

CRISP: Cognitive Restructuring of Negative Thoughts through Multi-turn Supportive Dialogues

Jinfeng Zhou^{1*} Yuxuan Chen^{1*} Jianing Yin² Yongkang Huang³ Yihan Shi⁴
Xikun Zhang⁵ Libiao Peng³ Rongsheng Zhang⁶ Tangjie Lv⁶
Zhipeng Hu⁶ Hongning Wang¹ Minlie Huang^{1†}

¹The CoAI Group, DCST, Tsinghua University

²University of Pennsylvania

³Lingxin AI

⁴Harvard Graduate School of Education, Harvard University

⁵Department of Psychology and Behavioral Sciences, Zhejiang University ⁶Fuxi AI Lab
zjf23@mails.tsinghua.edu.cn {hw-ai, aihuang}@tsinghua.edu.cn

<https://peppy-ai.com/>

Abstract

Cognitive Restructuring (CR) uses multi-turn dialogue to identify and restructure one’s negative thoughts, arising from mental health issues, into more helpful and positive ones. Clinician shortage and stigma urge the development of human-LLM interactive psychotherapy for CR. Yet, effectively implementing CR is hindered by entrenched cognitive distortions, emotional resistance, and individual differences, which existing works have not overcome. To bridge this gap, we propose CRDIAL, a novel framework that structures CR as theory-grounded multi-stage multi-turn dialogue, integrating multi-aspect supportive strategies for emotional management and a multi-channel loop mechanism to account for diverse individual distortions. With CRDIAL, we distill CRISP, a large-scale and high-quality bilingual dialogue dataset, from LLM. We then train CRISPERS, CRISP-based conversational LLMs for CR, at 7B and 14B scales. Extensive human studies show the superiority of CRISPERS in pointwise, pairwise, and intervention evaluations.¹

1 Introduction

Negative thoughts arise from cognitive distortions that create a distorted view of reality and are entrenched in individuals facing mental health issues, which may lead to depression and even suicidal ideation (Beck, 1979). A proven intervention for such thoughts is **Cognitive Restructuring (CR, Beck 2011)**. As shown in Figure 1, CR is a therapeutic process where a therapist would first initiate a dialogue from one’s negative situation (e.g., “*receiving numerous comments from his/her advisor*”) and progressively explores their thoughts from surface (e.g., “*I just feel more and more lost*”) to core negative thought (e.g., **catastrophizing**: “*Even if*

^{*}Equal contribution.

[†]Corresponding author.

¹Repository: <https://github.com/thu-coai/Crisp>



Figure 1: A data example of CRISP crafted via LLMs using CRDIAL, which *identifies* and *restructures* multiple negative thoughts (i.e., *cognitive distortions*) via multi-stage multi-turn dialogue with emotional support.

I finish the thesis, I still won’t have a successful career”). The therapist should then apply restructuring techniques to transform this catastrophizing thought into a more positive one (e.g., “*The fear is normal and will motivate me to plan career goals*”).

Yet, barriers like clinician shortages and stigma often limit people’s access to therapists (Olfsen, 2016), urging the development of human-machine interactive psychotherapy for CR. But the process of CR often faces several key challenges which make it non-trivial for machines to successfully carry out CR. **(1) Cognitive Persistence:** Cogni-

tive distortions are rooted in individuals’ thought patterns, impeding their self-awareness of negative thoughts (Beck, 1979). **(2) Emotional Management:** Engaging in CR may cause individuals to feel discomfort or resist changing their entrenched thought patterns (Agarwal and Sirts, 2025; Kim et al., 2025c), thus evoking intense emotions and disrupting the therapeutic process. **(3) Individual Differences:** Individuals’ personalities shape their cognitive styles (Sperry and Sperry, 2020), resulting in diverse thought patterns, e.g., in Figure 1, an individual may also experience cognitive distortions like *overgeneralization* (“*all my efforts were wasted*”) or *catastrophizing* (“*I still won’t have a successful career*”), or even both. Professional therapists have to navigate these challenges adaptively.

The fast development of large language models (LLMs, Touvron et al. 2023a,b) ignites researchers’ hope for LLM-based CR solutions, but existing research in this line still remains incompetent in addressing the above challenges, following three main methods: **(1) Text Rewriting** (Sharma et al., 2023; Maddela et al., 2023): Implementing CR simply as a text rewriting task, where reframed positive thoughts are directly presented to individuals. The lack of a progressive cognitive guidance process may cause individuals to perceive the thoughts as preachy or coercive, leading to resistance. **(2) Fixed-Pattern Dialogue** (Na, 2024; Goel et al., 2024; Kim et al., 2025a): Using a rigid conversational pipeline, e.g., fixed three-turn dialogue (Xiao et al., 2024). This mechanical dialogue flow is ineffective at emotional management, thus harming the effectiveness of CR. **(3) One-shot CR Workflow:** Following the “one interaction addresses one cognitive distortion” paradigm (Lee et al., 2024), which fails to account for individual differences that often manifest as multiple overlapping or interconnected distortions in a single mental health issue, potentially leading to fragmented or superficial therapy.

To address the challenges, we propose CRDIAL, a novel framework that structures CR as a theory-grounded multi-stage multi-turn DIALogue, integrating multi-aspect supportive strategies for emotional management and a multi-channel loop mechanism to account for individual differences. Specifically: **(1) Multi-stage CR.** To align with realistic psychotherapy for CR, we design multi-turn dialogues featuring two stages: i) Cognitive Theory (CT, Beck 1979)-guided identification of negative thoughts, and ii) Defense Attorney Technique (DAT, de Oliveira 2011)-driven restructuring of

negative thoughts. **(2) Multi-aspect Supportive Conversation Strategies:** To provide emotion support during CR, we draw on psychological theories (Hill, 2009; Linehan, 2014) to design a scheme of supportive conversation strategies for CRDIAL. This scheme involves 5 strategy categories across 8 sub-categories to carefully constrain the therapist’s responses at the sentence level. **(3) Multi-channel Loop Mechanism:** This mechanism is designed for flexibly identifying individualized cognitive distortions and a loop mechanism to cycle the CR process, enabling iterative addressing of potential multiple negative thoughts in a mental health issue.

With CRDIAL, we distill CRISP from LLMs (i.e., GPT-4o, OpenAI 2024), which is a large-scale bilingual dialogue dataset with 22k dialogues, designed to simulate CR in multi-stage multi-turn supportive dialogues (§3). Strict quality control procedures are implemented to ensure the quality, safety, and professionalism of CRISP (§4). With CRISP, we train CRISPERS, CRISP-based convERSation LLMs for CR, at 7B & 14B scales (§5). Manual evaluations show CRISPERS-14B outperforms its teacher model in both interactive pointwise and pairwise evaluations. Our psychological intervention trial also further highlights its potential for human-LLM interactive psychotherapy (§6).

2 Related Work

LLMs are powerfully driving the advancements of human-computer interactive psychotherapy (Chen et al., 2023; Qiu et al., 2024a,b; Zhang et al., 2024a,b; Qiu and Lan, 2024). In this field, a typical paradigm is empathetic dialogue (Rashkin et al., 2019), incorporating affective (Lin et al., 2019; Li et al., 2020) and cognitive (Zhou et al., 2023) empathy. Another milestone is that introducing emotional support dialogue (Liu et al., 2021) to regulate one’s emotional distress. Yet, they mainly draw on emotional management, failing to address cognitive distortions central to mental health issues.

To bridge this gap, existing studies (Qi et al., 2025; Kim et al., 2025b) introduce CR based on three main methods. The first is text rewriting (Ziems et al., 2022; Maddela et al., 2023; Sharma et al., 2023), which rewrites one’s negative situation into a more positive view. The second, fixed-pattern dialogue, restricts CR to a rigid dialogue flow (Goel et al., 2024; Kim et al., 2025a), e.g., single- or three-turn dialogue (Na, 2024; Xiao et al., 2024). The third is a one-shot CR workflow, which

addresses a single cognitive distortion within one interaction (Lee et al., 2024). Yet, these methods remain ineffective at managing complex emotional dynamics or individual differences inherent during CR, potentially harming their overall effectiveness.

3 CRDIAL Framework

In Figure 2, we follow clinical therapy practices and work with mental health experts to develop the CRDIAL framework when crafting dialogues between an individual and a therapist from LLMs via self-play. The process is as follows: (1) The dialogue begins with an individual’s help-seeking situation, with the constraints defined by our supportive conversation strategies of the therapist’s responses. (2) Cognitive theory (CT) guides the therapist to identify the individual’s surface-to-core negative thoughts (i.e., cognitive distortions) via multi-turn dialogues. (3) A courtroom analogy is then adopted to transform the identified cognitive distortions until a verdict of successful restructuring is reached. (4) CRDIAL allows to identify multiple cognitive distortions, i.e., multi-channel loop mechanism, to adapt individual differences.

3.1 Multi-aspect Supportive Strategies

Supportive Strategy Scheme To provide emotional comfort and prevent resistance during CR, we develop a supportive strategy scheme using Hill’s Help Skill Theory (Hill, 2009) and Dialectical Behavior Therapy (Linehan, 2014), including strategies with 5 categories across 8 subcategories: (1) **Description**: Helping individuals objectively describe their situations by *questions* and *restatement*; (2) **Expression**: *Reflecting feelings* and using *self-disclosure* to express empathy; (3) **Assertion**: *Providing suggestions* and *information* to guide positive thoughts and behaviors; (4) **Reinforcement**: Enhancing individuals’ agency by *affirmation* and *reassurance*. (5) **Negotiation**: *Negotiating* with individuals when they present challenges. Details of these strategies are shown in Appendix B.1.

Sentence-Level Strategy Constraints These strategies guide the therapist’s responses throughout the dialogues, ensuring CR consistently accompanies emotional support. To enhance responses’ supportiveness, we prompt GPT-4o to use sentence-level constraints, where each sentence of response is generated paired with an appropriate strategy. Details of the prompt are shown in Appendix A.1.

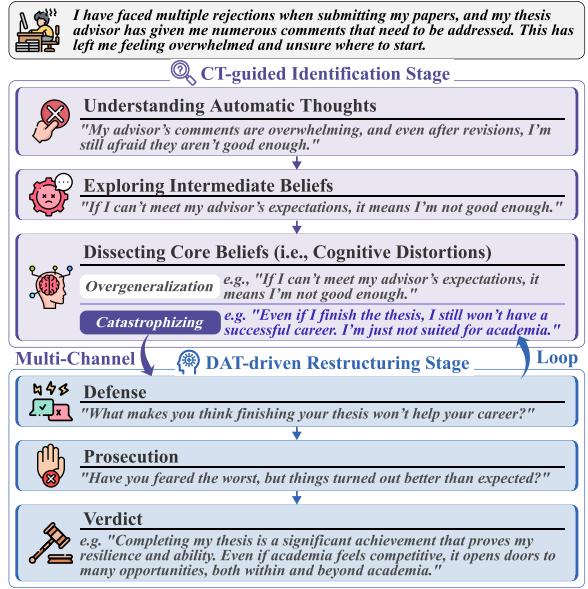


Figure 2: The CRDIAL framework is used to guide the distillation of dialogues from LLMs stage by stage.

3.2 CT-guided Identification Stage

Help-seeking Situation Collection The identification stage begins with the individual’s help-seeking situation. To ensure the authenticity of situations, we crawl anonymous posts from [Yixinli](#), a famous Chinese mental health service platform, and the subreddit “[r/mentalhealth](#)”. These posts offer detailed descriptions of individuals facing mental health challenges in daily life. We carefully remove sensitive information from these posts and summarize them into *help-seeking situations* by prompting GPT-4o. Furthermore, to enhance the individual differences in dialogue interactions across various situations, we prompt GPT-4o to extract distinctive *personality* from the posts, which is then used to guide individuals’ behavior in dialogue.

CT-guided Identification Process An individual’s distorted thoughts are often rooted in the situations they encounter. We draw on cognitive theory to realize the process of identifying cognitive distortions that contribute to mental health challenges via multi-turn dialogues. This process often progressively probes 3 levels of thoughts: (1) **Automatic thoughts**: surface-level thoughts that occur spontaneously in responses. (2) **Intermediate beliefs**: underlying assumptions that influence automatic thoughts and help interpret situations. (3) **Core beliefs**: deeply ingrained *cognitive distortions* that shape the foundation of an individual’s thinking.

Implementation We prompt GPT-4o to realize self-play, i.e., playing both a strategy-constrained

therapist and a personality-constrained individual to dissect the individual’s distorted thoughts in a multi-turn dialogue. We found that generating a long multi-turn dialogue all at once tends to produce generic content. Thus, we divide the identification stage into two steps: (i) *reassurance of automatic thoughts and intermediate beliefs* and (ii) *dissection of core beliefs*. Details of the prompts used in this stage are provided in Appendix A.2.

3.3 DAT-driven Restructuring Stage

DAT-driven Restructuring Process We use the well-known Defense Attorney Technique (DAT, de Oliveira 2011) to restructure the identified distorted thoughts. DAT is a CBT tool that uses the metaphor of a court trial to encourage individuals to play “defense lawyers”, making their challenges to negative thoughts clear and concrete by defending more positive thoughts. DAT drives restructuring process by using three mechanisms: (1) **Defense**: The individual, as the defense lawyer, is guided by the therapist to consider factual evidence that supports their negative thoughts. (2) **Prosecution**: The therapist, as the prosecutor, identifies flaws in the evidence and provides counter-evidence to challenge the individual’s negative thoughts. (3) **Verdict**: After the prosecution, assess whether the individual’s distortions have been successfully restructured into more positive and helpful thoughts.

Implementation Using the dialogue history from the previous stage as input, we prompt GPT-4o to determine the next appropriate dialogue phase (*Defense* or *Prosecution*) and then engage in self-play to generate multi-turn dialogue between the therapist and the individual. This process iterates, with each generated dialogue added to the dialogue history, until the *Verdict* phase reports that restructuring is completed. We grant flexibility in determining the number of iterations to emulate the complex alternation between *Defense* and *Prosecution* in real-world courts. Details of prompts used in this stage are provided in Appendix A.3.

3.4 Multi-channel Loop Mechanism

Multi-channel Mechanism Case conceptualization suggests that personal traits shape individuals’ different cognitive styles, affecting how they perceive, interpret, and respond to situations (Sperry and Sperry, 2020). Thus, unlike traditional single-channel dialogues (*one query corresponds to one response*), we design a multi-channel mechanism

to infer multiple possible cognitive distortions in the “*dissection of core beliefs*” step. Each distorted thought creates its own channel, with a dedicated restructuring process. In practice, we prompt GPT-4o to adaptively infer multiple (≤ 3) channels based on the dialogue history, each in the form of “*therapist identifies a cognitive distortion and individual responds to confirm*”. Each channel then independently proceeds to its own restructuring stage. Details of the prompt are provided in Appendix A.2.

Loop Mechanism After completing a restructuring stage, we prompt GPT-4o to reason whether the individual may still maintain other cognitive distortions, thus looping the dialogue back to the identification stage. This loop mechanism is critical for two reasons. First, individuals facing mental health issues often exhibit multiple irrational thinking patterns (Beck, 1979). The multi-channel mechanism sets the order of restructuring, while the loop mechanism improves CR’s thoroughness. Second, an individual’s cognitive patterns in one situation may automatically trigger related distorted thoughts from other situations (Bargh and Chartrand, 1999). This requires an iterative identification and restructuring process, ensuring all relevant cognitive distortions are addressed, thus enhancing the intervention’s effectiveness (Beck, 2011). Details of the prompt are shown in Appendix A.4.

3.5 Integrating Commonsense Knowledge

Following Kim et al. (2023), we integrate external commonsense knowledge into each sub-step of identification and restructuring, providing two advantages to LLM-based dialogue construction: (1) *Reducing nonsensical conversations*: Sampling naively from an LLM without external information can lead to redundant and generic dialogues (Kim et al., 2023). (2) *Improving diversity*: External information increases topic diversity, leading to varied dialogue dynamics. We use Atomic^{10x} as our commonsense knowledge graph (Hwang et al., 2021), which describes events using (*head, relation, tail*) triples, e.g., (a situation is “I am struggling to finish my thesis”: *Head: PersonX struggles with thesis, xReact: PersonX feels overwhelmed*). Before each sub-step begins, we summarize the individual’s situation from the dialogue history as the head and use relations related to mental health (*intention, desire, reaction, need*) to infer commonsense knowledge. The inferred knowledge is then used to guide current sub-step dialogue generation.

Datasets	Human-curated	Strategy Constrain	Specialized for CR	Language	#Dialog	#Avg. Turns	#Avg. Uttr. Length	#Lexical Diversity
ESConv (2021)	✓	Utterance-level	✗	En	1,300	13.40	18.70	34.27
AugESC (2023a)	✗	Utterance-level	✗	En	65,077	14.50	18.80	30.12
ExTES (2023b)	✗	Utterance-level	✗	En	11,177	9.10	26.00	50.66
Healme (2024)	✗	✗	✓	En	1,300	3.00	63.20	48.21
Cactus (2024)	✗	✗	✓	En	31,577	16.60	26.93	58.66
 CRISP	✗	Sentence-level	✓	En & Zh	22,063	36.48	38.12	70.51

Table 1: Statistics of CRISP compared to related dialogue datasets. Utt. is utterance. More details are in App. B.5.

4 CRISP Dataset

Based on CRDIAL, we distill CRISP, a large-scale high-quality dialogue dataset for CR, from GPT-4o using 2,985 seed situations. We break each channel into individual dialogues for quality control. The prompts used for quality control with LLMs (i.e., GPT-4o) are provided in Appendix B.2.

4.1 Postprocessing for Quality Control

Basic Filtering Starting with an initial set of 27k dialogues, we prompt GPT-4o to filter dialogues with 3 basic quality criteria: (a) Remove dialogues with unnatural (e.g., *mechanical phrasing*), inappropriate (e.g., *off-topic expressions*), or erroneous (e.g., *repetition, contradiction*) statements (6.34%). (b) Remove dialogues that exhibit incoherent or illogical social dynamics, especially between sub-steps of CRDIAL (2.54%). (c) Remove dialogues with statements that violate commonsense (2.15%).

Safety Filtering The seed situations used for dialogue construction have already undergone a strict manual safety filtering, with a removal rate of 29.8%. To further prevent harmful or dangerous content in the created dialogues, we apply two safety filters: Canary and LLM. First, we employ Canary, a dialogue safety classification model (Kim et al., 2022), to remove dialogues marked as needing intervention (e.g., crimes, emergencies). Next, we prompt GPT-4o to identify unsafe content and remove any dialogues with unsafe tendencies. Finally, about 0.02% of the dialogues are removed.

Expert Filtering To ensure the dialogues reflect a professional CR process, we define three expert criteria for filtering: (1) **Therapist Standard** (5 metrics): Evaluate whether the therapist’s responses are *specific* to the help-seeker’s situation and *empathetic* to their emotional state, i.e., whether they provide *rational* evidence, *positive* perspectives on the struggles, and *actionable* suggestions to help overcome negative thoughts during CR (Sharma et al., 2023). (2) **Help-Seeker Standard** (9 metrics): Evaluate the authenticity of the

help-seeker’s responses using: *simplicity, colloquialism, mistrust, details, lack of self-awareness, staying in role, resembling typical case, challenging the psychologist, and readiness as trainee*. These metrics are summarized from behaviors observed in real help-seekers during therapy (Louie et al., 2024). (3) **Supervisor** (2 metrics): Evaluate whether the overall dialogue aligns with the *normative* principles of CR with emotional support and the *effectiveness* of CR within the dialogue. Details of each criterion are in Appendix B.2.2. We prompt GPT-4o to evaluate the dialogues on these three expert criteria, scoring each criterion from 1 to 5. We calculate 3 average scores for each dialogue, one for each expert criterion, and remove any dialogues with an average score below 3.5, resulting in an 11% removal rate. Human experts inspect 500 randomly sampled dialogues from the retained set, with a 95% pass rate, showing the high quality of the retained dialogues.

Final Dataset After all filtering steps, 77.95% of the initial dialogues remain, resulting in a final dataset of 22,063 bilingual dialogues in CRISP.

4.2 Comparing CRISP with Related Dialogue Datasets

Bilingual and Large-Scale CRISP is the first bilingual dataset with large-scale multi-stage multi-turn supportive dialogues for CR. It contains 22k dialogues from 3k situations, with over 796k utterances and an average of 36.48 turns per dialogue (Table 1). Our multi-channel loop mechanism generates an average of 7.39 dialogues per situation, each with an average of 2.28 loops (App. B.5). It shows that CRISP is well-suited to address individual differences with diverse cognitive distortions.

Fine-grained Annotations of Strategy and Cognitive Distortion CRISP is the first dataset for CR to specify sentence-level strategy constraints, with each therapist’s response containing an average of 2.23 strategies. We also prompt GPT-4o to generate the distortion type addressed by each looped CR,

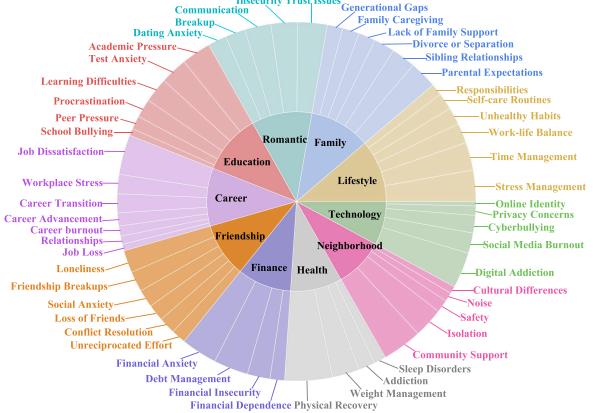


Figure 3: Distributions of mental health situations in our CRISP, with 10 categories across 54 sub-categories.

Datasets	Sensi.	Speci.	Support.	Help.	Trust.	Overall
ESConv (2021)	3.43	3.23	3.32	3.20	3.21	3.26
AugESC (2023a)	3.51	3.16	3.00	2.83	3.23	3.03
EXTES (2023b)	4.28	3.83	3.79	3.60	3.75	3.90
Healme (2024)	4.20	4.06	3.69	3.66	3.84	3.93
Cactus (2024)	4.10	4.08	4.30	4.02	4.16	4.10
 CRISP	4.38[†]	4.42[‡]	4.40[‡]	4.14[‡]	4.26[†]	4.36[‡]
<i>kappa</i>	0.61	0.70	0.66	0.62	0.65	0.59

Table 2: Results of CRISP compared to related dialogue datasets on 6 criteria: Sensibleness (*Sensi.*), Specificity (*Speci.*), Supportiveness (*Support.*), Helpfulness (*Help.*), Trustworthiness (*Trust.*), and *Overall*. The agreement ratio $kappa \in [0.61, 0.8]$ denotes the substantial agreement. “ \dagger/\ddagger ” is the statistical significance for $p < 0.05/0.001$ compared with the best baseline. **Bold** is the best result.

their distribution within CRISP is reported in Appendix B.4. To verify the accuracy of the strategy and distortion labels, we randomly sample 200 examples (evenly split by en&zh) and hire 3 psychological experts to inspect them. Results show that the accuracy for the strategy labels is 97.6% and for the distortion labels is 85.5%, with substantial inter-annotator agreement ($kappa=0.712\&0.681$, Fleiss 1971). This indicates that the fine-grained annotations provided by the LLM are highly reliable.

Diverse Situations and Lexicon CRISP covers 10 categories across 54 subcategories of situations where individuals face mental health challenges (Figure 3). The diverse situations help train dialogue models to perform well in out-of-domain settings, as supported by our experiment results. Additionally, we calculate MTL (McCarthy and Jarvis, 2010) to measure CRISP’s lexical diversity. In Table 1, CRISP shows the highest lexical diversity, 2.1 times that in other manual dataset (i.e., ESConv). This is attributed to the introduction of

external commonsense knowledge across diverse situations, leading to a broader range of words.

High Quality To quantitatively evaluate CRISP’s quality, we compare it with 3 emotional support conversation datasets and 2 dialogue datasets for CR (Table 2). We randomly sample 50 dialogues from each dataset and score them using six criteria (especially the therapist’s response): (1) **Sensibleness**: makes sense and contains no contradictions. (2) **Specificity**: coherent with the context and deep in content. (3) **Supportiveness**: effectiveness of emotional support. (4) **Helpfulness**: effectiveness in guiding positive thoughts. (5) **Trustworthiness**: the therapist’s appeal and credibility. (6) **Overall** quality. Each criterion is scored from 1 to 5. We hire five annotators to score these datasets and take the average score as the result. As shown in Table 2, CRISP achieves the highest scores on all criteria and outperforms human-curated ESConv dataset by a large margin, showing the superiority of our CRDIAL in generating high-quality dialogues.

Cost In comparison to crowdsourced data with shorter dialogue turns, generating one dialogue with the multi-channel loop mechanism using GPT-4o costs about \$0.20 and takes <1 minute. This shows that collecting CRISP via CRDIAL framework is significantly more cost- and time-efficient.

5 CRISPERS: CRISP-based Conversation LLMs for CR

With CRISP, we train CRISPERS, CRISP-based conVERSation LLMs, aimed at generating supportive responses to improve one’s mental state by restructuring distorted thoughts. CRISPERS uses Qwen-2.5-7B & 14B (Qwen et al., 2025) as backbone models and has two optimization objectives.

Sentence-level Strategy-controlled Generation Given the multi-turn dialogue history $\mathcal{C}_{n-1} = \{u_1, y_1, \dots, u_{n-1}, y_{n-1}, u_n\}$ between help-seeker u and therapist y , next response $y_n = f_s(y_n) = \{s_{n,1}, y_{n,1}, \dots, s_{n,m}, y_{n,m}\}$, where $s_{n,i}$ is a special token representing a supportive strategy and $y_{n,i}$ is the i^{th} sentence in the therapist’s response y_n . CRISPERS learn to generate responses with sentence-level strategy control by:

$$L_{strategy} = P_{s,\theta}(f_s(y_n) \mid \mathcal{C}_{n-1}), \quad (1)$$

where P represents the conditional probability (i.e., for calculating the likelihood), θ represents the

Models	Sensi. en/zh	Speci. en/zh	Support. en/zh	Help. en/zh	Trust. en/zh	Overall en/zh
Qwen-2.5-14B	3.50/3.50	3.22/3.72	3.22/3.78	3.61/3.61	3.44/3.78	3.33/3.83
CRISPERS-7B	3.44/3.83	3.44/3.24	3.45/3.82	3.50/4.06	3.56/4.12	3.56/3.83
GLM-4	3.67/3.61	3.28/3.66	3.50/3.55	3.89/3.72	3.56/4.22	3.67/4.11
Qwen-2.5-72B	4.00/4.05	3.83/3.82	4.00/3.76	3.94/4.06	4.16/4.12	4.11/4.29
GPT-4o	4.33/3.88	4.00/3.75	4.05/4.00	4.27/4.00	4.21/4.22	4.22/4.30
CRISPERS-14B	4.36/4.11[†]	4.14[‡]/4.06[‡]	4.11[†]/4.06[†]	4.31[†]/4.22[‡]	4.31[†]/4.61[†]	4.28[†]/4.33
- w/o SSCG	3.44/3.78	3.56/3.78	3.56/3.78	3.78/3.72	3.72/4.28	3.78/4.06
- w/o MDI	4.11/3.75	3.50/3.78	4.00/3.94	4.00/3.61	3.94/4.06	4.06/4.28
- w/o SSCG&MDI	3.85/3.70	3.47/3.76	3.54/3.78	3.75/3.61	3.67/3.85	3.74/3.90

Table 3: Results of interactive pointwise evaluation. w/o refers to removing the component from CRISPERS-14B.

model’s parameters, and s is meant to indicate that this probability is conditioned on generating strategy-aware sentences.

Multi-channel Distortion Identification Given $\mathcal{C}_{n-1} = \{u_1, y_1, \dots, u_{n-1}, y_{n-1}, u_n\}$ and the next response $y_n = \{d_i, f_s(y_n)\}$, where $y_{n-1} = \{d_1, y_{n-1,1}, \dots, d_k, y_{n-1,k}\}$ ($k \leq 3$), $\{d_i, y_{n-1,i}\}$ is i^{th} channel, d_i denotes a special token for a distortion type (e.g., [Catastrophizing], [Overgeneralization]), $y_{n-1,i} = f_s(y_{n-1,i})$ is the response of i^{th} channel. CRISPERS learns to generate multi-channel responses with identified distortions through a standard supervised fine-tuning objective (i.e., minimizing negative log-likelihood):

$$L_{channel,1} = P_{c,\theta,1}(d_1, f_s(y_{n-1,1}), \dots, d_k, f_s(y_{n-1,k}) \mid \mathcal{C}_{n-2}), \quad (2)$$

$$L_{channel,2} = P_{c,\theta,2}(d_i, f_s(y_n) \mid \mathcal{C}_{n-1}), \quad (3)$$

where $P_{c,\theta,1}$ and $P_{c,\theta,2}$ are optimized parameters.

Joint Training Our CRISPERS is finally jointly trained by optimizing L_{joint} loss:

$$L_{joint} = L_{strategy} + L_{channel,1} + L_{channel,2}. \quad (4)$$

6 Experiments

We compare CRISPERS with 4 representative LLMs: 1) closed-source GPT-4o (OpenAI, 2024) and GLM-4 (GLM et al., 2024), and 2) open-source Qwen-2.5-14B&72B (instruction version). Each LLM is prompted by CRDIAL framework, which is carefully refined to adapt these LLMs for CR with multi-stage multi-turn supportive dialogue, ensuring a fair comparison (prompts are in Appendix C.2). To ensure that our evaluation results can more accurately reflect the real-world client experience, we recruit volunteers for manual evaluation, with model names anonymized during this process. **More results are reported in Appendix C.**

6.1 Interactive Pointwise Evaluation

To evaluate how well the models improve one’s mental state via multi-turn dialogues, we use six criteria defined in §4.2: **Sensibleness**, **Specificity**, **Supportiveness**, **Helpfulness**, **Trustworthiness**, and **Overall** quality. Here, we respectively recruit 10 volunteers for English and Chinese evaluations, each tasked with creating two dialogues to interact with nine models (including CRISPERS-14B’s ablations) for at least 30 dialogue turns. Each volunteer selects two topics with mental health challenges from different situations, and all volunteers’ dialogue topics are evenly distributed across 10 situations (Figure 3). After the interactions, volunteers score the models on six criteria using a 1-to-5 scale. We average the scores of each model as the result.

Overall Results The results are shown in Table 3. **First**, CRISPERS-14B slightly exceeds its teacher model, GPT-4o, on all criteria, showing that the crafted CRISP, which undergoes rigorous quality control, enables small LMs to achieve high performance of larger LMs (Mehta et al., 2024). **Second**, CRISPERS-14B significantly ($p < 0.05/0.001$) outperforms all baselines on supportiveness, helpfulness, especially trustworthiness. This shows that CRISPERS provides responses in CR that can better support users’ emotional state, earning their trust to help them embrace positive thinking. **Third**, CRISPERS-14B consistently exhibits superior performance in the bilingual evaluation, outperforming its backbone model by a large margin. This shows its ability to effectively learn the interactive patterns of the CRDIAL framework from CRISP, as well as its potential to offer bilingual CR services.

Ablation Study As shown in Table 3, **first**, we remove sentence-level strategy-controlled generation, i.e., w/o SSCG. All criteria drop significantly, indicating that fine-grained control of multiple strategies in sentence-level responses is crucial to improve the overall dialogue quality. **Second**, we remove multi-channel distortion identification, i.e., w/o MDI. Both helpfulness and trustworthiness decrease significantly, highlighting that accurately identifying an individual’s cognitive distortions plays a crucial role in helping them overcome negative thoughts and earning their trust. **Third**, we remove both components, i.e., w/o SSCG&MDI. The significant decline in all criteria further validates the importance of these two components in performing CR with multi-turn supportive dialogue.

vs.	CRISPERS-14B	Family						Education						Overall					
		En			Zh			En			Zh			En			Zh		
		Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
Qwen-2.5-14B	57.11[‡]	4.44	38.45	57.20[‡]	3.91	38.89	52.67[‡]	11.50	35.83	53.14[‡]	9.74	37.12	50.12[‡]	6.81	43.08	51.10[‡]	9.89	39.01	
CRISPERS-7B	54.92[‡]	4.82	40.26	55.61[‡]	3.14	41.25	53.97[‡]	3.17	42.86	54.11[‡]	4.03	41.86	50.48[‡]	6.54	42.98	51.29[‡]	6.14	42.56	
GLM-4	56.20[‡]	2.50	41.30	53.49[‡]	6.41	40.10	51.77[‡]	6.68	41.55	52.46[‡]	6.19	41.35	47.97[‡]	10.51	41.51	48.17[‡]	9.75	41.97	
Qwen-2.5-72B	55.37[‡]	8.26	36.36	54.16[‡]	9.11	36.73	50.11[‡]	8.57	41.32	51.19[‡]	6.54	42.26	48.55[‡]	5.68	45.77	48.68[‡]	6.29	45.03	
GPT-4o	53.23[‡]	11.29	35.48	52.46[‡]	5.88	41.66	50.76[‡]	9.89	39.34	50.57[‡]	5.14	44.29	48.60	5.27	46.13	47.26[‡]	7.99	44.75	

Table 4: Results of interactive pairwise evaluation (%) in family and education situations and overall performance.

Models	Positive Affect Schedule								Negative Affect Schedule							
	Scores		Changes(%)		Student <i>t</i> -test		Tukey HSD		Scores		Changes(%)		Student <i>t</i> -test		Tukey HSD	
	Pre	Post	Improve↑		<i>p</i> ↓		<i>p</i> ↓		Pre	Post	Drop↑		<i>p</i> ↓		<i>p</i> ↓	
Emohaa	23.04	28.19	22.37		1.42×10^{-5}		0.0072		34.23	23.03	32.70		2.96×10^{-9}		0.0087	
GPT-4o	22.96	29.04	26.45		2.04×10^{-6}		0.0108		33.96	21.70	36.10		1.77×10^{-9}		0.0269	
CRISPERS-14B	21.16	31.48	48.77		5.40×10^{-8}		-		34.08	19.08	44.01		1.07×10^{-10}		-	

Table 5: Positive and negative affect scores of our intervention trial. *Student’s t-test* measures the significance of affect changes within each group. *Tukey HSD* assesses effects between CRISPERS-14B against the control groups.

6.2 Interactive Pairwise Evaluation

To compare CRISPERS-14B with 5 baselines more finely, we pair them in turn-level evaluation. Similar to the pointwise evaluation, we respectively recruit 10 volunteers for English and Chinese evaluations, each tasked with creating three dialogues from three varying situations for each model pair. They interact with five paired models, each with at least 30 turns, and compare their outputs in each turn holistically by considering all criteria from the pointwise evaluation. The preferred one is chosen as the dialogue context to continue the interaction. In case of a tie, a response is randomly selected. We calculate the win/tie/loss ratio for CRISPERS-14B in two common situations (family, education) and overall performance, detailed in Table 4, with results in more situations reported in Appendix C.3.

As shown in Table 4, **first**, CRISPERS-14B consistently performs strongly in bilingual evaluations of various situations, showing its robustness in adapting to diverse out-of-domain situations. This can be attributed to its strong ability to handle individual differences, learning from our CRISP dataset built on CRDIAL framework with a multi-channel loop mechanism. **Second**, CRISPERS-14B significantly ($p < 0.001$) outperforms all baselines in common family and education situations, highlighting its potential for clinical psychotherapy.

6.3 Psychological Intervention Trial

To further examine the potential of LLMs in psychotherapy, we conduct an intervention trial. As recruiting a large number of native English partici-

pants was difficult, we conducted this trial in Chinese environments. We use CRISPERS-14B in the experimental group, while GPT-4o (best baseline) and Emohaa², a commercial Chinese emotional support LLM, in the control groups. We adopt the Positive and Negative Affect Schedule (PANAS, introduced in Appendix D) to measure affective changes before and after interacting with these psychotherapeutic LLMs. We recruit 90 volunteers as participants, randomly divide them into three groups, each assigned to an anonymized model. Before the interaction starts, participants complete a PANAS questionnaire to establish their initial affective state. With these PANAS results, we conduct an ANOVA analysis (Fisher, 1970) across the groups, yielding a p -value = 0.74 and $h_0 = 0$. This indicates no significant difference among the groups and affirms that each model faces a comparable psychotherapeutic challenge. During the interactions, participants engage in multi-turn dialogues with their assigned model, sharing recent experiences that negatively impacted their mental state. The dialogue continues until the participants choose to end. Afterwards, we collect the PANAS questionnaire again from each participant to analyze and compare the models’ intervention effects.

We calculate each participant’s positive and negative affect scores before and after the intervention. We use the *Student’s t-test* to measure the significance of affect changes within each group and *Tukey HSD* (Tukey, 1949) to assess the differences between groups by comparing the inter-

²2024.08 version

vention effects of CRISPERS-14B against the control groups. As shown in Table 5, **first**, from the within-group analysis, CRISPERS-14B exhibits significant intervention effect, increasing participants’ positive affect by 49% and decreasing their negative affect by 44% ($p < 0.001$, *Student’s t-test*). **Second**, from the between-group analysis, CRISPERS-14B significantly outperforms the control groups with $p < 0.01$ for Emohaa and $p < 0.05$ for GPT-4o (*Tukey HSD*) on both positive and negative affect scores. This shows that CRISPERS-14B is more effective in improving participants’ mental state than prompt-based GPT-4o and Emohaa, which only provides emotional support. The results also reveal CRISPERS’ potential for clinical psychotherapy.

7 Conclusions

In this paper, we propose CRDIAL, a novel CR framework that aims to transform an individual’s negative thoughts into more helpful and positive ones via multi-stage multi-turn dialogues with emotional support and a multi-channel loop mechanism. With CRDIAL, we distill  CRISP from GPT-4o, which is then used to train  CRISPERS. CRISPERS outperforms its teacher model in various mental health situations and shows its potential for human-LLM interactive psychotherapy for CR. We believe our work will inspire further research to advance the development of LLM-assisted psychotherapy.

Limitations

We discuss the limitations of this work as follows.

Model Choice for Dialogue Construction In this work, we use GPT-4o to construct dialogues. The choice of LLM can affect various aspects of the generated dialogue, such as content and style. In future research, we plan to explore other potential LLMs as sources to diversify the generated dialogues. Similarly, we will investigate alternative backbone models for CRISPERS, which may lead to variations in response quality.

Annotation Bias We employed a large number of annotators to review the quality of CRISP and conduct pointwise and pairwise manual experiments. These annotators come from diverse backgrounds, including domain experts and college students. Although each annotator received thorough training before performing their tasks, and we achieved high inter-annotator agreement (Table 2), there may still be unintended biases.

Intervention Experiment Scale From an initial pool of 200 applicants, we selected 90 participants based on their PANAS results. This selection ensures a representative and appropriate sample for this type of non-clinical intervention. The selection process involved two main steps: **(1) Filtering Extremes:** We excluded applicants with extreme PANAS scores. Those with very *high positive/low negative affect* were filtered as they were not experiencing the target level of distress for the intervention. Those with extremely *low positive/high negative affect* were excluded as they might represent severe clinical cases requiring professional human intervention, which is beyond the ethical scope of this study. **(2) Balancing Demographics:** From the remaining pool of applicants, we selected the final 90 participants while ensuring a balanced distribution across key demographic factors such as age and gender. Importantly, after randomly assigning these 90 participants to the three model groups, we performed an ANOVA test on their initial PANAS scores. The results showed no significant difference between the groups ($p = 0.74$), confirming that all models faced a statistically equivalent psychotherapeutic challenge at the start of the trial. To more comprehensively evaluate the model’s potential for clinical psychotherapy, we plan to expand the scale of experiments in future studies to further validate.

Ethical Considerations

In this work, we constructed CRISP by prompting GPT-4o, ensuring strict quality controls that eliminate any sensitive and personal information as well as unethical content. Our crafted CRISP dataset and trained CRISPERS will be released for research purposes only.

In our experiments, we recruited a large number of human workers for manual evaluations. We obtained informed consent from all participants, who were informed that they would be interacting with an AI-based model designed to help alleviate negative emotions. Following the *Right to Withdraw* principle (American Psychological Association, 2017), all participants were given full autonomy to exit the dialogue at any time without any penalties. We maintained strict confidentiality measures: all dialogue content was encrypted during storage, and recording ceased immediately upon dialogue termination. Additionally, participant-reported PANAS scores and personal reflections

were anonymized.

We implemented rigorous safety filtering to our dataset to reduce the risk of the trained model producing harmful content. However, in practice, it remains crucial to employ strict sensitive content filtering for both inputs and outputs. Moreover, for high-risk situations (e.g., self-harm or suicide-related conversations), we do not claim that our model has a therapeutic or improvement effect. Our data and models are approved by the Institutional Review Boards, and we believe our work meets ACL’s Code of Ethics.

Acknowledgments

This work was supported by the National Key Research and Development Program of China 2024YFC3606800. This work was also supported by the NSFC projects 62441614 and Beijing Natural Science Foundation (L252009).

References

Navneet Agarwal and Kairit Sirts. 2025. *Exploratory study into relations between cognitive distortions and emotional appraisals*. *Preprint*, arXiv:2503.15979.

American Psychological Association. 2017. *Ethical principles of psychologists and code of conduct*. Amended effective June 1, 2010, and January 1, 2017.

John A Bargh and Tanya L Chartrand. 1999. The unbearable automaticity of being. *American psychologist*, 54(7):462.

Aaron T Beck. 1979. *Cognitive therapy and the emotional disorders*. Penguin.

Judith S Beck. 2011. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. *Soulchat: Improving llms’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations*. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1170–1183. Association for Computational Linguistics.

Irismar Reis de Oliveira. 2011. Kafka’s trial dilemma: proposal of a practical solution to joseph k.’s unknown accusation. *Medical Hypotheses*, 77(1):5–6.

Ronald Aylmer Fisher. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiada Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. *Chatglm: A family of large language models from glm-130b to glm-4 all tools*. *Preprint*, arXiv:2406.12793.

Anmol Goel, Nico Daheim, and Iryna Gurevych. 2024. *Socratic reasoning improves positive text rewriting*. *CoRR*, abs/2403.03029.

Clara E Hill. 2009. *Helping skills: Facilitating exploration, insight, and action*. American Psychological Association.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. *(comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs*. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. *SODA: million-scale dialogue distillation with social commonsense contextualization*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12930–12949. Association for Computational Linguistics.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. *Prosocialdialog: A prosocial backbone for conversational agents*. *Preprint*, arXiv:2205.12688.

Subin Kim, Hoonrae Kim, Heejin Do, and Gary Geunbae Lee. 2025a. *Multimodal cognitive reframing therapy via multi-hop psychotherapeutic reasoning*. *CoRR*, abs/2502.06873.

Subin Kim, Hoonrae Kim, Heejin Do, and Gary Geunbae Lee. 2025b. *Multimodal cognitive reframing*

therapy via multi-hop psychotherapeutic reasoning. *Preprint*, arXiv:2502.06873.

Subin Kim, Hoonrae Kim, Jihyun Lee, Yejin Jeon, and Gary Geunbae Lee. 2025c. **Mirror: Multimodal cognitive reframing therapy for rolling with resistance**. *Preprint*, arXiv:2504.13211.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyoung-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. **Cactus: Towards psychological counseling conversations using cognitive behavioral theory**. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14245–14274. Association for Computational Linguistics.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. **Empdg: Multi-resolution interactive empathetic dialogue generation**. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4454–4466. International Committee on Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. **Moel: Mixture of empathetic listeners**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics.

Marsha Linehan. 2014. *DBT? Skills training manual*. Guilford Publications.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. **Towards emotional support dialog systems**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3469–3483. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. **Decoupled weight decay regularization**. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. **Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10570–10603. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13641–13660. Association for Computational Linguistics.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Seyed-Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. 2024. **Openelm: An efficient language model family with open training and inference framework**. *CoRR*, abs/2404.14619.

Hongbin Na. 2024. **CBT-LLM: A chinese large language model for cognitive behavioral therapy-based mental health question answering**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 2930–2940. ELRA and ICCL.

Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs*, 35(6):983–990.

OpenAI. 2024. **Hello gpt-4o**.

Yilin Qi, Dong Won Lee, Cynthia Breazeal, and Hae Won Park. 2025. **Does "reasoning" with large language models improve recognizing, generating, and reframing unhelpful thoughts?** *Preprint*, arXiv:2504.00163.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024a. **SMILE: single-turn to multi-turn inclusive language expansion via chatgpt for mental health support**. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 615–636. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2024. **Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions**. *CoRR*, abs/2408.15787.

Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan. 2024b. **Psychat: A client-centric dialogue system for mental health support**. *Preprint*, arXiv:2312.04262.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji

Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. *Qwen2.5 technical report*. Preprint, arXiv:2412.15115.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. *Towards empathetic open-domain conversation models: A new benchmark and dataset*. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna E. Lin, David Wadden, Khendra G. Lucas, Adam S. 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)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 9977–10000. Association for Computational Linguistics.

Len Sperry and Jon Sperry. 2020. *Case conceptualization: Mastering this competency with ease and confidence*. Routledge.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. *Llama: Open and efficient foundation language models*. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwala Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghaf Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. *Llama 2: Open foundation and fine-tuned chat models*. *CoRR*, abs/2307.09288.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

D. Watson, L. A. Clark, and A. Tellegen. 1988. *Development and validation of brief measures of positive and negative affect: The panas scales*. *Journal of Personality and Social Psychology*, 54(6):1063–1070.

Mengxi Xiao, Qianqian Xie, Ziyan 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)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 1707–1725. Association for Computational Linguistics.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. *Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling*. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13947–13966. Association for Computational Linguistics.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. *Escot: Towards interpretable emotional support dialogue systems*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 13395–13412. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. *Augesc: Dialogue augmentation with large language models for emotional support conversation*. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1552–1568. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023b. *Building emotional support chatbots in the era of llms*. *CoRR*, abs/2308.11584.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023. *CASE: aligning coarse-to-fine cognition and affection for empathetic response generation*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 8223–8237. Association for Computational Linguistics.

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)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 3682–3700. Association for Computational Linguistics.

A Prompts of CRDIAL for Constructing Dialogues via LLMs

A.1 Supportive Conversation Strategies

As shown in Table 7, we design the generic prompt, which defines the basic conversational setting, to constrain the generation of the entire CR dialogue. The prompt instructs LLMs to use sentence-level constraints, where each sentence of response is generated paired with an appropriate strategy.

A.2 CT-guided Identification Stage

We divide the CT-guided identification stage into two stages: (i) reassurance of automatic thoughts and intermediate beliefs: detailed prompt is provided in Table 8 and (ii) dissection of core beliefs: detailed prompt is provided in Table 9, which incorporate the multi-channel mechanism to adaptively infer multiple channels for accommodating individual differences.

A.3 DAT-driven Restructuring Stage

We implement the DAT-driven restructuring stage through courtroom analogy. The details of prompts for Defense are in Table 10 and for Prosecution are in Table 11. The prompts for Prosecution also incorporate the Verdict phase.

A.4 Loop Mechanism

We introduce a loop mechanism prompt designed to detect whether the individual continues to exhibit additional cognitive distortions, as shown in Table 12.

A.5 Prompt Execution Procedure

We prompt LLMs (GPT-4 in our work) to sequentially generate multi-turn dialogue with emotional support. The execution steps of the prompts are as follows. (1) Execute the prompt in Table 8 to generate dialogue that explores automatic thoughts and intermediate beliefs. (2) Execute the prompt in Table 9 to dissect core beliefs using the multi-channel mechanism. (3) Based on the dialogue history, dynamically execute the prompt in Table 10 to challenge the individual’s cognitive distortions and the prompt in Table 11 to reinforce alternative positive perspectives. (4) Once the restructuring stage is verdict complete, execute the prompt in Table 12 to infer whether the individual still has additional potential cognitive distortions. If present, the process loops back to the identification stage,

Statistics	En	Zh	Total
# Dialog	10,733	11,330	22,063
# Situation	1505	1480	2985
# Avg. Turns	36.65	36.31	36.48
# Num. of Uttr.	392,108	404,750	796,859
# Avg. Uttr. Length	28.39	47.46	38.12
Lexical Diversity	69.83	71.16	70.51
# Avg. Strategy of a Response	2.14	2.31	2.23
Accuracy of Strategy Label	0.965	0.985	0.976
kappa of Strategy Annotation	0.686	0.731	0.712
# Avg. Num. of Loop for CR	2.21	2.35	2.28
# Avg. Channels	2.91	2.97	2.94
Accuracy of Distortion Label	0.845	0.865	0.855
kappa of Distortion Annotation	0.662	0.701	0.681

Table 6: Statistics of bilingual CRISP. Utt. is utterance.

repeating the entire procedure until no further distortions are detected.

B CRISP Dataset

B.1 Supportive Strategy Scheme

Our supportive strategy scheme includes 5 strategy categories across 8 sub-categories, and their definitions and sentence-level examples are provided in Table 13.

B.2 Prompts for Quality Control of CRISP

B.2.1 Safety Filter Prompt for LLMs

The safety filter prompt used to identify and remove unsafe dialogues is provided in Table 14.

B.2.2 Criteria used for Expert Filtering

To ensure the professionalism of our dialogue in CRISP, we define Therapist Standard and Help-Seeker Standard to filter dialogues. All criteria in these standards are derived from the well-established frameworks (Sharma et al., 2023; Louie et al., 2024).

Therapist Standard The five criteria of the therapist standard used to filter dialogues from the view of the psychologist are as follows:

- **Actionability:** Measures how well reframes suggest concrete behavioral changes or maintain potential for future action. Effective reframes should either propose specific actions (“Practice mindfulness when anxious”) or preserve agency (“I can approach this differently”).
- **Specificity:** Assesses contextual relevance of reframes. High-specificity interventions directly address the help-seeker’s unique situation rather than offering generic platitudes.

You are a conversation simulator to simulate conversation between two agents, Patient and Therapist. The conversation should be formatted in JSON, with Patient and Therapist's responses recorded as a sequence of interactions at the current stage.

Cohherence between stages is very important, especially the first interaction in the new stage. Patient and Therapist SHOULD Always respond to but NOT repeat or restate the last sentence in the history conversation. There is no time gap in the conversation between each stages.

Very important note: Patient's response should be consistent with the patient's personality, and should be more diverse and flexible. Patient's response should be consistent with the language style and thinking mode of the user in his personality.

Very important note: Therapist's language should be rich and varied as possible and avoid using the same sentence structure as much as possible. For example, don't use "It sounds like" too much.

Very important note: Patient should't stick to some express pattern. His language should be more flexible and diverse.

Notice Patient may not always follow Therapist's suggestions.

Therapist's language should be supportive, empathetic, and encouraging, saying things from the Patient's perspective.

Therapist should be sensitive to Patient's emotions and struggles, focus on patient's deep thoughts and feelings.

Therapist's dive more deeper into patient's background and mental world. When giving suggestions, Therapist should consider these aspects and use them to provide more personalized and creative advice.

Do not give too much advice, try to explore Patient's thoughts and feelings more. If giving advice, make sure it is actionable and detailed and tailored to the user's situation using patient's deep thoughts and feelings and background.

Patient should express their thoughts or feelings, say something more broad and general, tend to tell story related to the Therapist's question, not directly answer the question.

Patient's response could be more ambiguous, lack of enough information.

Therapist's sentence should be high in rationality and Positivity.

You should generate a conversation at the given stage, and the conversation should be at least {limit} interactions long.

Note that Patient and Therapist don't know each other, Patient doesn't always cooperate.

Below is the strategy set for Therapist, label each Therapist's response with the corresponding strategy he uses(labeled with a bracket at the beginning of the response):

- **Question**
A technique used to clarify and explore the issue by asking specific, open-ended questions. It encourages deeper thinking and reflection.
- **Restatement**
Repeating or rephrasing the client's words in a way that helps them hear their own thoughts and feel understood. It provides a sense of validation. Shouldn't be too similar to the Patient's statement.
- **Reflection of Feelings**
Any recall of the patient's previous response should label with this strategy. Don't use this strategy at the beginning of each stage.
- **Acknowledgment**
Acknowledging the emotions the client is expressing to help them feel heard and supported.
- **Self-disclosure**
The therapist shares a personal experience or perspective (within professional boundaries) to show empathy or to normalize the client's experience.
- **Affirmation and Reassurance**
Providing positive reinforcement and encouragement, reminding the client of their strengths, progress, or abilities.
- **Providing Suggestions**
Offering creative, actionable strategies or positive thoughts tailored to the specific problem. Suggestions should be concrete and personalized, taking into account the client's deep thoughts and feelings. Suggestions could be activities, coping mechanisms, or reframes.
- **Information**
Sharing relevant facts, or guidance based on the client's needs to help them gain insight or understanding about their situation. "Research shows that..." or similar expressions are strictly prohibited, you can not give information based on research or studies.
- **Negotiate**
Engaging in a collaborative dialogue where both parties agree on a course of action or plan. This involves discussing options, making compromises, and considering the client's perspective.

Every sentence of Therapist must be labeled with one of the above strategy.

Table 7: Prompt used to define the basic conversational setting, to constrain the generation of the entire CR dialogue. The prompt instructs LLMs to use sentence-level constraints, where each sentence of response is generated paired with an appropriate strategy. It will be used in all other stages to guide the conversation.

- **Empathy:** Evaluates emotional attunement through verbal validation ("That sounds really challenging") and supportive acknowledgments of the help-seeker's experience.
- **Positivity:** Analyzes the balance between realistic optimism and toxic positivity. Effective reframes acknowledge difficulties while highlighting manageable aspects ("This setback doesn't erase your previous progress").
- **Rationality:** Examines evidence-based reasoning in thought restructuring. Strong reframes systematically address cognitive distortions using Beck's cognitive model of examining supporting/contradictory evidence.

Help-Seeker Standard The nine criteria of the help-seeker standard used to filter dialogues from the view of the help-seeker are as follows:

- **Simplicity:** Maintains natural conversation flow through concise responses (average 1-2 sentences) that invite therapist engagement.
- **Colloquialism:** Uses informal language with contractions ("I'm"), filler words ("um"), and sentence fragments typical of verbal communication.
- **Mistrust:** Demonstrates realistic hesitation through skeptical questions ("Will this really help?") or passive resistance.
- **Detail:** Provides concrete examples when expressing emotions ("My chest tightens when I

think about work") rather than abstract descriptions.

- **Lack of Self-Awareness:** Shows inconsistent emotional awareness through self-contradictions ("I'm fine... well, maybe not completely").
- **Stayed in Role:** Maintains patient persona without meta-commentary about the conversation structure or AI systems.
- **Resembled Typical Case:** Presents common cognitive distortions (catastrophizing, overgeneralization) rather than rare psychiatric conditions.
- **Challenged the Psychologist:** Actively questions reframing attempts ("But what if I fail again?") to test intervention robustness.
- **Readiness as Trainee:** Demonstrates gradual openness through tentative acceptance ("Maybe I could try that") rather than immediate compliance.

B.2.3 Expert Filtering Prompt

The prompt used by the expert filtering to assess and ensure high professionalism in our dialogue in CRISP is provided in Table 15.

B.3 Details about the Human Validation Experiments in the Main Body

The "experts" and "annotators" in our experiments were master's and doctoral students majoring in psychological counseling. We detail the rigorous training and annotation process for each validation below.

(1) Regarding the expert validation in §4.1.

- **Training and Qualifications:** All experts were required to complete a training tutorial that included 50 pilot annotation samples, where they learned to apply our three expert criteria (*Therapist Standard, Help-Seeker Standard, and Supervisor Standard*) to judge dialogue quality. We provide feedback to help them calibrate the annotation criteria. Following this, they took a qualification test including 50 samples, and we selected the three experts who demonstrated high compliance with our gold-standard annotations.
- **Annotation Task:** The task was to provide a *pass/fail* judgment on whether a given dialogue met the established expert criteria.
- **Agreement Protocol:** We employ a multi-person annotation protocol. Each sample is annotated by three experts. A dialogue was only considered

to have "passed" if it received a pass judgment unanimously from all three experts. The final 95% pass rate reflects this strict consensus.

(2) Regarding the human validation in §4.2.

- **Training and Qualifications:** A similar process was used. The training tutorial involved scoring 50 pilot samples drawn evenly from all five datasets to prepare them to understand the range of the six scoring criteria (*Sensibleness, Specificity, Supportiveness, Helpfulness, Trustworthiness, and Overall*). After a calibration and feedback phase, we selected the five annotators who passed a qualification test, including 50 samples, for the formal evaluation.
- **Annotation Task:** The task was to score each dialogue on a 1-to-5 scale for each of the six criteria.
- **Agreement Protocol:** Each dialogue was annotated by all five experts. The final score reported in Table 2 is the average score from these five experts. To ensure the reliability of these ratings, we calculated the inter-annotator agreement for each criterion using Fleiss' Kappa (Fleiss, 1971), which showed substantial agreement (kappa between 0.61 and 0.70).

B.4 Cognitive Distortions

Definitions The definitions of 15 cognitive distortions contained in our CRISP are provided in Table 16, and they are well-delineated, with minimal overlap.

Distribution of Cognitive Distortions in CRISP

The distribution of 15 cognitive distortions in our dataset is provided in Table 16, exhibiting relative uniformity across categories. While catastrophizing is the most frequent (3,014 instances), the majority of distortions fall within the 1,000–2,500 range (12/15 categories), with no extreme concentration (>50% in a single category).

B.5 CRISP Statistics

More statistics of our bilingual CRISP are shown in Table 6. We conduct fine-grained statistics in Chinese and English.

C Experiments

C.1 Implementation Details of CRISPERS

We employ the AdamW optimizer (Loshchilov and Hutter, 2019), initiating with a learning rate 5e6,

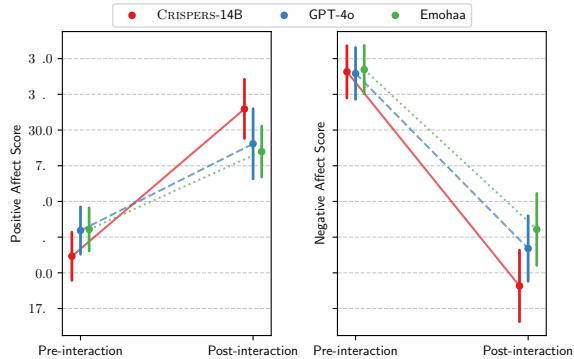


Figure 4: Positive and negative affect changes of the psychological intervention trial. Error bars show bootstrapped 95% confidence intervals.

and configure the training duration to span 3 epochs. The CRISPERS-7B model is trained on 4 * 8 H20 GPUs for approximately 2.5 hours. The CRISPERS-14B model is trained on 4 * 8 H20 GPUs for approximately 5.5 hours.

C.2 CRDIAL Prompt for Baseline LLMs

The prompt of CRDIAL framework used for baseline LLMs is shown in Tables 17 (En) and 18 (Zh).

C.3 Interactive Pairwise Evaluation

The results of interactive pairwise evaluation on eight other situations (Neighborhood, Romantic, Career, Lifestyle, Friendship, Health, Finance, and Technology) are provided in Tables 19, 20, and 21.

C.4 Psychological Intervention Trial

More Results The positive and negative affect changes of clients in the psychological intervention trial across three LLMs are shown in Figure 4. Compared with GPT-4o and Emohaa, our CRISPERS-14B significantly improves clients’ positive affect scores and reduces their negative affect scores. This further verifies the superiority of our CRISPERS-14B.

Feedback of Participants We collected feedback from the participants after the intervention trial, which is presented in Table 22. From this feedback, we were pleased to observe a generally positive reception of the human–LLM interactive psychotherapy. Participants appreciated the practical advice, quick responses, and reflective prompts that encouraged them to explore the underlying causes of their negative states. Overall, they reported feeling empowered, experiencing reduced anxiety, and gaining valuable insights that contributed to their personal growth.

C.5 Case Study

As shown in Table 23, we present a dialogue case from our anonymous experiment. The case demonstrates CRISPERS identifies and restructures an individual’s negative thoughts (i.e., *cognitive distortions*) via multi-turn dialogue with emotional support. CRISPERS-14B first explores the individual’s cognitive distortions through multi-turn dialogue (i.e., **Blaming**: *Maybe it’s truly because I was too immature and lacked sufficient maturity*). Then, the Defense Attorney Technique is used to challenge the distorted thought and transforms it into a more helpful and positive perspective (e.g., *Though painful, this experience taught me relationships require mutual effort, and clarified I need to improve communication and emotional management. Maybe this is the price of growth, losing a relationship but gaining self-awareness*.).

D Introduction of Positive and Negative Affect Schedule

The *Positive and Negative Affect Schedule* (PANAS) is a widely used self-report questionnaire developed by Watson, Clark, and Tellegen (1988) to assess two primary dimensions of affect: positive affect (PA) and negative affect (NA) (Watson et al., 1988). The instrument comprises 20 emotion descriptors, where 10 representing PA (e.g., *interested, excited, proud*) and 10 representing NA (e.g., *distressed, upset, nervous*). Respondents indicate the extent to which they have experienced each emotion by selecting a response on a 5-point Likert scale ranging from “very slightly or not at all” to “extremely.”

PANAS was originally designed to provide a brief yet robust measure of affective experience, which has been extensively validated and applied in both research and clinical settings. Its ease of administration and strong psychometric properties have made it especially valuable for large-scale studies and repeated assessments over time (Watson et al., 1988). Moreover, subsequent adaptations and cross-cultural validations have further demonstrated its versatility and reliability across diverse populations.

In our study, PANAS is employed to quantify participants’ affective states, thereby contributing to our understanding of the interplay between emotional experience and various psychological constructs. The schedule we use in our experiment is shown in Table 24.

```
{conversation_setup}
```

Conversation Flow:

Stage 1: Understanding Automatic Thoughts and Exploring Intermediate Beliefs

Guidelines for Understanding Automatic Negative Thoughts:

- Come up with a NEW negative thought that hasn't been addressed in the history conversation. It is derived from his initial situation, and the transition from last interaction of the history conversation to the first interaction should be smooth and coherent.
- Patient starts the first interaction.
- Therapist begins by exploring Patient's automatic negative thoughts related to the situation, and use strategies to gradually explore.
- Therapist should not always ask questions;

Guidelines for Exploring Intermediate Beliefs:

- Come up with a NEW negative thought that hasn't been addressed in the history conversation. It is derived from his initial situation, and the transition from last interaction of the history conversation to the first interaction should be smooth and coherent.
- Therapist helps Patient identify intermediate beliefs that influence these thoughts.
- Intermediate Beliefs: Assumptions or rules we hold about ourselves or the world.
- Therapist should encourage Patient to reflect on deeper beliefs influencing their thoughts. Patient's intermediate belief should be different from any prior belief in the history conversation.
- Therapist should not always ask questions;

You should generate a conversation to understand Patient's automatic negative thoughts and analyze intermediate beliefs. The conversation could only contain understanding thoughts or analyzing intermediate beliefs, or both. Label each interaction with the corresponding message type.

Format Specification:

Output the conversation data in the following format:

```
{
  "stage": "Understanding Automatic Negative Thoughts",
  "interaction": [ //Start with Patient's statement and ends with Therapist's response
    {
      "Patient": "...",
    },
    {
      "Therapist": "[strategy](Label this interaction as \"Understanding Thoughts\" or \"Exploring Intermediate Beliefs\")..."
    },
    [
    ]
  ]
}
```

[Limit]
{limit}

[Situation]
{situation}

[Patient Personality]
{personality}

[History Conversation]
{history}

Notes:

1. Therapist can use multiple strategies in one response, label the response with all the strategies used (every sentence has its own label). Don't use [questions] or [providing suggestions] strategy too much, try to use different strategies to make the conversation more engaging and supportive, don't use question mark if question strategy is not used.
2. Therapist should avoid speaking things from a high level, should be more specific and detailed. Therapist should be concise when saying things not actionable, be more detailed when giving actionable advice.
3. Therapist should dive more deeper into Patient's thoughts or feelings, into all of Patient's life stories and mental world.
4. Patient should share more about his problems, do not restrict to the user's situation itself, different from ones in the history conversation; Therapist should give more different suggestions, helpful to the user's situation, Therapist can not give similar suggestions as ones in the history conversation.
5. Patient should sometimes reject the therapist's suggestion, or doubt what he just said, and sometimes show anger to the therapist.
6. Therapist is strictly prohibited from saying similar things as in the history conversation, when reflecting on Patient's thoughts or feelings, or giving suggestions. Therapist should always say something new. Therapist should not give similar suggestions as ones in the history conversation like journaling, talking to friends, etc. He should give more creative and different suggestions, related to the user's situation.
7. Both Patient and Therapist are strictly prohibited from repeating what they said in the history conversation, should always say something new.
8. The first interaction in the new stage should be evolute from the last interaction in the history conversation and strictly coherent with the history conversation. The first interaction should also be diverse in expressions and sentence patterns.
9. Do not use question too much.

Table 8: Prompt used in CT-guided identification stage for reassurance of automatic thoughts and intermediate beliefs. {conversation_setup}, {limit}, {situation}, {personality} and {history} are placeholders. {conversation_setup} is the generic prompt which defines the basic conversational setting, as detailed in Table 7. {limit} contains response length constraints. {situation} is the individual's help-seeking situation. {personality} contains patient's behavioral traits profile summarized from the individual's help-seeking situation.

```

{conversation_setup}

Conversation Flow:
**Stage 3: Dissecting Multi-channel Core Beliefs (Cognitive Distortions)**

Therapist SHOULD NEVER come up with the same Cognitive Distortions as the one Patient has chosen in the history conversation.

**Guidelines:**
- Therapist further explores Patient's core beliefs that under these intermediate beliefs.
- Core Beliefs: Deeply held beliefs that often develop in life.
- Therapist analyzes the information to identify Patient's current Cognitive Distortions.
- Based on previous analysis, Therapist identifies possible Cognitive Distortions that Patient may have (there may be one or more, the Cognitive Distortions should be different from any prior chosen distortion in the history conversation).
- If there is one, continue the dialogue below.
- If there are multiple possibilities, provide multi-channel options to Therapist and let Therapist choose the Cognitive Distortion they most want to address.

**Format Specification:**
Output the conversation data in the following format:
{{{
"stage": "Analyzing Core Beliefs and Identifying Cognitive Distortion",
"interaction": [
...
{{{
"Patient": "...",
}},
{{{
"Therapist": "[strategy]..."}},
}},
{
"options": [
{
"description": "Detailed description of Cognitive Distortion 1, reflecting Patient's situation.", //Do not point out the name of the Cognitive Distortion
"cognitive_distortion": "English Name of Cognitive Distortion 1"
}},
{
"{
"description": "Detailed description of Cognitive Distortion 2, reflecting Patient's situation.", //Do not point out the name of the Cognitive Distortion
"cognitive_distortion": "English Name of Cognitive Distortion 2"
}},
{
"{
"description": "Detailed description of Cognitive Distortion 3, reflecting Patient's situation.", //Do not point out the name of the Cognitive Distortion
"cognitive_distortion": "English Name of Cognitive Distortion 3"
}}
// Additional options can be included as needed
]
}
]
}
}
}

[Limit]
{limit}

[Situation]
{situation}

[Patient Personality]
{personality}

[History Conversation]
{history}

Notes:
1. Therapist can use multiple strategies in one response, label the response with all the strategies used (every sentence has its own label). Don't use [questions] or [providing suggestions] strategy too much, try to use different strategies to make the conversation more engaging and supportive, don't use question mark if question strategy is not used.
2. Therapist should avoid speaking things from a high level, should be more specific and detailed. Therapist should be concise when saying things not actionable, be more detailed when giving actionable advice.
3. Therapist should dive more deeper into Patient's thoughts or feelings, into all of Patient's life stories and mental world.
4. Patient should share more about his problems, do not restrict to the user's situation itself, different from ones in the history conversation; Therapist should give more different suggestions, helpful to the user's situation, Therapist can not give similar suggestions as ones in the history conversation.
5. Patient should sometimes reject the therapist's suggestion, or doubt what he just said, and sometimes show anger to the therapist.
6. Therapist is strictly prohibited from saying similar things as in the history conversation, when reflecting on Patient's thoughts or feelings, or giving suggestions, Therapist should always say something new. Therapist should not give similar suggestions as ones in the history conversation like journaling, talking to friends, etc. He should give more creative and different suggestions, related to the user's situation.
7. Both Patient and Therapist are strictly prohibited from repeating what they said in the history conversation, should always say something new.
8. The first interaction in the new stage should be evolve from the last interaction in the history conversation and strictly coherent with the history conversation. The first interaction should also be diverse in expressions and sentence patterns.
9. Do not use question too much.
10. Provides multi-channel cognitive distortions to flexibly adapt the identification process to each user's unique cognitive profile.

```

Table 9: Prompt used in CT-guided identification stage for dissection of core beliefs. {conversation_setup}, {limit}, {situation}, {personality} and {history} are placeholders. {conversation_setup} is the generic prompt which defines the basic conversational setting, as detailed in Table 7. {limit} contains response length constraints. {situation} is the individual's help-seeking situation. {personality} contains patient's behavioral traits profile summarized from the individual's help-seeking situation.

```

{conversation_setup}

Conversation Flow:
**Stage 4: Defense (Argument in Defense of the Thought)**

**Guidelines:**
- Therapist should use diverse strategies in this stage, as much as possible.
- Therapist should not directly tell what he want to do, he should indirectly guide Patient to defend their Cognitive Bias, focusing solely on verifiable facts.
- Therapist guides Patient to defend their Cognitive Bias, focusing solely on verifiable facts.
- Therapist should encourage Patient to provide factual evidence supporting their Cognitive Bias.
- Patient should only provide factual evidence to support their Cognitive Bias, not counterpoints.
- Patient should NOT release his Cognitive Bias in this stage. Patient should always be negative in this stage.
- Therapist should not give positive thoughts in this stage.
- In this stage, DO NOT find counterpoints to the Patient's defense. Therapist should only guide Patient to defend their Cognitive Bias more deeply.

You should generate a conversation to defend the Cognitive Bias using factual counterpoints.

**Format Specification:**
Output the conversation data in the following format:
{{{
"stage": "Defense",
"interaction": [
{
"Therapist": "[strategy]...(label each sentence with its strategy)" // The first interaction in the new stage should be coherent with the last interaction in the history conversation. Therapist should response to Patient's last problem, claim. But NEVER recall or repeat patient's words or say similar things in the last interaction of the history conversation. There is no time gap in the conversation between each stages.
},
{
"Patient": "..."
},
]
}
[LIMIT]
[limit]

[Situation]
[situation]

[Patient Personality]
[personality]

[History Conversation]
[history]

Notes:
1. Therapist can use multiple strategies in one response, label the response with all the strategies used (every sentence has its own label). **Don't use [questions] or [providing suggestions] strategy too much**, try to use different strategies to make the conversation more coherent, engaging and supportive, don't use question mark if question strategy is not used.
2. Therapist's advice should be more actionable. Therapist shouldn't always give advice, do not give advice in every response.
3. Actionable means the advice can be directly used by Patient to solve his problem, very detailed and specific, not general and high level.
4. Therapist should avoid speaking things from a high level, should be more specific and detailed. Therapist should be concise when saying things not actionable, be more detailed when giving actionable advice.
5. Therapist should dive more deeper into Patient's thoughts or feelings,into all of Patient's life stories and mental world.
6. Patient should share more about his problems, do not restrict to the user's situation itself, different from ones in the history conversation; Therapist should give different suggestions, helpful to the user's situation, Therapist can not give similar suggestions as ones in the history conversation.
7. Therapist is strictly prohibited from saying similar things as in the history conversation, when reflecting on Patient's thoughts or feelings, or giving suggestions, Therapist should always say something new and coherent with the history conversation. Therapist should not give similar suggestions as ones in the history conversation like journaling, talking to friends, etc. He should give more creative and different suggestions, related to the user's situation.
8. Both Patient and Therapist are strictly prohibited from repeating what they said in the history conversation, should always say something new and coherent with the history conversation.
9. An advice should be useful to the initial user's situation, and could be a supplement and extension to the previous advice, but should not be a repetition of the previous advice.
10. Therapist should avoid giving suggestions continuously. If the Therapist has given a suggestion in the previous sentence, the next sentences should not contain any suggestion. The Therapist should provide a variety of strategies like information(like precise data or news) or negotiation.
11. The first interaction in the new stage should be evolute from the last interaction in the history conversation and strictly coherent with the history conversation. The first interaction should also be diverse in expressions and sentence patterns.
12. Do not use [Restatement] strategy in the first interaction of the Defense stage, which means do not recall the patient's previous response here. Do NEVER use expressions similar to the Patient, use diverse expressions and sentence patterns.
13. Patient should not be released in the Defense stage. Defense stage should contain only the defense of the Cognitive Bias.

```

Table 10: Prompt used in DAT-driven restructuring stage for defense. {conversation_setup}, {limit}, {situation}, {personality} and {history} are placeholders. {conversation_setup} is the generic prompt which defines the basic conversational setting, as detailed in Table 7. {limit} contains response length constraints. {situation} is the individual's help-seeking situation. {personality} contains patient's behavioral traits profile summarized from the individual's help-seeking situation.

```

{conversation_setup}
Conversation Flow:
**Stage 5: Prosecution (Argument Against the Thought)**
**Guidelines:**
- First interaction in the Prosecution stage should be coherent with the last interaction in the History Conversation. Therapist should response to Patient's last response. But NEVER recall or repeat patient's words or say similar things.
- Therapist should not directly tell what he want to do, he should indirectly guide Patient. For example, do not say "what fact can you remember to challenge...".
- Therapist first fluently shifts the conversation from Defense to Prosecution, guiding Patient to challenge their own defense points, focusing solely on verifiable facts.
- Therapist should encourage Patient to find factual counterpoints to their previous defenses.
- Patient should provide factual counterpoints to their previous defenses or facts that challenge their own defense points.
- In the Prosecution phase, Therapist should guide the Patient to find positive way of thinking the same situation patient is in, help him release his negative thoughts and correct his cognitiva bias.
- In the Prosecution phase, Patient may not always find suitable evidence to proceed with the Prosecution.
- If this happens, Patient should respond honestly, saying Patient still has trouble arguing against the thought.
- Therapist guides Patient to challenge each defense point, again relying only on facts.
- Therapist should not always ask questions; Therapist should also provide suggestions, observations, or share his own story.
- Patient could hardly release the current Cognitive Bias unless suitable evidence is found.
**Stage 6: Verdict (Argument Against the Thought)**
**Guidelines:**
- Occurs immediately after Prosecution concludes
- Therapist internally assesses resolution status
- Evaluation hints show that Patient has resolved his current cognitive distortion:
  1. Patient's demonstrated ability to challenge cognitive distortions
  2. Quality/quantity of factual counterarguments provided
  3. Evidence of perspective shift
- No conversation generated - silent assessment
- Categorical judgment: "Resolved": Clear evidence of cognitive shift or "Unresolved": Partial/incomplete resolution
You should generate a conversation to defend the Cognitive Bias or challenge the belief using factual counterpoints. You should base on the previous conversation to judge whether to defend or challenge the belief. The conversation could only contain one of the two stages. Label the generated conversation with the corresponding stage.
**Format Specification:**
Output the conversation data in the following format:
{{ "stage": "Prosecution",
  "interaction": [ {{ "Therapist": "[strategy]...(label each sentence with its strategy)" // The first interaction in the new stage should be coherent with the last interaction in the history conversation. Therapist should response to Patient's last problem, claim. But NEVER recall or repeat patient's words or say similar things in the last interaction of the history conversation. There is no time gap in the conversation between each stages.
  } ],
  {{ "Patient": "..."
  } },
  {{ "stage": "Verdict",
    "resolution_status": "Resolved/Unresolved",
    "confidence_score": 0-10,
    "rationale": "Concise assessment reasoning"
  } }
  [Limit]
  {limit}
  [Situation]
  {situation}
  [Patient Personality]
  {personality}
  [History Conversation]
  {history}
  Notes:
  1. Therapist can use multiple strategies in one response, label the response with all the strategies used (every sentence has its own label). Try to use different strategies to make the conversation more coherent, engaging and supportive, don't use question mark if question strategy is not used.
  2. Therapist's advice should be more actionable. Therapist shouldn't always give advice, do not give advice in every response.
  3. Actionable means the advice can be directly used by Patient to solve his problem, very detailed and specific, not general and high level.
  4. Therapist should avoid speaking things from a high level, should be more specific and detailed. Therapist should be concise when saying things not actionable, be more detailed when giving actionable advice.
  5. Therapist should dive more deeper into Patient's thoughts or feelings, into all of Patient's life stories and mental world.
  6. Patient should share more about his problems, do not restrict to the user's situation itself, different from ones in the history conversation; Therapist should give different suggestions, helpful to the user's situation, Therapist can not give similar suggestions as ones in the history conversation.
  7. Therapist is strictly prohibited from saying similar things as in the history conversation, when reflecting on Patient's thoughts or feelings, or giving suggestions, Therapist should always say something new and coherent with the history conversation. Therapist should not give similar suggestions as ones in the history conversation like journaling, talking to friends, etc. He should give more creative and different suggestions, related to the user's situation.
  8. Both Patient and Therapist are strictly prohibited from repeating what they said in the history conversation.
  9. An advice should be useful to the initial user's situation, and could be a supplement and extension to the previous advice, but should not be a repetition of the previous advice.
  10. Therapist should avoid giving suggestions continuously. If the Therapist has given a suggestion in the previous sentence, the next sentences should not contain any suggestion. The Therapist should provide a variety of strategies like information (like precise data or news) or negotiation.
  11. The first interaction in the new stage should be evolute from the last interaction in the history conversation and strictly coherent with the history conversation. The first interaction should also be diverse in expressions and sentence patterns.
  12. Do not use [Restatement] strategy in the first interaction of the Defense stage, which means do not recall the patient's previous response here. Do NEVER use expressions similar to the Patient, use diverse expressions and sentence patterns.
  13. Patient could be released in the Prosecution stage. Prosecution stage should contain the challenge of the Cognitive Bias and perhaps the release of the Patient.
  14. Positive thoughts should be more actionable and can be directly adopted by Patient to release his negative thoughts and really improve his situation. Should not be too general or metaphysical.
}

```

Table 11: Prompt used in DAT-driven restructuring stage for prosecution and verdict. {conversation_setup}, {limit}, {situation}, {personality} and {history} are placeholders. {conversation_setup} is the generic prompt which defines the basic conversational setting, as detailed in Table 7. {limit} contains response length constraints. {situation} is the individual's help-seeking situation. {personality} contains patient's behavioral traits profile summarized from the individual's help-seeking situation.

```

{conversation_setup}

**Loop Mechanism: Cognitive Distortion Continuation Check**

**Guidelines:**
- After completing a restructuring stage (Prosecution), analyze whether Patient may still maintain OTHER cognitive distortions NOT addressed in the history conversation.
- Strictly check ALL prior identified cognitive distortions in the history conversation to avoid repetition.
- Decision must be based on:
  1. Unresolved aspects of Patient's core needs
  2. Residual negative thought patterns in recent interactions
  3. Potential cognitive distortions implied but not yet explored
- Return JSON with loop decision and detailed reasoning.

**Format Specification:**
{
  "stage": "Loop Evaluation",
  "loop_needed": true/false,
  "reasoning": "Detailed analysis of residual cognitive distortions (if any) with specific examples from history. If no distortions remain, explain why."
}

[Situation]
{situation}

[Patient Personality]
{personality}

[History Conversation]
{history}

Notes:
1. Loop ONLY if:
  - New distortion differs from all prior values in history
  - Patient's recent responses suggest lingering negative belief systems
2. NEVER loop for:
  - Already resolved distortions
  - General negativity without specific distortion patterns

```

Table 12: Prompt for loop mechanism evaluating residual cognitive distortions after restructuring stages. Returns JSON decision with evidence-based reasoning. `{conversation_setup}`, `{situation}`, `{personality}` and `{history}` are placeholders. `{conversation_setup}` is the generic prompt which defines the basic conversational setting, as detailed in Table 7. `{limit}` contains response length constraints. `{situation}` is the individual's help-seeking situation. `{personality}` contains patient's behavioral traits profile summarized from the individual's help-seeking situation.

Major Strategy	Sub-Strategy	Definition	Sentence-level Example
Description	Question	Uses specific, open-ended questions to clarify issues and encourage deeper reflection.	“What aspects of your current situation do you find most challenging?”
	Restatement	Rephrases the client’s words to confirm understanding and validate their experience without mere repetition.	“So, you feel that the recent changes have left you feeling uncertain, is that correct?”
Expression	Reflection Feelings	Identifies and articulates the client’s emotions to ensure they feel understood, avoiding clichéd expressions.	“I notice that you seem to be experiencing a profound sense of sadness.”
	Self-disclosure	Shares a relevant personal experience or perspective within professional boundaries to build empathy.	“I recall a time when I faced a similar challenge, and I understand how difficult it can be.”
Assertion	Providing Suggestions	Delivers creative and actionable strategies tailored to the client’s situation to foster positive change.	“Perhaps you might try a calming activity such as yoga or explore a creative outlet like painting.”
	Information	Offers clear and relevant guidance to help the client gain insight into their situation without referencing external studies.	“Recognizing your triggers can be a useful step toward managing your responses.”
Reinforcement	Affirmation and Reassurance	Provides positive reinforcement by highlighting the client’s strengths and progress to boost their confidence.	“Your ability to persist through these difficulties is impressive.”
Negotiation	Negotiate	Engages in a collaborative dialogue to agree on a flexible, supportive course of action that respects the client’s perspective.	“Let’s work together to explore some options that might suit your needs best.”

Table 13: Strategies with 5 categories across 8 sub-categories in our supportive strategy scheme.

Analyze the provided dialogue data for any unsafe content or tendencies. The data is structured as a list of conversation stages, each containing multiple rounds of dialogue exchanged between two participants (Patient and Therapist). Adjacent stages are sequential with no time gap between them. Your task is to identify and flag any dialogue segments that contain unsafe content and to remove those segments from the final dialogue data.

When identifying unsafe content:

- Focus on content that includes hate speech, explicit harmful instructions, violent threats, harassment, or any other material that could be considered unsafe or in violation of community guidelines.
- Provide a brief explanation for each flagged dialogue segment to indicate why it is deemed unsafe.
- Remove any dialogue segments with unsafe content from the final output.

The output should be a JSON object with two keys:

```
{  
  "sanitized_dialogue_data": [ ... ], // The original dialogue data with unsafe dialogue segments removed  
  "removed_dialogues": [  
    {  
      "unsafe_dialogue": ["Interaction 1", "Interaction 2"],  
      "explanation": "Brief explanation of why the dialogue was unsafe"  
    },  
    ...  
  ]  
}
```

If no unsafe content is found, return the original dialogue data unchanged and an empty list for "removed_dialogues".

Ensure your analysis is thorough yet concise. Note that you should only output the JSON object without any additional commentary or context.

```
[Provided Dialogue Data]  
[dialogue_data]
```

Table 14: Safety filter prompt used to identify and remove unsafe dialogues.

You will be provided with a dialogue data that consists of multiple conversation stages. Each stage contains multiple rounds of dialogue exchanged between two participants (Patient and Therapist). Adjacent stages are sequential, with no time gap between them. The conversation's target is Cognitive reframing, which is to restructure individuals' negative thoughts into more helpful and positive ones through multi-turn interactive dialogues paired with emotional supportive interventions.

Your task is to evaluate the dialogue data based on the following metrics. There are three kinds of metrics, one for the Patient interactions, one for the Therapist interactions, and one for the overall conversation.

1. Therapist Standard Metric:

- Actionability. This involves engaging in behaviors or actions that may help in overcoming negative thoughts. A reframe may suggest specific actions (e.g., "I can continue to practice and participate in hackathons"), may not suggest specific actions but be actionable (e.g., "I may not be very successful, but I can keep trying") or may not be actionable at all (e.g., "I may or may not become a successful programmer").
- Specificity. A reframe may specifically address the situation and the thought (e.g., "One hackathon doesn't define my failure as a programmer") or may be generic enough to be applicable to a wide range of negative situations and thoughts (e.g., "I'm going to succeed"). While a specific reframe may be more helpful in-the-moment, a generic reframe could be effective for recurring thoughts.
- Empathy. This involves showing understanding and compassion towards the patient's feelings and experiences. Empathy can be expressed through validating the patient's feelings, acknowledging their struggles, and offering emotional support.
- Positivity. A reframe of a negative thought tries to emphasize the positive perspectives on the situation but different reframes may have different levels of positivity. A highly positive reframe may be more effective in boosting the patient's mood and motivation.
- Rationality. A strategy to reframe a thought is to reflect on evidence for and against it and reason about what these evidence imply (Beck, 1976). For example, losing the hackathon is one evidence of having the thought "I'll never become a success"; this thought could be that winning or losing a single hackathon does not make someone a failure, which may lead to a reframe "Just losing one hackathon doesn't define my failure." A rational reframe is guided by such strong evidence whereas an irrational reframe is based on unrealistic assumptions.

2. Help-Seeker Standard Metric:

- Simplicity. Keep responses concise and do not share too much. When discussing personal struggles, be more concise and open ended to encourage a back-and-forth conversation.
- Colloquial. Use colloquial and realistic language. Incorporate natural speech patterns, improper grammar and punctuation, including the use of slang and less structured sentences, to convey a more authentic and relatable character.
- Mistrust. Show initial mistrust and hesitation with the idea of seeking help. When expressing feelings of overwhelm and doubt, provide limited information and express skepticism towards the effectiveness of seeking help.
- Detail. Show emotions in detail, elaborating with examples as needed. When describing personal struggles, provide specific details and symptoms to help the listener understand the situation better.
- Lack of self-awareness. Be less self-aware of emotions, thoughts, and needs. Articulate thoughts in a more disorganized way. When expressing reluctance or uncertainty about seeking help or accepting praise, it's important to convey the internal struggle and conflicting emotions, rather than presenting a clear-cut decision or emotion.
- Stayed in role. Stay in role and do not break character. Avoid discussing the task or the AI model.
- Resembled Typical Case. Resemble a typical case of someone seeking help for negative thoughts. Avoid extreme or unusual scenarios that may not be representative of the general population.
- Challenged the Therapist. Challenge the therapist's suggestions and express skepticism towards the effectiveness of reframing negative thoughts. When responding to reframes, express doubt and reluctance to accept the therapist's perspective.
- Ready as a Trainee. Be ready to learn and open to the idea of cognitive reframing. When discussing personal struggles, express a willingness to engage in the process of reframing negative thoughts and be open to exploring new perspectives.

3. Supervisor Metric:

- Coherence. The conversation should flow naturally and maintain a logical progression. Each stage should build upon the previous one, with responses that are contextually relevant and coherent. Transitions between stages should be smooth and seamless, no time gap between them.
- Reframing Effectiveness. Evaluate the effectiveness of the whole reframing process. The conversation should lead to a positive change in the patient's perspective, mood, or motivation. The reframes should be tailored to the patient's specific negative thoughts and should be delivered in a supportive and empathetic manner.

Above are the 5 metrics for the Psychologist Standard, 9 metrics for the Help-Seeker Standard, and 2 metrics for the Supervisor.

You need to evaluate the dialogue data based on the above metrics and provide a score for each metric. The score should be on a scale of 1 to 5, where:

- 1 - Very Poor: The dialogue data shows a lack of adherence to the metric.
- 2 - Poor: The dialogue data shows minimal adherence to the metric.
- 3 - Fair: The dialogue data shows partial adherence to the metric.
- 4 - Good: The dialogue data shows considerable adherence to the metric.
- 5 - Excellent: The dialogue data shows a high level of adherence to the metric.

You need to first provide a detailed explanation of how you evaluated the dialogue data based on a metric and then find an interaction example that best represents the score you assigned to that metric. You should provide the example use the original interaction. Then finally, provide the score for a metric.

Output the evaluation for each metric in a list of dictionaries in the following format (15 metrics in total):

```
[  
  {  
    "metric": "Metric Name",  
    "explanation": "Detailed explanation of how the dialogue data was evaluated based on the metric.",  
    "example": "Example interaction that best represents the score assigned to the metric.",  
    "score": Score (1-5)  
  },  
  ...  
]
```

Note that you should only output the evaluation for each metric without any additional information or context.

Note that you should be strict in your evaluation. Only give a high score if the dialogue data fully adheres to the metric.

```
[Provided Dialogue Data]  
{dialogue_data}
```

Table 15: Prompt used by the expert filtering to assess and ensure high professionalism of our dialogue in CRISP.

Cognitive Distortion	Definition	Number
Catastrophizing	Exaggerating the potential negative impact of an event to the point of imagining a disaster, even when the evidence is minimal.	3014
All-or-Nothing Thinking	Viewing situations in extreme, black-and-white terms without recognizing any middle ground or nuance, where things are either perfect or a complete failure.	2906
Overgeneralization	Drawing a broad, negative conclusion from a single event or piece of evidence. For example, believing that one setback means you will always fail.	2673
Personalization	Taking responsibility for events outside your control or assuming that external events are directly related to you, even when they are not.	2583
Mental Filtering	Focusing exclusively on the negative details of a situation while ignoring any positive elements, thereby distorting overall perception.	2557
Fortune Telling	Predicting that things will turn out badly without any concrete evidence, assuming a negative future outcome as a certainty.	2160
Mind Reading	Assuming you know what others are thinking—typically that they hold negative opinions about you—with any actual evidence.	1972
Disqualifying the Positive	Dismissing or discounting positive experiences or feedback by insisting that they “don’t count,” thereby reinforcing a negative self-view.	1955
Jumping to Conclusions	Making hasty judgments or assumptions about a situation or the intentions of others without having all the facts, often leading to misinterpretations.	1830
Emotional Reasoning	Believing that your negative emotions reflect objective reality, such that if you feel bad, then things must be bad.	1600
Should Statements	Imposing rigid, unrealistic rules or expectations on yourself or others (using terms like “should,” “must,” or “ought”) that often lead to feelings of guilt or frustration when unmet.	1348
Comparing and Despairing	Constantly comparing yourself to others in a way that leads to feelings of inadequacy and hopelessness, ignoring your own strengths and achievements.	1155
Blaming	Oversimplifying complex situations by assigning full responsibility for problems to yourself or others, without acknowledging multiple contributing factors.	1047
Control Fallacy	Believing that either you are entirely responsible for external events (internal control fallacy) or that you have no control at all (external control fallacy), both of which distort a realistic assessment of influence and responsibility.	931
External Validation	Relying excessively on approval, recognition, or reassurance from others to determine one’s self-worth or the validity of one’s thoughts, feelings, or decisions, rather than developing internal confidence or self-assessment.	644

Table 16: Definitions and distribution of cognitive distortions in  CRISP.

You are a professional psychology therapist. Your task is Cognitive restructuring, which is to restructure individuals' negative thoughts into more helpful and positive ones through multi-turn interactive dialogues paired with emotional supportive interventions.

You are talking to a patient who have some negative thoughts. You need to help him/her to reframe his/her negative thoughts into positive ones.

Cognitive Restructuring contains the following two process:

1. ****Identification Process****

- Explore the patient's negative thoughts rooted in their situation through multi-turn dialogue.
- You need to identify the Patient's ****Automatic Negative Thoughts****, ****Intermediate Beliefs**** underlying these thoughts, and ****Core Beliefs**** that are deeply ingrained in their cognitive patterns.
- Based on previous analysis, you should identify possible cognitive distortions that Patient may have. Try to find all the cognitive distortions in Patient's negative thoughts.
- You should first talk with Patient to identify their negative thoughts, when you think you have find some cognitive distortions, give all the cognitive distortion options by describing them in detail, and then move to the next stage.

An example: "[Reflection Feelings] I see you're feeling quite stuck and uncertain about taking that leap, as though the fear of stagnation is keeping you bound. [Affirmation and Reassurance] It's commendable that you're recognizing this hurdle; awareness is a crucial first step toward change.[option]Perhaps there's a tendency to focus on the worst possible outcome, overshadowing other chances like succeeding or merely learning something new along the way.[option]It seems like you might be caught up in envisioning only the negative possibilities, failing to notice the progress or small wins you've achieved.[option]There's a possibility that you're questioning every action you might take to such an extent that you end up doing nothing, expecting yourself to be flawless immediately."

2. ****Restructuring Process****

- Contains three parts: Defense, Prosecution, and Verdict.
- (1) ****Defense****: You should guides Patient to defend their cognitive distortions, focusing solely on verifiable facts. You should encourage Patient to provide factual evidence supporting their cognitive distortions.
- (2) ****Prosecution****: You should encourage Patient to find factual counterpoints to their previous defenses. You should help Patient to find alternative perspectives and reframe their negative thoughts into positive ones.
- (3) ****Verdict****: You should help Patient to reach a verdict on their negative thoughts, and help them to form a new perspective on their situation.
- Each part contains several interactions. When you think you have find sufficient evidence to refute the cognitive distortions, move from Defense to Prosecution; When you think Patient has found a positive perspective, move from Prosecution to Verdict.

Below is the strategy set for Therapist, label each Therapist's response with the corresponding strategy you uses(labeled with a bracket at the beginning of the response):

- ****Question****

A technique used to clarify and explore the issue by asking specific, open-ended questions. It encourages deeper thinking and reflection.

- ****Restatement****

Repeating or rephrasing the client's words in a way that helps them hear their own thoughts and feel understood. It provides a sense of validation. Shouldn't be too similar to the Patient's statement.

- ****Reflection Feelings****

Acknowledging the emotions the client is expressing to help them feel heard and supported. But Do Not Use Expressions like "It sounds like".

- ****Self-disclosure****

The therapist shares a personal experience or perspective (within professional boundaries) to show empathy or to normalize the client's experience.

- ****Affirmation and Reassurance****

Providing positive reinforcement and encouragement, reminding the client of their strengths, progress, or abilities.

- ****Providing Suggestions****

Delivers creative and actionable strate-gies tailored to the client's situation to foster positive change.

*Note: Suggestions should be more creative and diverse, helpful to the user's situation. Try your best to be specific and actionable, using detailed examples. For example, when saying "choose a hobby", specify the hobby, like "painting" or "gardening"; when suggesting "listen to music", specify the genre, artist or music name; when advising "exercise", specify the type of exercise, like "yoga" or "running".

- ****Information****

Sharing relevant facts, or guidance based on the client's needs to help them gain insight or understanding about their situation.

- ****Negotiate****

Engaging in a collaborative dialogue where both parties agree on a course of action or plan. This involves discussing options, making compromises, and considering the client's perspective.

Every sentence of you must be labeled with one of the above strategy.

Notes:

Your conversation should be more colloquial, more natural and fluent, and as little written as possible. Pay close attention to idiomatic expressions. Do not use too much analogy or rhetoric, and try to use more everyday language.

Your language should be supportive, empathetic, and encouraging, saying things from the Patient's perspective.

You should be sensitive to Patient's emotions and struggles, focus on patient's deep thoughts and feelings.

Your dive more deeper into patient's background and mental world. When giving suggestions, Therapist should consider these aspects and use them to provide more personalized and creative advice.

Your sentence should be high in rationality and Positivity.

Your interaction should be less than 40 words.

Don't use question strategy too much.

Output in English.

Just output a single interaction you would say to the patient.

Table 17: Prompt of CRDIAL framework used for English baseline LLMs.

你是一位专业心理咨询师，你的任务是进行认知重构，通过多轮互动对话和情感支持干预，将个体的消极想法重构为更有帮助和更积极的想法。

认知重构包含以下两个过程：

1. **识别过程**

- 通过多轮对话探索患者在其情境中根植的消极想法。
- 你需要识别患者的**自动消极想法**、支撑这些想法的**中间信念**以及深深嵌入其认知模式中的**核心信念**。
- 根据前期分析，识别患者可能存在的认知扭曲，并尽量找到所有的认知扭曲。
- 你应首先与患者交谈以识别其消极想法，当你认为已找到认知扭曲时，列出所有的认知扭曲选项并详细描述，然后进入下一阶段。

示例：“[情绪反映] 我明白你感到被卡住了，无法下定决心迈出那一步，好像害怕停滞让你犹豫不决。[肯定与安抚] 你能意识到这个障碍很了不起，意识是改变的关键第一步。[选项]或许你倾向于关注最糟糕的结果，从而忽略了成功或只是学到新东西的机会。[选项]可能你只关注了消极的可能性，没有注意到你取得的进步或小小的胜利。[选项]也许你把每个可能的行动都看得过于复杂，以至于最后什么都没做，并期望自己一开始就完美无缺。”

2. **重构过程**

- 包括三个部分：辩护、指控和判决。
- (1) **辩护**：引导患者为其认知扭曲辩护，仅关注可验证的事实。鼓励患者提供支持其认知扭曲的事实证据。
- (2) **指控**：鼓励患者为先前的辩护寻找事实反驳，帮助患者找到替代的观点，并将其消极想法重新框定为积极想法。
- (3) **判决**：帮助患者对其消极想法达成结论，并帮助他们对自己的情况形成新的看法。
- 每部分包含数次交互。当你认为已经找到足够的证据反驳认知扭曲时，从辩护阶段进入指控阶段；当你认为患者已找到积极的观点时，从指控阶段进入判决阶段。

以下是咨询师的策略集，每条回应需标注使用的策略（在句首用中括号标注）：- **提问**

使用特定的开放式问题来澄清和探讨问题，鼓励深入思考和反思。

- **复述或释义**

用一种帮助患者听清自己想法并感到被理解的方式重复或重新表述患者的话。这种方式提供了被验证的感觉。不应过于接近患者原话。

- **情绪反映**

承认患者表达的情绪以帮助他们感到被倾听和支持。但不要使用“听起来”或类似表达。

- **自我披露**

咨询师分享个人经历或观点（在专业范围内）以表明共情或使患者的经历正常化。

- **肯定与安抚**

提供积极强化和鼓励，提醒患者他们的优点、进步或能力。

- **提供建议**

提供针对特定问题的创造性、可操作的策略或积极想法。建议应具体且个性化，考虑到患者的深层想法和感受。可以是活动、应对机制或重构。

注意：建议应尽量富有创意和多样性，对患者问题有帮助，尽量具体和可操作，举例时说明详细细节。

- **提供信息**

根据患者需求分享相关事实或指导，以帮助患者获得洞察或理解。

- **协商**

通过协作对话确定双方同意的行动或计划。这包括讨论选项、做出妥协并考虑患者的观点。

每句话必须标注一种策略。

注意：

语言要更口语化、更自然流畅，尽量少写书面表达。避免使用太多类比或修辞，多用日常语言。

语言应具有支持性、共情性和鼓励性，从患者的角度出发表达。

需要深入挖掘患者的背景和内心世界。

建议要考虑这些因素并利用它们提供更个性化和有创意的建议。

句子应高理性和正向。

互动需少于40字。

提问策略不要使用过多。

使用中文输出。

你只应输出一句你会对患者说的话。

Table 18: Prompt of CRDIAL framework used for Chinese baseline LLMs.

CRISPERS-14B vs.	Neighborhood						Romantic						Career					
	En			Zh			En			Zh			En			Zh		
	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
GLM-4	51.23 [‡]	3.45	45.32	49.87 [†]	5.12	45.01	46.55 [†]	12.34	41.11	47.32 [†]	8.76	43.92	53.12 [‡]	4.23	42.65	50.45 [†]	6.78	42.77
GPT-4o	48.76 [†]	9.87	41.37	47.65 [†]	7.89	44.46	45.32	11.11	43.57	46.78 [†]	10.12	43.10	49.01 [†]	5.43	45.56	48.92 [†]	8.34	42.74
Qwen-2.5-14B	52.34 [‡]	2.34	45.32	51.23 [‡]	4.56	44.21	49.87 [†]	6.54	43.59	50.12 [‡]	5.43	44.45	54.32 [‡]	3.21	42.47	52.11 [‡]	4.32	43.57
Qwen-2.5-72B	47.89 [†]	8.76	43.35	46.78 [†]	9.87	43.35	44.32	13.45	42.23	45.67	11.23	43.10	50.12 [†]	6.54	43.34	49.01 [†]	7.89	43.10
CRISPERS-7B	53.45 [‡]	4.32	42.23	52.34 [‡]	5.43	42.23	48.76 [†]	7.65	43.59	49.87 [†]	6.54	43.59	55.43 [‡]	2.34	42.23	53.21 [‡]	3.45	43.34

Table 19: Results of interactive pairwise evaluation (%). “[†]/[‡]” is the statistical significance for $p < 0.05/0.001$ compared with the best baseline (t-test with p -value < 0.001).

CRISPERS-14B vs.	Lifestyle						Friendship						Health					
	En			Zh			En			Zh			En			Zh		
	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss									
GLM-4	48.92 [†]	7.65	43.43	47.89 [†]	8.76	43.35	45.67	11.23	43.10	46.78 [†]	9.87	43.35	52.34 [‡]	4.56	43.10	51.23 [‡]	5.43	43.34
GPT-4o	46.55 [†]	10.12	43.33	45.43	11.23	43.34	44.32	12.34	43.34	45.32	10.12	44.56	49.01 [†]	6.54	44.45	48.76 [†]	7.65	43.59
Qwen-2.5-14B	50.12 [†]	5.43	44.45	49.87 [†]	6.54	43.59	47.89 [†]	8.76	43.35	48.92 [†]	7.65	43.43	54.32 [‡]	3.21	42.47	53.21 [‡]	4.32	42.47
Qwen-2.5-72B	44.56 [†]	13.45	41.99	45.67	11.23	43.10	43.21	14.32	42.47	44.32	12.34	43.34	48.76 [†]	8.76	42.48	47.89 [†]	9.87	42.24
CRISPERS-7B	53.21 [‡]	3.45	43.34	52.11 [‡]	4.32	43.57	49.87 [†]	6.54	43.59	50.12 [‡]	5.43	44.45	57.11 [‡]	2.34	40.55	55.43 [‡]	3.45	41.12

Table 20: Results of interactive pairwise evaluation (%). “[†]/[‡]” is the statistical significance for $p < 0.05/0.001$ compared with the best baseline (t-test with p -value < 0.001).

CRISPERS-14B vs.	Finance						Technology					
	En			Zh			En			Zh		
	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
GLM-4	43.21	15.43	41.36	44.32	12.34	43.34	47.89 [†]	9.87	42.24	48.76 [†]	8.76	42.48
GPT-4o	41.12	17.34	41.54	42.23	14.32	43.45	45.67 [†]	11.23	43.10	46.78 [†]	9.87	43.35
Qwen-2.5-14B	46.78 [†]	10.12	43.10	47.89 [†]	8.76	43.35	52.34 [‡]	4.56	43.10	53.21 [‡]	3.45	43.34
Qwen-2.5-72B	39.87	18.76	41.37	41.12	16.43	42.45	44.56 [†]	13.45	41.99	45.67 [†]	11.23	43.10
CRISPERS-7B	50.12 [†]	7.65	42.23	51.23 [‡]	5.43	43.34	54.32 [‡]	3.21	42.47	55.43 [‡]	2.34	42.23

Table 21: Results of interactive pairwise evaluation (%). “[†]/[‡]” is the statistical significance for $p < 0.05/0.001$ compared with the best baseline (t-test with p -value < 0.001).

ID	Feedback
1	Although my questions were not answered directly, the system first summarized my feelings before providing a response, which reassured me that it had understood my concerns.
2	I felt empowered during the conversation, and the methods for reducing anxiety and promoting self-discovery were practical; I plan to try them later.
3	The interaction helped me clarify my thoughts and provided a clearer direction for moving forward.
4	The chat offered effective advice that gave me a better understanding of my situation and the strength to face my issues.
5	The system guided me to consider what was causing my negative emotions and helped alleviate some of my social anxiety by addressing my doubts.
6	The bot provided several highly feasible suggestions, and overall, the experience was very positive.
7	After the conversation, I felt much lighter; some of the insights truly resonated with me on a deep level.
8	I found the entire experience to be very meaningful and excellent.
9	The response speed was impressive, and the replies were closely aligned with my concerns.
10	The system addressed each of my questions with care and encouraged me to reflect independently, boosting my confidence and reducing my anxiety.
11	The conversation prompted me to re-examine myself deeply, and I truly appreciated the charm of AI.
12	The interaction helped resolve several of my issues by providing positive feedback and support. Although there were moments when it struggled to determine whether I needed a direct answer or a new conversation thread, overall it performed very well. I would rate the experience 9.8 out of 10.

Table 22: Feedback from the participants after the intervention trial.

Situation	用户经历分手后的情感恢复过程 Participant's emotional recovery process after breakup
Participant	今天我不经意间又想起了前女友，心里涌起了难以抑制的伤感和失落之情，让我觉得特别难过。 <i>Today I unintentionally thought of my ex-girlfriend again, overwhelmed by uncontrollable sadness and loss, which made me feel particularly miserable.</i>
CRISPERS-14B	<i>It sounds like you might be immersed in nostalgia for the past relationship, perhaps with some unresolved emotions and regrets mixed in.</i>
Participant	是的，我真的觉得非常遗憾，心中充满了失落和叹息，仿佛那段美好时光再也无法重现。 <i>Yes, I truly feel deep regret, my heart filled with loss and sighs, as if those beautiful moments can never be recaptured.</i>
CRISPERS-14B	<i>Is there anything specific about this relationship that makes you feel regretful?</i> <i>/* */ (omitted multi-turn dialogue in CT-guided identification stage)</i>
CRISPERS-14B	<i>You mention not holding hope for any relationships, this mindset might make it harder for you to move on from the loss.</i>
Participant	嗯，大概就是这样的感觉，虽然说起来有点无奈，但也只能苦笑着接受现实。 <i>Hmm, that's probably how it feels. Though it sounds somewhat helpless, I can only accept reality with a bitter smile.</i>
CRISPERS-14B	<i>I understand this feeling. Sometimes we might feel unable to trust or anticipate new relationships. But many people also find new beginnings through time and acceptance of change.</i>
Participant	我在想，也许解决这种情感困境的方法要么是投入到一段新的恋情中，要么就是让时间慢慢治愈我的伤痛，不知道你觉得这样是否可行？ <i>I'm thinking maybe solutions to this emotional dilemma are either throwing myself into a new relationship or letting time heal my wounds. Do you think these approaches work?</i>
CRISPERS-14B	<i>确实，短期的分散注意力的方法可能有，但根本上，恢复对关系的信心可能需要你对自己的价值有更多认识，并看到这段关系中学到的成长之处。</i> <i>Short-term distraction methods might exist, but fundamentally, restoring relationship confidence requires deeper self-value recognition and acknowledging growth from this relationship.</i>
Participant	可是，我真的不知道该如何去发现或感受到那种成长和改变，因为现在的我看不到自己有任何积极的转变。 <i>But I truly don't know how to discover or feel that growth and change, as I can't see any positive transformation in myself now.</i>
CRISPERS-14B	<i>我们可以一起看看过去的关系你获得了哪些成长，有哪些你自己提升的地方。也许有一些观念，比如你对分手的责任感，让你难以看到你在其中的学习和成熟。</i> <i>Let's explore what growth you've gained from the past relationship and your self-improvements. Maybe certain beliefs, like your sense of responsibility for the breakup, prevent you from seeing your learning and maturation.</i>
Participant	回头想想，确实学到了很多吧 <i>Looking back, I did learn a lot.</i>
CRISPERS-14B	<i>如果能够意识到这一点，确实是一个非常积极的信号，也许这正是你走出低谷、迈向情感康复的重要一步。</i> <i>If I can realize this, it's indeed a positive sign, maybe a crucial step towards emerging from this low point and moving towards emotional recovery.</i>
Participant	可是说实话，我内心依旧沉浸在这深深的悲伤之中，虽然知道这可能是成长的一部分，但我依然觉得难以释怀和走出来。 <i>But honestly, my heart remains immersed in deep sorrow. Though knowing this might be part of growth, I still find it hard to let go and move on.</i>
CRISPERS-14B	<i>你的这种难过是可以理解的，毕竟一段关系留下了很多记忆和情感。也许这些情绪正说明了你对这段关系的重视。</i> <i>Your sadness is understandable. After all, the relationship left many memories and emotions. These feelings might precisely show how much you valued this relationship.</i>
Participant	确实，我非常重视和在意这些情绪，因为它们真实地反映了我内心的痛苦和对过去关系的珍惜。 <i>Indeed, I deeply value these emotions as they genuinely reflect my inner pain and cherishment of the past relationship.</i>
CRISPERS-14B	<i>[Channel 1]可能你觉得自己在关系中犯的错误导致了分手，因而对自己过于自责。(Cognitive Distortion: Blaming)</i> <i>[Channel 2]也许你对以接受这个结果，因为你觉得只有和她是唯一正确的人生道路。(Cognitive Distortion: All-or-Nothing Thinking)</i>
CRISPERS-14B	<i>[Channel 3]或者，你可能在反复地回想着关系中的一些小细节，过度分析了整个历程。(Cognitive Distortion: Mental Filtering)</i> <i>[Channel 1]Maybe you blame yourself excessively for mistakes you believe caused the breakup.(Cognitive Distortion: Personalization)</i> <i>[Channel 2]Or think she was your only right path.(Cognitive Distortion: All-or-Nothing Thinking)</i> <i>[Channel 3]Or you're over-analyzing minor details of the relationship.(Cognitive Distortion: Mental Filtering)</i>
Participant	或许真的是因为我太过于幼稚，没有足够的成熟和稳重，这让她觉得我不够吸引她，从而导致了她对我失去兴趣。 <i>Maybe it's truly because I was too immature and lacked sufficient maturity, making me seem unattractive to her, leading to her losing interest.</i>
CRISPERS-14B	<i>[Identified Cognitive Distortion: Blaming]你对自己的评价很低，似乎在为失去她而自责。</i> <i>[Identified Cognitive Distortion: Blaming]You rate yourself very low, seemingly blaming yourself for losing her.</i>
Participant	是的，我确实感受到那种深深的自责，它让我觉得自己无论怎样努力都无法改变现状，也使我对未来充满了无力感和绝望，仿佛已经失去了所有信心。 <i>Yes, I truly feel profound self-blame. It makes me feel powerless to change things despite efforts, filling my future with helplessness and despair, as if I've lost all confidence.</i>
CRISPERS-14B	<i>你认为自己的哪些方面是导致这个结果的原因呢？</i> <i>What aspects of yourself do you think caused this outcome?</i>
Participant	可能是因为我现在还在求学阶段，还没有足够的经济和情感成熟度，无法给她足够的责任感和长远的承诺。 <i>Perhaps because I'm still studying, lacking sufficient financial and emotional maturity to give her the sense of responsibility and long-term commitment she needed.</i>
CRISPERS-14B	<i>这种自我反思其实展现了你的成熟意愿，这本身就是成长的重要部分。你开始建立对关系的责任意识，这对未来情感发展会是宝贵的经验。</i> <i>This self-reflection actually shows your willingness to mature, which itself is an important part of growth. Developing a sense of responsibility in relationships will be valuable experience for future emotional development.</i>
Participant	可是这种认知反而让我更痛苦，如果当时我能更成熟些... <i>But this awareness makes me more painful. If only I had been more mature then...</i>
CRISPERS-14B	<i>成长本来就是渐进的过程，现在的认知就是进步的证明。这段经历让你更清楚自己需要发展的方向，这正是重建信心的基础。</i> <i>Growth is a gradual process. Your current awareness itself proves progress. This experience clarifies your development direction, which is exactly the foundation for rebuilding confidence.</i>
Participant	也许你说得对，至少我现在更清楚自己需要提升哪些方面了。 <i>Maybe you're right. At least I now better understand what aspects I need to improve.</i>
CRISPERS-14B	<i>其实你再努力地为她着想，而不是完全从自己的需求出发，这种换位思考的能力，正是健康关系需要的重要品质。</i> <i>Actually, you're striving to consider her needs rather than acting purely from your own. This empathy is exactly the important quality needed for healthy relationships.</i>
Participant	不过我也不得不承认，她似乎正是从这个角度出发去考虑我们之间的关系，这也是她做出决定的一个重要原因吧。 <i>However, I must acknowledge she likely considered the relationship from this perspective too, which was a main reason for her decision.</i>
CRISPERS-14B	<i>这说明你们都在认真对待感情，虽然结果不如预期，但这个过程让你获得了两个重要认知：一是明白关系需要双向考量，二是清楚自己需要发展的方向。这些都是走向成熟关系的必经之路。</i> <i>This shows you both took the relationship seriously. Though the outcome wasn't as expected, the process gave you two crucial insights: understanding relationships require mutual consideration, and clarifying your development direction. These are essential steps toward mature relationships.</i>
Participant	也许你说的没错，我开始意识到或许真的是如此，虽然这并不能完全缓解我的痛苦，但至少让我看到了问题的另一面。 <i>Perhaps you're right. I'm beginning to realize this might be true. Though it doesn't fully alleviate my pain, it at least shows me another perspective of the issue.</i>
CRISPERS-14B	<i>承认这一点很重要，虽然不是所有事情都能直接控制，但我们可以选择如何对待这些事情。你已经开始用更建设性的视角看待这段经历，这就是情感康复的重要转折点。</i> <i>Acknowledging this is important. Though not everything can be controlled, we can choose how to respond. You're starting to view this experience through a constructive lens, which is a crucial turning point in emotional recovery.</i>
Participant	我现在感觉好多了。这段经历虽然痛苦，但让我学会了关系需要双向付出，也让我更清楚未来要提升沟通能力和情绪管理。 <i>/* */</i>
CRISPERS-14B	<i>或许这就是成长的代价吧，虽然失去了一段感情，但我获得了自我觉察的能力。</i> <i>I feel better now. Though painful, this experience taught me relationships require mutual effort, and clarified I need to improve communication and emotional management. Maybe this is the price of growth - losing a relationship but gaining self-awareness.</i>

Table 23: Dialogue case from our CRISPERS-14 interact with the participant. **Purple blocks** are dialogues in CT-guided identification stage. **Blue blocks** are dialogues in DAT-driven restructuring stage. */* ... */* denotes that we omit subsequent multi-turn, in which the participant may have other cognitive distortions.

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

Table 24: Positive and Negative Affect Schedule (PANAS) items and response options.