a psychotherapeutic conversation. In this context, a "turn" is the same as a "stage." The action expected from the client is to follow the psychotherapist's instructions, and the actions required for the psychotherapist at each stage are as follows.

²Under AffectNet's license, M2CoSC dataset is partially available at <https://github.com/nobel-postech/M2CoSC>.

³We used version gpt-4-0613 of the GPT-4 API and version gpt-4-1106-vision-preview of the GPT-4 Vision API.

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

Table 1: Dataset statistics for M2CoSC. *# of Dialogue* 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.

1. **Introduction:** The AI psychotherapist expresses empathy and encourages the client to explore their problems.
2. **Exploration:** The AI psychotherapist guides the client to distinguish their thoughts from their situation.
3. **Brainstorming:** 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:** The AI psychotherapist first recognizes the client’s effort to explore other possibilities and presents specific and rational suggestions for the client.

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

The statistics for the M2CoSC dataset are summarized in Table 1. The M2CoSC dataset contains a total of 429 conversations, each consisting of exactly four turns.

3.2 Dataset Cleansing

To ensure the quality of the M2CoSC dataset, we conducted manual data cleansing with the three

	Image-Dialogue Consistency
	(0-2)
Train	1.472
Test	1.667

Table 2: *Image-Dialogue Consistency* on the M2CoSC dataset.

	Empathy	Logical Coherence	Guidance	Overall
	(0-3)	(0-3)	(0-3)	(0-3)
Human	2.929	2.980	2.879	2.929
GPT-4	2.920	2.930	2.400	2.420

Table 3: Human and GPT-4 evaluation results on M2CoSC test set.

native speakers, focusing on four aspects: *Client-clarity*, *Client-role*, *Therapist-role*, and *Image-Dialogue Consistency* (see Appendix B for detailed criteria). The *Image-Dialogue 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. Table 2 indicates a considerable correlation between the client’s facial expressions and their verbal responses in our M2CoSC dataset.

We hired three native English speakers through Upwork⁴, a crowdsourcing platform, to support this cleansing process.

3.3 Dataset Quality Validation

To further validate the quality of the M2CoSC, we evaluate the test set of the M2CoSC 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 them to manually evaluate the test set according to the detailed guidelines. We also employed GPT-4 for evaluation, feeding it the scorecard criteria and the dialogues from the test set.

The results in Table 3 indicate that both the human evaluator and GPT-4 provided high scores, showing similar scoring tendencies except for "Guidance." GPT-4 tended to give lower scores in this aspect, possibly attributable to its higher degree of expectation for problem-solving than human evaluators (Chiu et al., 2024). Despite this discrepancy, the overall consistency in the other criteria supports the dataset’s reliability.

⁴<https://www.upwork.com/>

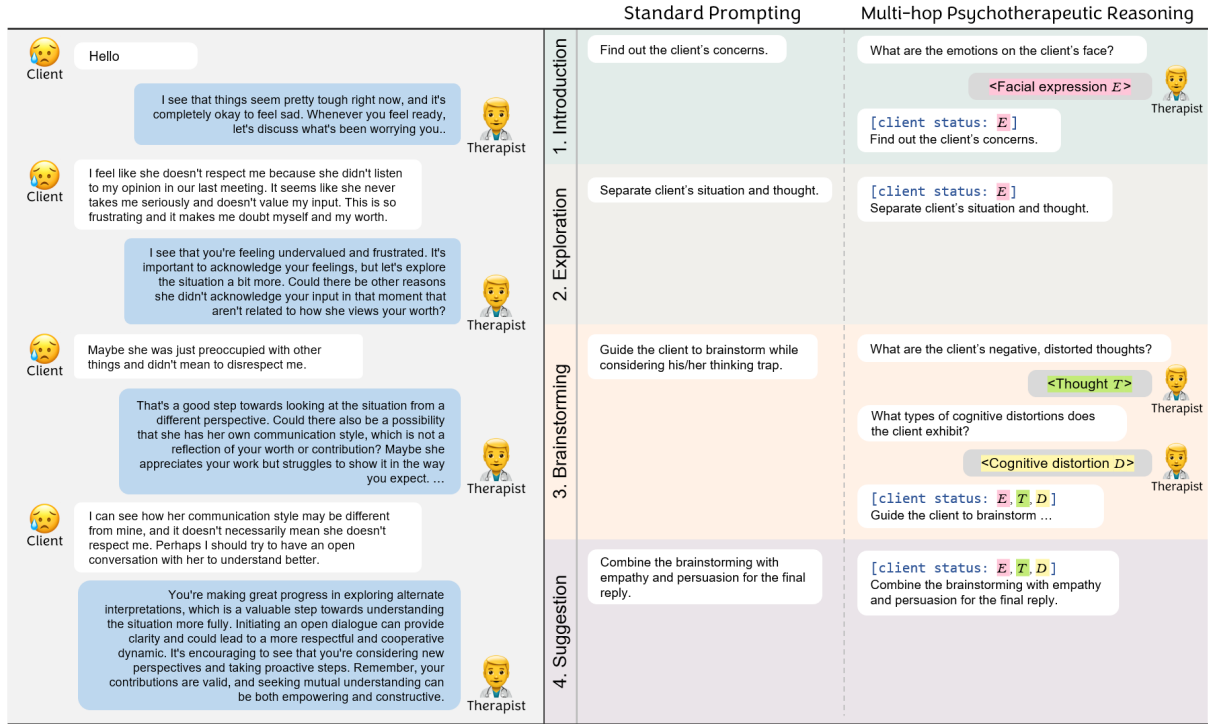


Figure 3: Comparison of standard prompting and multi-hop psychotherapeutic reasoning. The multi-hop approach integrates the client’s emotional and cognitive state (facial expressions, thoughts, and cognitive distortions) at each step of the intervention. The conversation on the left shows the therapist’s replies, which correspond to the four stages—Introduction, Guidance, Brainstorming, and Suggestion—outlined on the right.

3.4 Multi-hop Psychotherapeutic Reasoning

To ensure that the interventions are tailored to the client’s needs, professional psychotherapists typically first understand the client’s state and then provide interventions grounded on that (Greenberg and Safran, 1989; Rice and Elliott, 1996). 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. Initially, the AI therapist detects the client’s state and then generates a response based on it, as illustrated in Figure 3. In this work, we focus on three major aspects of the client’s states—*facial expression*, *thought*, and *thinking traps*—which accumulate over the stage of the conversation. Each piece of evidence is identified at the appropriate stage. The detected evidence is included in the client’s states, and the states are fed to the AI therapist as the prompt for the next evidence detection.

4 Experiments

4.1 Settings

Baseline models. We utilize two representative models for our experiments: LLaMA2-chat-7b

(Touvron et al., 2023)⁵, which is widely used in text generation tasks, and LLaVA-v1.5-7b⁶, renowned for vision-related tasks. For simplicity, we refer to LLaMA2-chat-7b as LLaMA2 and LLaVA-v1.5-7b as LLaVA in this work. In addition, we denote the versions of LLaMA2 and LLaVA that were trained on the M2CoSC dataset as CS-LLaMA2 and CS-LLaVA, respectively. When multi-hop psychotherapeutic reasoning is applied, we add MH to their names.

Hyper-parameters. Both LLaMA2 and LLaVA were fine-tuned with LoRA (Hu et al., 2022) on the M2CoSC dataset using default settings, except for the number of epochs. For LLaMA2, we used the official open-source models from Hugging Face, and for LLaVA, we followed the official code defaults⁷. We split the M2CoSC train set into 80:20 training and validation subsets to select the optimal epoch based on validation performance. All models were trained with $4 \times \text{A100-80GB GPUs}$,

⁵<https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

⁶<https://huggingface.co/liuhaotian/llava-v1.5-7b>

⁷Default settings from <https://github.com/haotian-liu/LLaVA/tree/main>

		Empathy	Logical Coherence	Guidance	Overall	Avg.
Baselines	LLaMA2	2.665*	2.390*	1.600*	1.540*	2.218*
	LLaVA	2.640*	2.570*	1.790*	1.740*	2.333*
Ours	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 w/ MH	2.980	2.960	2.510	2.490	2.817

Table 4: Dialogue-level assessment results evaluated by GPT-4 using a role-playing approach with an AI client. Stars (*) next to values indicate a significant difference compared to CS-LLaVA w/ MH, with a p-value < 0.05 determined by the paired t-test.

using a batch size of 32 per GPU.

4.2 Evaluators

GPT-4. Recent research has shown that the evaluation of natural language generation (NLG) models using GPT-4 closely aligns with human evaluations. Therefore, 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). Also, in conversation models, Zheng et al. (2023) showed that GPT-4 achieves high agreement with human judgment in evaluations, releasing the corresponding judging prompt and the used codes⁸. Building on this research, we evaluated the AI therapists using GPT-4 (API version)⁹ for evaluation in two ways:

- **Score assessment:** We adopt a three-dimensional scoring system for 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 and vice versa or if it’s a tie for all possible pairs.¹⁰

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

⁸We utilize prompts from https://github.com/lm-sys/FastChat/tree/main/fastchat/llm_judge

⁹We used the gpt-4-0613 version of the GPT-4 API.

¹⁰To ensure fairness and prevent position bias, we tested each case twice, swapping the positions each time.

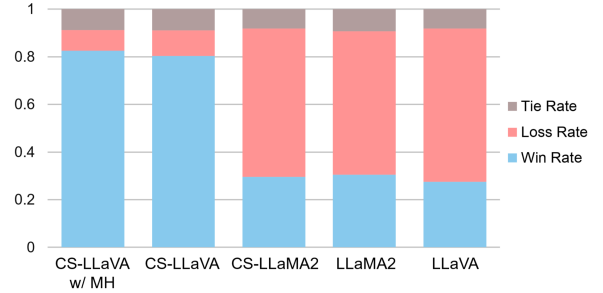


Figure 4: Dialogue-level win rates assessed by GPT-4. Detailed numerical results are provided in Appendix D.1.

5 Results and Discussions

For reliable comparison, we compared the performance of both LLMs and VLMs with two test scenarios: dialogue-level evaluation and stage-level evaluation. The dialogue-level testbed, which has been used in prior research, allows us to observe how interventions are carried out throughout conversations. However, relying solely on this testbed makes it difficult to accurately compare the AI therapists’ abilities due to the variability of the AI client. To better assess interventions in terms of empathy, logical coherence, and rationality, we also conducted a stage-level evaluation. This approach enabled us to compare therapists’ interventions more precisely by analyzing turn-level performance on the M2CoSC test set, using consistent contextual input across the models.

5.1 Dialogue-level Evaluation

In this scenario, we employ ChatGPT (API version)¹¹ as an AI client to test our approach in AI-to-

¹¹We used the gpt-3.5-turbo-0125 version of the ChatGPT API.

		Introduction			Exploration			Brainstorming			Suggestion		
		Emp.	Coh.	Gui.	Emp.	Coh.	Gui.	Emp.	Coh.	Gui.	Emp.	Coh.	Gui.
Baselines	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*
Ours	CS-LLaVA	1.87*	1.99	0.92*	2.15	2.24*	1.64	2.11*	2.25*	1.68	2.54	2.61	1.71
	CS-LLaVA w/ MH	2.11	2.16	1.02	2.23	2.39	1.60	2.27	2.39	1.79	2.59	2.67	1.80

Table 5: Stage-wise assessment results as evaluated by GPT-4 at each stage on the M2CoSC testset. **Emp.**, **Coh.**, and **Gui.** represent Empathy, Logical Coherence, and Guidance, respectively. Stars (*) next to values indicate a significant difference compared to CS-LLaVA w/ MH, with a p-value < 0.05 as determined by the paired t-test.

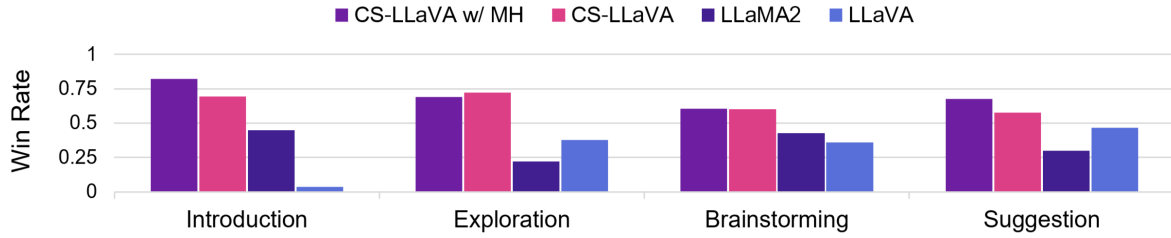


Figure 5: Stage-wise win rates assessed by GPT-4 at each stage of the M2CoSC benchmark. Numerical results are provided in Appendix D.2.

AI scenarios. For prompting to AI client, we leverage 100 resources which are used as base resources from the test set, originally sourced from Sharma et al. (2023) and Mollahosseini et al. (2019). The AI client’s role aligns with our data construction method, and we use the same prompts throughout the process. To evaluate the performance difference between using only the text modality and incorporating image information, we also compared the results of CS-LLaMA2. For CS-LLaMA2, only the text modality was used without incorporating image information.

Table 4 shows the dialogue-level assessment results evaluated by GPT-4. Our M2CoSC dataset with the LLaVA family of models led to significant improvements across all aspects. By integrating multi-hop psychotherapeutic reasoning with three implicit evidences—*facial expressions*, *thoughts*, and *thinking traps*—the models achieved enhancements across all evaluation aspects, with a particularly remarkable improvement in empathy. These results support our hypothesis that understanding the client’s emotional state before responding leads to more tailored and compassionate interactions.

LLaMA2, in contrast, shows minimal improvement when trained on the M2CoSC dataset, primarily due to the absence of visual information during the Introduction stage, which hampers effective training and results in subtle change. These findings validate our hypothesis that models in-

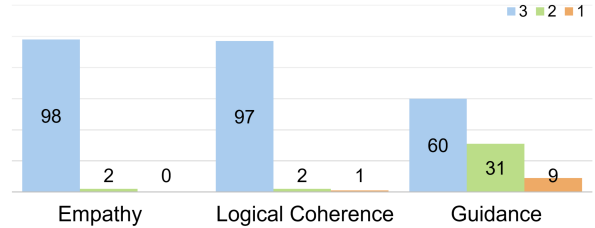


Figure 6: Dialogue-level score assessment ratio for CS-LLaVA w/ MH evaluated by GPT-4

tegrating multimodal information in counseling conversations possess superior therapeutic qualities compared to models relying solely on text. Consequently, we shifted our focus to the LLaVA family for further analysis instead of assessing CS-LLaMA2 in stage-level evaluations, except during human evaluation settings.

Similar trends were observed in the pairwise comparison (Figure 4). Despite showing the lowest performance, LLaVA exhibited a significant improvement when training on the M2CoSC and applying our multi-hop psychotherapeutic reasoning method, achieving the highest performance in CS-LLaVA w/ MH. No significant performance difference was noted between LLaMA2 and CS-LLaMA2 due to the absence of visual information, further highlighting the impact of multimodal integration in counseling conversations.

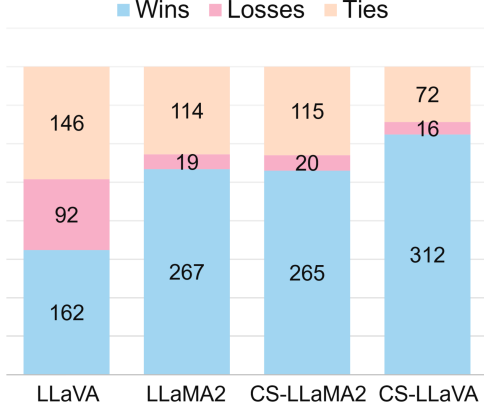


Figure 7: All stages aggregated win rates of CS-LLaVA w/ MH against other models, as evaluated by two psychotherapists on the M2CoSC test set. The domain experts are asked to choose which intervention is better at assessing the given intervention.

Error analysis Examining the aspect-wise results of our CS-LLaVA w/ MH, we found that the guidance received lower scores than the other aspects (Figure 6). Hence, we further conducted a detailed error analysis, where the primary (six out of nine) cases were the AI therapist’s failure to offer forward-looking strategies. Specifically, the AI therapist could not provide future-oriented strategies, hindering its ability to assist clients in preventing similar possible distortions. Detailed examples and additional qualitative analysis are described in Appendix E and F.

5.2 Stage-level Evaluation

In this testbed, each AI therapist responds to the same dialogue history, allowing us to directly compare their interventions. To ensure reliability, we conducted evaluations using both GPT-4 and two human psychotherapists, and the evaluation was carried out at the turn level for each stage. As in the dialogue-level evaluation, we report score assessment and win rate results.

Throughout both assessment results and pairwise comparisons, the M2CoSC dataset had a noticeable impact across all stages. Multi-hop psychotherapeutic reasoning outperformed the standard prompting method across most stages; however, in the Exploration stage, the guidance performance was slightly lower. This is likely because this stage emphasizes a deeper exploration of the client’s situation rather than offering direct suggestions.

Stage-wise assessment results evaluated by GPT-4 (Table 5) reveal that the score distribution is lower than that of the dialogue-level evaluation, as it as-

		Win Rate (%)			
		Intro.	Explo.	Brain.	Sugg.
Baselines	LLaMA2	90.0	75.5	68.0	90.0
	LLaVA	97.5	98.0	85.5	70.0
Ours	CS-LLaMA2	88.0	77.5	70.5	88.0
	CS-LLaVA	71.0	46.0	61.0	57.0

Table 6: Stage-wise win rates of CS-LLaVA w/ MH against other models, as evaluated by two psychotherapists at each stage of the M2CoSC testset.

sesses intervention at the turn level rather than the entire conversation. One key finding is the substantial performance improvement of the multi-hop psychotherapeutic reasoning process observed in the Introduction stage. The results demonstrate that our attempt to initially detect and incorporate the client’s emotions led to more empathetic and coherent interactions. Figure 5 illustrates the stage-wise pairwise comparison results among the four models, as judged by GPT-4. Note that LLaVA, without our method, had the lowest win rate during the Introduction stage. This is attributable to its difficulty in expressing empathy using the client’s non-verbal evidence, as it had not yet learned to effectively integrate multimodal information into conversations. The results highlight the importance of teaching models to recognize and utilize such information in counseling phases.

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 CS-LLaVA with standard prompting and other baselines, as evaluated by two domain experts. All stages aggregated results in Figure 7 exhibit that CS-LLaVA w/ MH achieved the highest wins and significantly fewer losses. Stage-wise results in Table 6 show that CS-LLaVA w/ MH outperforms across all stages. However, in the Exploration stage, CS-LLaVA performed similarly to CS-LLaVA w/ MH, likely due to the nature of the stage, which focuses more on exploring the client’s situation than providing suggestions.

Overall results strengthen the essence of integrating multi-hop psychotherapeutic reasoning, particularly in stages where understanding and responding to emotional and cognitive states is critical.

6 Conclusion

In this paper, we explore 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 take an initial step toward enhancing the therapeutic capabilities of AI therapists by incorporating non-verbal cues, particularly facial expressions, into the intervention process. To address the challenge of restricted access to real client data in the field of psychology, which hinders research efforts, we synthetically create a novel multimodal conversational cognitive reframing dataset called M2CoSC. Our experiments across two test scenarios, dialogue- and stage-level evaluations, exhibit significant improvements in the therapeutic capabilities of VLMs when using M2CoSC. The proposed multi-hop psychotherapeutic reasoning strategy, which integrates facial expressions, thoughts, and thinking traps, demonstrates superior performance in providing clients with empathetic, logically coherent, and specific rational suggestions.

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.

While we used benchmark images for facial expression recognition, capturing real clients' facial expressions can be challenging and may influence the consultation's content. Moreover, our study only utilized facial images as the source of non-verbal information, which presents a limitation compared 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.

Another important consideration is that facial expression recognition can vary across cultural contexts. Different cultures express and interpret emotions in distinct ways, which can affect the accuracy and fairness of emotion recognition models. Such cultural differences may introduce biases in how

virtual clients' emotions are understood and addressed, meaning our findings might not generalize well across diverse populations. Future research should work on mitigating these biases and adapting emotion recognition models to better account for cultural diversity.

Moving forward, 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.

Ethics Statement

This study explores multimodal cognitive reframing therapy constructing a synthetic dataset, M2CoSC. Importantly, no real client data was used in this work; all information in the dataset was generated synthetically.

To construct the dataset, we conducted a thorough data cleansing process with the assistance of three native English speakers, compensating them \$0.13 per data entry. For human evaluation, we engaged three professional psychotherapists. One was responsible for evaluating the quality of the M2CoSC test set, while the other two conducted pairwise comparisons. We compensated \$0.80 per conversation for the dataset evaluation and \$0.0625 per entry for the pairwise comparisons.

Additionally, to adhere to the AffectNet license, images attached in this paper were not sourced from the AffectNet dataset; instead, all images were created using DALL-E 3 (Betker et al., 2023). Furthermore, we obtained consent for all research participants, including annotators and evaluators, to ensure adherence to the license.

To address privacy concerns while complying with the license, we only partially release our M2CoSC dataset, with full access requiring an AffectNet license. Specifically, we provide synthetic dialogues paired with image IDs.

Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2025-RS-

2024-00437866, 47.5%) and Smart HealthCare Program funded by the Korean National Police Agency(KNPA) (No. RS-2022-PT000186, 47.5%), and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH), 5%).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Aaron T. Beck. 1970. [Cognitive therapy: Nature and relation to behavior therapy](#). *Behavior Therapy*, 1(2):184–200.
- Aaron T. Beck. 1987. [Cognitive therapy and the emotional disorders: A. t. beck](#). *British Journal of Psychiatry*, 150:870 – 871.
- Aaron T. Beck. 1988. [Love is never enough](#).
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8.
- Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *arXiv preprint arXiv:2401.00820*.
- Claibourne Counseling. 2024. [What are the five stages of counseling?](#) Accessed: 2024-10-08.
- 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.
- Leslie S Greenberg and Jeremy D Safran. 1989. Emotion in psychotherapy. *American psychologist*, 44(1):19.
- 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.
- Adam O Horvath and Lester Luborsky. 1993. The role of the therapeutic alliance in psychotherapy. *Journal of consulting and clinical psychology*, 61(4):561.
- 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.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- 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, Slideshare. Retrieved: 2016-07-09.
- Michael J Lambert and Dean E Barley. 2001. Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory, research, practice, training*, 38(4):357.
- 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.
- William E. Powles. 1974. [Beck, aaron t. depression: Causes and treatment](#). philadelphia: University of pennsylvania press, 1972. pp. 370. \$4.45. *American Journal of Clinical Hypnosis*, 16:281–282.

- The Psychol. 2024. [Five stages of the counseling process](#). Accessed: 2024-10-08.
- 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*.
- Laura N Rice and Robert Elliott. 1996. *Facilitating emotional change: The moment-by-moment process*.
- Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, 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.
- 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.
- 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*.
- Susan R Walen, Raymond DiGiuseppe, and Windy Dryden. 1992. *A practitioner's guide to rational-emotive therapy*. Oxford University Press.
- 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](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. [MentalLlama: 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.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao 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*.
- 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, Dublin, Ireland. Association for Computational Linguistics.

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.

C Details for Human evaluator

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.

Instructions for M2CoSC evaluation We provided domain experts with detailed instructions for evaluating the M2CoSC test set, as illustrated in Figure 13.

Instructions for human pairwise comparison

We provided the other two experts with instructions for conducting human pairwise comparisons, as shown in Figure 14. The evaluation sheet provided to them included only Dialogue 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 Detailed Numerical Results

This section provides the numerical results corresponding to the performance analyses discussed in the main text.

D.1 Dialogue-level Pairwise Comparison

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 w/ MH showed a win rate exceeding 50%, outperforming all models.

D.2 Stage-level Pairwise Comparison

In Table 9 and Table 10, we have provided the numeric results as additional data for Figure 5. When comparing each model pair, CS-LLaVA w/ MH showed a win rate that surpassed other models, similar to the dialogue-level testbed results.

E Case Study

We conducted additional analysis on test cases to compare our approaches with the LLaMA2.

The comparisons are illustrated in three figures: Figures 8, 9, and 10 show the full conversations between LLaMA2, CS-LLaVA, and CS-LLaVA w/ MH with an AI client, respectively. All three conversations were generated using the same source from the M2CoSC test set, where the client exhibits cognitive distortions, specifically overgeneralization.

In the Introduction stage, LLaMA2 primarily offers unconditional consolation, as it lacks the ability to draw on client-specific information. In contrast, both CS-LLaVA and CS-LLaVA w/ MH demonstrate a more targeted empathy by tailoring their facial expression.

When it comes to suggestions, LLaMA2 tends to focus on generic advice, like "having an open conversation with a friend," without utilizing specific cognitive reframing techniques. On the other hand, CS-LLaVA and CS-LLaVA w/ MH encourage the client to consider alternative viewpoints. Additionally, CS-LLaVA w/ MH goes even further by prompting the client to reflect on past instances where they may have made similar cognitive errors.

F Error Analysis

Based on the dialogue-level evaluation results (see Figure 6), we analyzed the cases where guidance received lower scores, specifically those rated as 1.

F.1 Failure to Provide Future Strategies

A common issue with our CS-LLaVA w/ MH model was its inability to offer strategies that would help clients address similar challenges in the future.

For instance, in Figure 11, the therapist successfully reframed the client's feelings about oversharing but neglected to suggest how to handle similar situations moving forward.

In such cases, the therapist could have recommended practical strategies, such as discussing personal issues openly with colleagues to gauge their comfort levels. Providing forward-looking strategies is essential for helping clients develop resilience and effective coping mechanisms for recurring issues.

F.2 Inadequate Crisis Management

In situations where clients are experiencing severe emotional distress, our CS-LLaVA w/ MH struggled to respond flexibly to the crisis, such as by recommending immediate crisis intervention. For example, in Figure 12, a client expressed suicidal thoughts, but the therapist failed to suggest professional help or provide resources for dealing with a crisis. This represents a serious limitation in ensuring safe guidance, especially in high-risk situations.

The staged approach we used to enhance logical consistency in the LLM's counseling conversations mandates that the therapist adhere to a predefined role at each session stage. However, this rigidity impedes the model's ability to adapt during crises. Furthermore, this staged conversation assumes that clients are willing to engage openly with the therapist; if clients display resistance or strong defense mechanisms, the model may fail to deliver effective support.

Future research should aim to improve the AI's flexibility in crisis situations while preserving the logical coherence central to its design. Addressing these limitations is crucial for enhancing the AI therapist's capacity to provide meaningful, practical, and safe guidance, especially during critical moments.

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-Dialogue 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 M2CoSC. *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 w/ MH	Win Rate
LLaMA2	-	52.551	49.495	9.694	9.794	30.485
LLaVA-7b	47.449	-	49.485	6.566	7.071	27.481
CS-LLaMA2	50.505	50.516	-	10.309	7.071	29.592
CS-LLaVA	90.306	93.434	89.691	-	47.959	80.357
CS-LLaVA w/ MH	90.206	92.929	92.929	52.041	-	82.061

Table 8: Numerical results of dialogue-level pairwise comparison of five models, evaluated using GPT-4.

Win Rate (%)				
	Introduction	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	72.119	60.256	57.617
CS-LLaVA w/ MH	82.101	68.978	60.478	67.671

Table 9: Stage-wise win rates of four models on the M2CoSC testset, evaluated using GPT-4.

	LLaMA2	LLaVA-7b	CS-LLaVA	CS-LLaVA w/ MH	Win Rate
LLaMA2	-	54.096	27.562	22.877	34.676
LLaVA-7b	45.904	-	25.710	21.676	31.179
CS-LLaVA	72.438	74.290	-	47.067	64.848
CS-LLaVA w/ MH	77.123	78.324	52.933	-	69.677

Table 10: Numerical results of stage-level pairwise comparison of four models on the M2CoSC testset, evaluated using GPT-4.

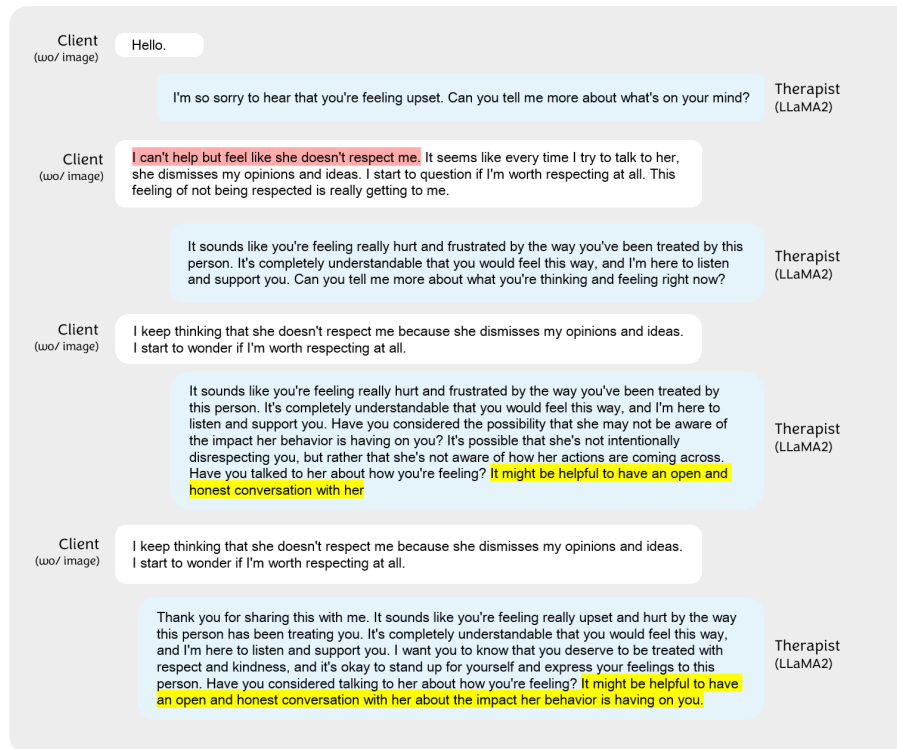


Figure 8: A case study between an AI Client (ChatGPT) and LLaMA2. The red shading highlights the client's distorted thoughts, while the yellow shading emphasizes the key components of the therapist's intervention.

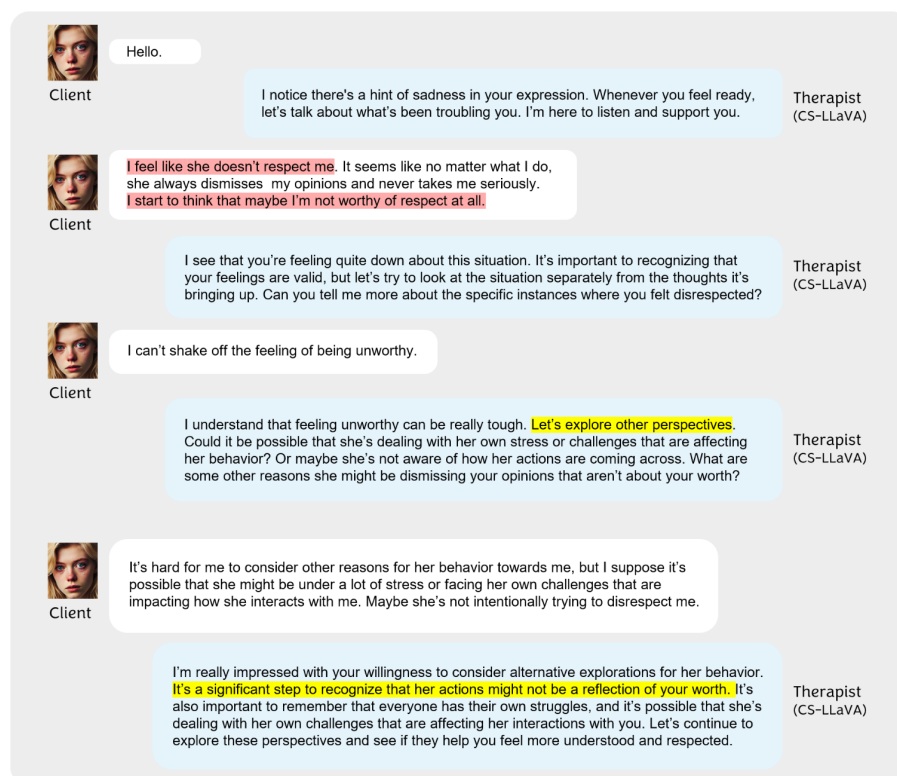




Figure 9: A case study between an AI Client (ChatGPT) and CS-LLaVA.




Client

Hello.




Client

It feels like everyone always disrespects me. 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.




Client

I'll try to think about instances where people have shown respect towards me, even if it didn't feel that way initially.




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
(CS-LLaVA w/ MH)

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?




Therapist
(CS-LLaVA w/ MH)

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?



Therapist
(CS-LLaVA w/ MH)

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. Consider this: might there have been times when someone was trying their best to be respectful, but the communication got misinterpreted? 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.



Therapist
(CS-LLaVA w/ MH)

You've made a significant step by recognizing that there may have been instances of support from colleagues that you might have missed. 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. 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 10: A case study between an AI Client (ChatGPT) and CS-LLaVA w/ MH.

Client: Hello.

Therapist (CS-LLaVA w/ MH): I notice there's a look of concern on your face. Whenever you feel ready, let's talk about what's been causing you to feel fearful. Remember, this is a safe place to share your thoughts and feelings.

Client: I feel like everyone finds me annoying and boring. I always end up oversharing and saying the wrong things. I can't seem to stop myself and I'm terrified that people will judge me for it.

Therapist (CS-LLaVA w/ MH): I see that you're feeling quite anxious about how others perceive you. Let's take a moment to distinguish between the situation itself—that is, social interactions—and the thoughts you have about it. Your thoughts are just that: thoughts, not facts. Can you tell me about a specific instance of oversharing that happened recently?

Client: I recently talked about a personal issue at a company meeting, and I worried afterwards that I shouldn't have shared so much.

Therapist (CS-LLaVA w/ MH): It's understandable to feel uneasy after sharing personal information, but it's also important to remember that everyone has their own insecurities. Perhaps your colleagues understood more than you thought and appreciated your honesty. They might even have been sympathetic or found your story engaging. Can you think of any positive feedback you might have received, or times when someone else shared personally and it didn't lead to judgement?

Client: I can recall a few instances where others have shared personal stories in meetings, and it didn't lead to any negative judgment. Actually, some people seemed to appreciate the honesty and bravery in opening up.

Therapist (CS-LLaVA w/ MH): I'm really impressed with how you've challenged the negative thoughts around oversharing. It's a sign of strength to be able to see that not all sharing leads to judgment, and it's great that you recognize this. Remember, everyone has their own insecurities, and they're not always visible. It's also important to remind yourself that being open can be a way of connecting, not just with others but also with yourself. Keep in mind that your honesty is valuable, and it's okay to express yourself as you are.

Figure 11: A failure case study of failure to provide future strategies. The case received scores of 3 for Empathy, 3 for Logical Coherence, and 1 for Guidance.

Client: Hello.

Therapist (CS-LLaVA w/ MH): I notice there's a look of sadness on your face. Whenever you feel ready, I'm here to listen and support you through whatever is weighing on you. What's been happening that's caused you to feel this way?

Client: I feel so overwhelmed and hopeless. I can't see a way out of this pain. I keep thinking that I want to die because it feels like there's no other option to escape this suffering.

Therapist (CS-LLaVA w/ MH): I see the pain and fear in your eyes, and I hear the despair in your words. It's important to recognize that these feelings are overwhelming, but they are not the entirety of your situation or your thoughts. Let's try to take a step back and look at the moments of your life that are causing you pain. Can you tell me about one specific event that has been happening recently that may be contributing to these feelings?

Client: I've been feeling so overwhelmed by the stress at work lately. It feels like no matter how hard I try,

Therapist (CS-LLaVA w/ MH): I see that you're feeling overwhelmed, and that's a completely valid feeling. Let's try to explore other thoughts that may be there too. Could there be moments at work where things have gone right, despite the stress? Maybe you've learned something valuable, or perhaps you've done something that helped a colleague. Even small victories are important in such situations. Is there a hobby or a side project you're passionate about that you can think about? Sometimes, finding a balance between work and other aspects of life can give us a new perspective.

Client: You know, now that you mention it, there was a project at work that I was really proud of last month. Despite all the stress, I managed to lead a successful team effort that resulted in positive feedback from our clients. I hadn't thought about that in a while. And outside of work, I do enjoy painting in my free time. It's a great stress reliever and helps me unwind. Thank you for reminding me of these positive moments. I guess there are some bright spots even in the midst of all the chaos.

Therapist (CS-LLaVA w/ MH): I'm really proud of you for recognizing those positive moments, like leading your team to success and enjoying painting as a hobby. It's important to acknowledge these accomplishments and pleasures in your life. They may seem small in the face of your current challenges, but they are significant. They are evidence that you have strengths and interests outside of the stress you're experiencing. Whenever you can, try to focus on these positive activities and remember your successes, as they are a testament to your resilience and ability to navigate through difficult times.

Figure 12: A failure case study of inadequate crisis management; highlights client's expression of suicidal thoughts. The case received scores of 3 for Empathy, 3 for Logical Coherence, and 1 for Guidance.

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 troubling 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 haven'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.



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 13: 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?



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?	It's understandable to feel upset after mistakes, but let's consider that making mistakes is a part of learning and growth. Can you think of any improvements you've made or lessons you've learned from these mistakes?	
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 have communicated your feelings and boundaries more effectively. By exploring 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? Perhaps you're being harsh on yourself by using words like "mistake". Reflect on what you've learned 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 the 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 "frustrated", 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 hair right now. Whenever you feel ready, we can talk about what is causing you to feel this way. Remember, this is a safe 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?	
	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 14: Instruction for human pairwise comparison.