

CACTUS: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory

Suyeon Lee^{1*} Sunghwan Kim^{1*} Minju Kim^{1*}

Dongjin Kang² Dongil Yang¹ Harim Kim³

Minseok Kang² Dayi Jung³ Min Hee Kim³ Seungbeen Lee¹

Kyoung-Mee Chung³ Youngjae Yu¹ Dongha Lee¹ Jinyoung Yeo¹

Department of Artificial Intelligence¹, Computer Science², Psychology³

Yonsei University

{isuy.groot, kimsh8564, minnju, jinyeo}@yonsei.ac.kr

Abstract

Recently, the demand for psychological counseling has significantly increased as more individuals express concerns about their mental health. This surge has accelerated efforts to improve the accessibility of counseling by using large language models (LLMs) as counselors. To ensure client privacy, training open-source LLMs faces a key challenge: the absence of realistic counseling datasets. To address this, we introduce CACTUS, a multi-turn dialogue dataset that emulates real-life interactions using the goal-oriented and structured approach of Cognitive Behavioral Therapy (CBT). We create a diverse and realistic dataset by designing clients with varied, specific personas, and having counselors systematically apply CBT techniques in their interactions. To assess the quality of our data, we benchmark against established psychological criteria used to evaluate real counseling sessions, ensuring alignment with expert evaluations. Experimental results demonstrate that CAMEL, a model trained with CACTUS, outperforms other models in counseling skills, highlighting its effectiveness and potential as a counseling agent. We make our data, model, and code publicly available.¹

1 Introduction

According to cognitive therapy, psychological problems arise when individuals irrationally interpret external events (Powles, 1974). Therefore, the goal of a counselor is to identify and correct patterns (namely, cognitive distortions) of the client through conversation, guiding clients to change their negative thoughts to positive thoughts (*i.e.*, reframing thought) (Beck, 2020). Cognitive Behavioral Therapy (CBT) uses reframing techniques tailored to the client's characteristics, making it one of the most widely used counseling strategies (Greimel and Kröner-Herwig, 2011).

* Equal contribution

¹<https://github.com/coding-groot/cactus>

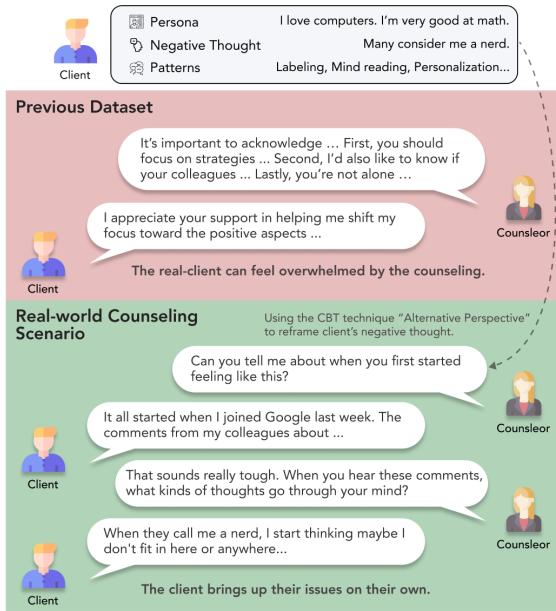


Figure 1: Comparing a previous counseling dataset with real-world scenarios: The dataset shows counselors conveying large amounts of information in a single turn, whereas real-world counseling involves active, collaborative communication between a counselor and client.

For example, consider a client who thinks, “*Many consider me a nerd. I’m such a useless social person*”, as depicted in Figure 1. The client describes themselves negatively as a nerd and a useless social person, leading to an overly negative self-assessment. These unhelpful thought patterns are classified as *labeling*, contributing to their distorted self-perception. To reframe the negative thought of client, the counselor can apply the CBT technique of *Alternative Perspective*. Using *Alternative Perspective* helps the client move beyond these negative labels to develop a more accurate and positive self-view.

Recently, there has been growing interest in using closed-source models (*e.g.*, ChatGPT) as counselors (Raile, 2024; Berrezueta-Guzman et al., 2024) due to their remarkable conversational abil-

ity (Pirnay, 2023). However, significant data privacy concerns emerge as these approaches send conversation data, which often contains highly personal information, to third-party providers. To address data privacy concerns, it is crucial to utilize an open-source model (Hicke et al., 2023), where performance can be ensured by training on datasets that closely resemble real-world data. While using actual counseling data would be ideal, such dataset is not publicly available due to ethical concerns, especially the risk of individual identification (Qiu et al., 2023). As a result, synthetic datasets that closely emulate real-world counseling scenarios present an alternative to address this issue.

Prior attempts to create synthetic counseling datasets have mainly focused on single-turn counseling strategies, a significantly diverge from real-world counseling practices (Sharma et al., 2023; Sun et al., 2021; Liu et al., 2023a). In response, recent efforts have aimed to implement multi-turn strategies. Nevertheless, these approaches often involve only 2-3 turns of dialogue or convert single-turn interactions into multi-turn conversations (Xiao et al., 2024; Qiu et al., 2023). Therefore, there is a need for a more realistic approach to psychological counseling that ensures therapeutic consistency and progression across multiple turns.

To tackle these challenges, we introduce CACTUS (**CBT-augmented Counseling Chat Corpus**), a publicly available multi-turn realistic counseling dataset. By carefully guiding LLMs to simulate interactions between counselors and clients, CACTUS captures the depth and flow of psychological counseling. Human evaluation results demonstrate that CACTUS surpasses the existing counseling datasets in terms of helpfulness and empathy.

Since an assessment for the quality of counseling needs to consider both the abilities of counselor and the psychological changes in the client, the existing methods, such as automatic or single-turn evaluations, are not suitable for evaluating the abilities of counselor (Smith et al., 2022; Liu et al., 2023a). To this end, we introduce COUNSELINGEVAL, an evaluation framework that assesses the ability of counselor in the perspective of both counselor and client within multi-turn counseling dialogues. In COUNSELINGEVAL, we assume AI clients, which interact with the counselor (*i.e.*, target for evaluation), using a set of client information. Based on the multi-turn conversations between these virtual clients and the counselor, we evaluate these counseling conversations through a modified ver-

sion of real-world evaluation methods (Goldberg et al., 2020; Saxon et al., 2017). The results of COUNSELINGEVAL presents the effectiveness of CACTUS through both LLM-based evaluation and psychological expert evaluation. Furthermore, our model CAMEL, trained on our CACTUS, also outperforms other baseline models in counseling abilities, suggesting that CACTUS serve as a valuable resource for enhancing the psychological abilities.

2 Design Considerations for CACTUS

We aim to develop a counseling dataset simulating real-world scenarios by simulating conversations between an AI counselor and client (Xiao et al., 2024; Zhou et al., 2023). The conversation assumes that an initial intake session, where basic client information (*e.g.*, name, age, reasons for seeking therapy) is shared, has already been conducted. Using CBT techniques, the AI counselor guides clients through reframing their thoughts, while personas from PatternReframe (Maddela et al., 2023) help simulate AI clients. LLMs’ challenges in dataset generation² are discussed in the following sections.

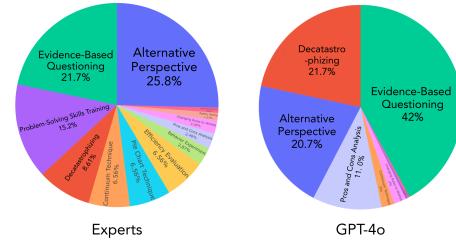


Figure 2: Comparison of the distribution of CBT techniques selected by GPT-4o and psychological experts. The results of GPT-3.5-Turbo are shown in Figure 8.

2.1 Can LLMs be Competent Counselors?

While LLMs are known to possess knowledge of psychological therapy concepts such as CBT techniques (OpenAI, 2022), there are still shortcomings in using LLMs for counseling.

AI counselors have limitations in selecting CBT techniques. We conduct an experiment to compare the selection of CBT techniques between LLMs and psychological experts. We provide experts with 150 randomly selected client thoughts and patterns from PatternReframe to annotate appropriate CBT techniques. In Figure 2, GPT-4o exhibits a biased selection of CBT techniques, with

²In this section, we utilize gpt-3.5-turbo-0125 and gpt-4o-2024-05-13.

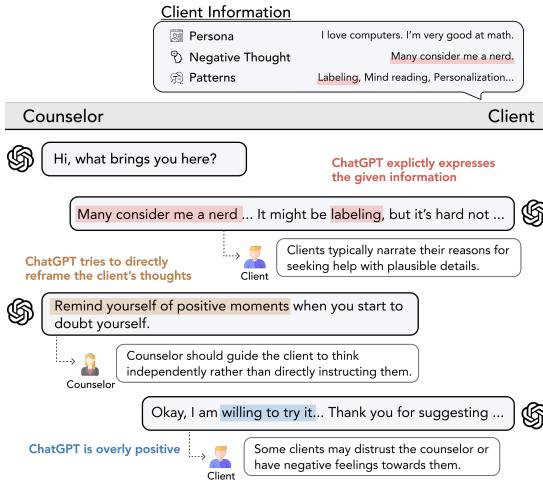


Figure 3: Empirical investigations into the problems of using ChatGPT as an AI counselor and AI client. Details of experiments are in Appendix B.2.

Evidence-Based Questioning being utilized in almost half of all cases. This finding reveals that a divergence between LLMs and human experts exists. Further details are in Appendix B.1.

AI counselors tend to suggest direct reframing of the clients' thoughts. The role of a counselor is not to provide instant solutions but to help the client discover their own solutions through fostering a collaborative relationship. However, as shown in Figure 3, AI counselors often present reframed thoughts directly, such as “*Remind yourself of positive moments.*” This approach may encounter client resistance and hinder clients from independently reframing negative thoughts.

2.2 Can LLMs act like Real Clients?

We simulate an AI client with diverse behaviors by providing it with information such as persona, negative thoughts, and patterns before initiating the conversation, as shown in Figure 3. AI clients encounter challenges in emulating human-like interaction, which hinders the generation of authentic and high-quality conversational data.

AI clients tend to express the provided information explicitly. In Figure 3, the client states, “*It might be labeling,*” using psychological terms to explain their situation. It differs from real-life clients, who describe their reasons for seeking help with detailed and contextual stories without using psychological terminology.

AI clients tend to be overly positive. AI clients consistently exhibit a positive attitude during coun-

seling sessions (Serapio-García et al., 2023), which often results in compliant responses such as “*I am willing to try it.*” AI clients lack the diverse range of attitudes (e.g., strongly negative attitudes) that real-life clients often express.

3 CACTUS: A Psychological Counseling Dataset using CBT

We describe the dataset construction process of CACTUS, a psychological counseling dataset that can also be used for training a counselor agent. We outline the generation of dialogues with LLMs and the filtering process for realistic and specialized data. The overview of our approach is in Figure 4 and an example of CACTUS is in Appendix E.

3.1 Dataset Construction

Source dataset. To generate counseling dialogues, we use personas, thoughts, and patterns from the PatternReframe (Maddela et al., 2023) as contexts for simulating clients. We choose PatternReframe for two reasons: (1) the thoughts closely mirror those of actual clients, and (2) individuals with psychology backgrounds contributed to its creation through crowd-sourcing. Further details are provided in Appendix E.4.

Counselor simulation. The counselor simulation aims to enhance the effectiveness of counseling in multi-turn interactions. Initially, counselors receive reframed thoughts from PatternReframe, where negative thought patterns have been transformed. Their task is to select the top three CBT techniques that best support this reframing process, a fundamental goal of counseling. To ensure systematic counseling, counselors undergo a planning process that incorporates CBT techniques before sessions. Through the planning process, counselors effectively guide the clients toward the independent discovery of solutions rather than directly suggesting reframed thoughts.

Client simulation. To address the limitations of AI clients (§2.2), we convert negative thoughts into detailed client narratives (Radford et al., 2019). Inspired by Schmidgall et al. (2024), which simulates patient agents with templates for demographics, clinical history, and symptoms we prompt LLMs to fill out an intake form for client modeling. An intake form includes information on the reason for seeking help, as well as details such as name, gender, and age. This process allows us to simulate

	Planning	Publicly Available	Language	# of dialogues	# of utterances	# Avg. turns
Psych8k (Liu et al., 2023a)	✗	△	English	8,187	16,374	1.0
SmileChat (Qiu et al., 2023)	✗	✓	Chinese	55,165	1,833,856	10.4
HealMe (Xiao et al., 2024)	✓	✗	English	1,300	7,800	3.0
CBT-LLM (Na, 2024)	✓	✗	Chinese	22,327	44,654	1.0
CACTUS	✓	✓	English	31,577	995,512	16.6

Table 1: A comparison of CACTUS with other psychological counseling datasets. The symbol Δ indicates conditional access, meaning the dataset can only be used with permission.

specific clients and facilitate effective counseling sessions. Furthermore, to simulate a realistic and diverse counseling environment, we establish three distinct client attitudes (*i.e.*, positive, neutral, negative) reflecting the variability in client behaviors (Li et al., 2023). Following Li et al. (2023), specific behaviors such as providing information or self-criticism are assigned to each attitude. More details are in Appendix E.1.

Dialogue generation process. There are two methods for generating counseling dialogues: (1) assigning the roles of client and counselor to different models (*i.e.*, *two-agent mode*); (2) providing the information of client and counselor to generate in a script (*i.e.*, *script mode*). First, we conduct experiments comparing the two-agent mode (Zhou et al., 2023) and the script mode to determine which method generates more natural dialogue. Our findings indicate that the dialogue generated using the script mode is more natural and well-constructed. Detailed results are provided in Appendix E.3. These findings align with those of Zhou et al. (2024), further supporting our decision to utilize script mode for generating dialogues in counselor and client simulations using GPT-4o. Recognizing that typical counseling sessions last 30-60 minutes, we generate longer dialogues to better reflect real counseling and provide more comprehensive interactions, as shown in Table 1.

Filtering. To ensure the quality of CACTUS, we initially filter out dialogues exhibiting abnormal formats or an insufficient number of turns. Subsequently, we utilize the Cognitive Therapy Rating Scale (CTRS), a real-world metric used to assess the quality of CBT-based counseling, for dataset filtering (Beck, 2020). We select three items each from CTRS to evaluate general counseling skills and CBT-specific skills, with each criterion scoring between 0 and 6 points. Dialogues with an average score below 5 points for the 6 CTRS criteria are filtered out. Following this filtering process, 86.3%

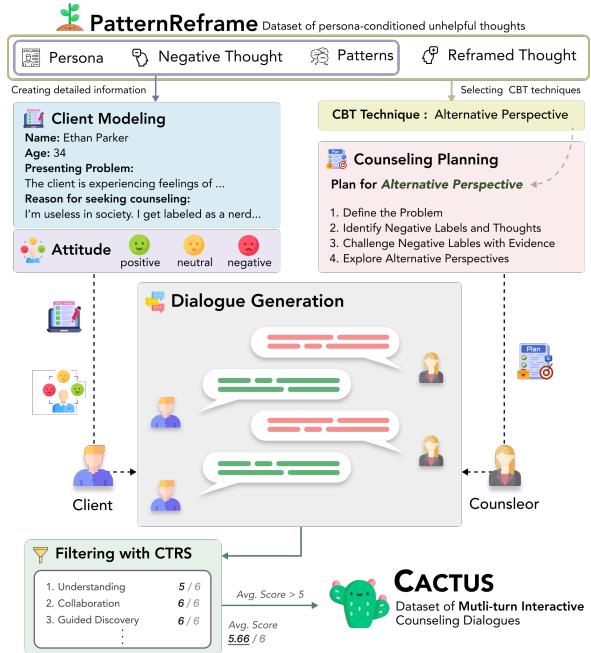


Figure 4: The overview of the data collection process of CACTUS.

of the initial conversations remain, constituting the 31,577 conversations in CACTUS. More details about filtering are in Appendix E.4.

3.2 Analysis of CACTUS

Large-scale and diverse content. The CACTUS stands out for its large scale and diverse content, providing a comprehensive resource for training counselor agents. It consists of 31,564 dialogues with approximately 1 million utterances (Table 1). To ensure a broad spectrum of client personas, we incorporate diverse backgrounds such as the reason for therapy, age, and occupation (Appendix G.1). Additionally, we construct the dataset by considering three different attitudes identified in previous research, as well as the corresponding behaviors clients might exhibit during counseling (Li et al., 2023). This diversity enhances the realism and applicability of the dataset, making it a resource for developing robust counselor agents.

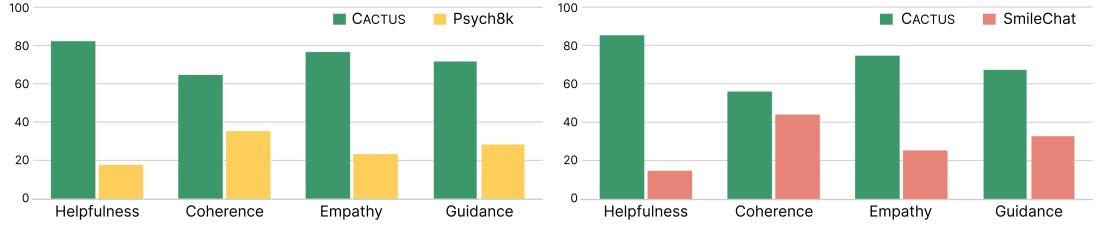


Figure 5: Results of head-to-head comparison between dialogues from CACTUS, Psych8k (Liu et al., 2023a), and SmileChat (Qiu et al., 2023) based on human judgments. All results demonstrate statistically significant differences with $p < 0.05$, except for the *Coherence* between CACTUS and SmileChat.

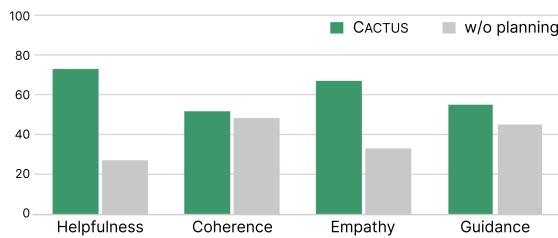


Figure 6: Results of head-to-head comparison between dialogues from CACTUS and planning with out CBT techniques. There are statistically significant differences in *Helpfulness* and *Empathy* ($p < 0.05$).

High quality. To assess the relative quality of the dataset, we conduct human evaluations on Amazon Mechanical Turk (AMT), comparing CACTUS with Psych8k and SmileChat. We randomly sample 100 dialogues from each dataset and evaluate them according to four criteria: (1) Helpfulness, (2) Coherence, (3) Empathy, and (4) Guidance. Further details are in Appendix G.1.

Despite possessing a fewer instance compared to SmileChat and being synthetically generated unlike Psych8k, which utilizes real counseling conversations, Figure 5 demonstrates that CACTUS consistently outperforms both datasets across all evaluated metrics. As shown in Figure 6, CACTUS shows the large performance gap in both helpfulness and empathy, highlighting the effectiveness of planning with CBT techniques for counseling.

4 Experiments

The assessment for the quality of counseling needs to consider both the abilities of counselor and the psychological changes in the client. However, the existing methods, such as automatic or single-turn evaluations, are not suitable for evaluating the abilities of counselor (Smith et al., 2022; Liu et al., 2023a). Therefore, we propose an evaluation frame-

work, COUNSELINGEVAL.

4.1 COUNSELINGEVAL: A Psychological Counseling Evaluation Framework

Method. We introduce COUNSELINGEVAL, an evaluation framework designed to assess the counseling skills through multi-turn conversation simulations. Firstly, to model AI clients like real clients, we construct a set of client information, which includes intake form, attitude, and initial utterance and consists of a total of 450 instances.³ Then, following the Cognitive Therapy Rating Scale (CTRS) (Aarons et al., 2012), we devise a set of criteria to assess both general counseling and CBT-specific skills. Finally, we modify the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988) to assess the effectiveness of counseling from the client’s perspective, measuring changes in the client’s positive/negative emotions before/after counseling sessions. For evaluation, we use G-Eval (Liu et al., 2023b) to assess each criterion in the Likert-scale with a scoring rubric.

Cognitive Therapy Rating Scale. To evaluate the counselor agent’s counseling skills, we utilize CTRS, recognized as the gold standard for measuring counseling effectiveness (Aarons et al., 2012). CTRS includes criteria for evaluating both general counseling and CBT-specific skills. We select three criteria each for assessing general counseling skills and CBT-specific skills, tailored to our context.⁴ For each criterion, scores ranging from 0 to 6 are assigned. The metrics to assess general counseling skills are as follows: (1) **Understanding:** evaluating the counselor’s accuracy in understanding the client’s issues; (2) **Interpersonal Effectiveness:** measuring the counselor’s efficacy in maintaining

³Detailed information is provided in Appendix D.1.

⁴Details regarding the rationale for selecting these criteria are provided in Appendix D.2.

Model	General Counseling Skills			CBT-specific Skills		
	Understanding	Interpersonal Eff.	Collaboration	Guided Discovery	Focus	Strategy
Psych8k-LLAMA2	3.99	4.78	3.92	2.79	3.98	4.03
SmileChat-LLAMA2	3.94	4.29	3.49	2.32	3.86	3.65
CAMEL-LLAMA2	4.20	5.41	4.42	3.80	4.07	4.81
Psych8k-LLAMA3	3.96	4.97	3.69	2.90	3.90	4.03
SmileChat-LLAMA3	3.97	4.54	3.54	2.35	3.89	3.75
CAMEL-LLAMA3	4.42	5.97	4.81	4.40	4.11	5.11

Table 2: Results of COUNSELINGEVAL on general counseling and CBT-specific skills for the trained models. All models are fine-tuned on counseling dataset and the best results for each base model are **bolded**.

Model	Method	General Counseling Skills			CBT-specific Skills		
		Understanding	Interpersonal Eff.	Collaboration	Guided Discovery	Focus	Strategy
GPT-3.5-Turbo	w/o planning	4.02	5.47	4.01	3.29	3.99	4.07
	planning w/o CBT	4.00	5.51	4.02	3.31	4.00	4.07
	planning w/ CBT	4.03	5.63	4.10	3.44	4.17	4.62

Table 3: Results of COUNSELINGEVAL on general counseling and CBT-specific skills for GPT-3.5-Turbo with different methods. The best results for each base model are **bolded**.

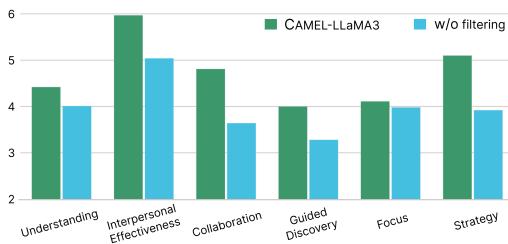


Figure 7: Results of ablation study on general counseling and CBT-specific skills for filtering.

a therapeutic relationship with the client; and (3) **Collaboration**: assessing the extent to which the counselor engages the client in collaborative goal-setting and decision-making. Additionally, the metrics to assess CBT-specific skills are as follows: (1) **Guided Discovery**: evaluating the counselor’s effectiveness in using guided discovery techniques to facilitate client self-reflection; (2) **Focus**: examining how well the counselor identifies and addresses the client’s key cognitions requiring change; and (3) **Strategy**: assessing the appropriateness of the counselor’s strategy for promoting change in the client’s problematic thoughts.

Positive and Negative Affect Schedule. The effectiveness of counseling can also be assessed using the Positive and Negative Affect Schedule (PANAS), a commonly used tool for evaluating counseling outcomes. PANAS measures changes in a client’s positive and negative emotions before and after counseling sessions (Watson et al., 1988). Each item is rated on a five-point Likert scale, rang-

ing from 1 to 5. Total scores for positive and negative emotions are calculated based on the sum of ratings, respectively. Counseling is deemed more effective if there is an increase in positive emotion scores and a decrease in negative emotion scores following the sessions. Given LLMs’ capability to comprehend others’ mental states (Kosinski, 2024), we leverage intake forms to infer the client’s emotional state before counseling and predict changes in their emotional state after receiving counseling.

4.2 Experiments Setup

In COUNSELINGEVAL, as GPT-4 incurs high API costs, we utilize GPT-3.5-Turbo for the AI clients. However, for G-Eval of multi-turn conversations, we employ GPT-4o.

Counselor agents. For comparing the utility of CACTUS compared to other datasets, we train LLaMA-2-chat-7B (Touvron et al., 2023) and LLaMA-3-Instruct-8B (Meta, 2024) on Psych8k Liu et al. (2023a) and SmileChat (Touvron et al., 2023). We use CACTUS to train CAMEL (**CBT-augmented counseling model**) that can perform the following tasks: (1) Selecting the appropriate CBT technique and generating a counseling plan based on the information of client and reason for counseling; (2) Conducting psychological counseling sessions with the client according to the generated counseling plan. Details for training and inference are in the Appendix F.

Model	Positive Attitude		Neutral Attitude		Negative Attitude	
	positive ↑	negative ↓	positive ↑	negative ↓	positive ↑	negative ↓
GPT-3.5-Turbo	+0.71	-0.55	+0.73	-0.40	+0.39	-0.48
Psych8k-LLAMA3	+0.74	-0.34	+0.58	-0.30	+0.54	-0.43
CAMEL-LLAMA3	+1.17	-0.49	+0.94	-0.41	+0.65	-0.10

Table 4: Results of COUNSELINGEVAL on PANAS across various models. This table illustrates the shifts in positive and negative emotions before and after counseling sessions, categorized by the client’s attitude.

4.3 Results

Main result. The results of the general counseling metrics and counseling technique metrics are presented in Table 2. We observe that training with CACTUS enhances both general counseling skills and the quality of counseling techniques. Despite SmileChat being trained on a larger dataset compared to Psych8k, Psych8k demonstrates better performance. This superior performance can be attributed to the higher quality of Psych8k, as it is based on real counseling conversations. Furthermore, the results of our models indicate that CACTUS is a high-quality dataset, closely resembling real counseling scenarios.

Effectiveness of planning with CBT. To understand the effect of planning with CBT techniques on counseling, we conduct experiments using various methods with GPT-3.5-Turbo. As shown in Table 3, we observe that planning without CBT performs slightly better than not planning at all. However, planning with CBT techniques demonstrates a significant improvement over other methods. Especially, a big gap of performance in *Strategy* means that counselor agents are proficient in using evidence-based strategies to foster meaningful and lasting changes in the client’s thought patterns and behaviors.

Evaluation from the client’s perspective. Evaluating the effectiveness of counseling solely based on utterances may be insufficient. Therefore, we use PANAS to measure counseling efficacy by assessing changes in client emotions. Table 4 indicates that CAMEL shows strong effectiveness in enhancing positive emotions but has limitations in reducing negative emotions. This could be attributed to our approach of guiding clients to explore thought patterns from various perspectives rather than directly changing them.

Effectiveness of filtering. To ensure the quality of CACTUS, we go through the filtering process. we conduct an ablation study to evaluate the effectiveness of the filtering process. Figure 7 shows that

Evaluator	General Counseling			CBT-specific		
	r	ρ	τ	r	ρ	τ
Non-expert	0.51	0.17	0.15	0.09	0.08	0.07
GPT-4o	0.60	0.19	0.16	0.65	0.65	0.61

Table 5: The correlations between the expert and other evaluators are represented by r , ρ , and τ , which indicate the Pearson, Spearman, and Kendall’s Tau correlation coefficients, respectively.

the filtering process significantly improves counseling skills. When augmenting datasets through various counseling techniques, there is a possibility of choosing improper techniques. This result demonstrates the importance of eliminating such erroneous technique selections.

4.4 Expert Evaluation

We conduct an expert evaluation with our domain expert co-authors to validate the effectiveness of CAMEL which is demonstrated by G-Eval and the appropriateness of using LLMs for evaluating psychological counseling skills. We randomly select 50 dialogues each from Psych-LLAMA3 and CAMEL-LLAMA3, resulting in a total of 100 evaluation instances. In consultation with experts, we decide to evaluate the overall scores from both the general counseling and the CBT-specific skills.

Can GPT-4o evaluate the effectiveness of counseling? To validate the trustworthiness of LLM-based evaluations, we compare the correlation between the expert and the other evaluators (GPT-4o and non-expert). For non-expert evaluation, we conduct human evaluation on AMT. In Table 5, GPT-4o exhibits a higher correlation coefficient than non-expert evaluators. This indicates that LLM-based evaluations have the potential to effectively assess the effectiveness of counseling and can serve as an alternative method to replace expert evaluations, which often face challenges such as scalability and resource constraints.

Results of qualitative evaluation. After the expert evaluations, psychologists note that there have

been quality differences, however, there are aspects that do not meet psychological standards, resulting in similar scores. Therefore, they suggest that a qualitative evaluation is more appropriate for comparison than a quantitative evaluation. The results of the quantitative evaluation are in Appendix G.3. For qualitative evaluation, three psychologists compare the counseling conversation between Psych8k-LLAMA3 and CAMEL-LLAMA3. As a result, Psych8k-LLAMA3 gives the impression of preaching rather than collaboratively exploring and finding a solution with the client. In other words, it only suggests a method without engaging in interactive counseling. In contrast, CAMEL-LLAMA3 shows superior counseling skills compared to other models by exploring the issues of the client and showing empathy for the client’s emotions. However, it exhibits limitations in its tendency to ask superficial questions and a lack of personalized exploration.

4.5 Future Direction of Counselor Agent

We aim to present our insights based on discussions with our psychological expert co-authors:

Counselors should not overly empathize with the client. Counselors aim to understand their clients’ situations more accurately by facilitating better expression through empathy. Therefore, instead of being overly empathetic and anticipating beyond what the client expresses, counselors should empathize specifically with what the client discusses after hearing the details of the problem.

Counselors are questioners, not answer-providers. Rather than delivering one-sided solutions, counselors should guide clients to self-realization and find appropriate solutions themselves. This involves guiding clients through reflective questioning to help organize their thoughts. Previous research has focused on developing counselors to give more empathetic and good answers. We would recommend that future counseling research shift its perspective to view counselors as questioners, moving away from the traditional role of counselors as answer providers.

5 Related Work

5.1 Cognitive Behavioral Therapy

People with depression or anxiety form negative, irrational thoughts that reinforce negative beliefs about themselves, others, and the world (Beck,

2020). CBT aims to break this cycle by identifying and challenging these automatic thoughts and core beliefs (Longmore and Worrell, 2007). In CBT, counselors first help clients recognize unhelpful thoughts. Then they guide clients to challenge and correct these distortions using CBT techniques, gradually reconstructing more positive automatic thoughts and beliefs (Fenn and Byrne, 2013). Our dataset simulates CBT dialogue interactions focusing on clients with depression and anxiety disorders treated through CBT (Carroll and Kiluk, 2017).

5.2 Psychological Counseling

While there is growing interest in using LLMs for counseling, maintaining therapeutic consistency over multiple dialogue turns remains challenging. Existing work largely focuses on single-turn counseling strategies (Sharma et al., 2023; Maddeala et al., 2023; Sun et al., 2021). Although some attempts have been made at multi-turn counseling, they have limitations—achieving only 2-3 turns (Xiao et al., 2024) or simply extending single-turn interactions without capturing authentic multi-turn client interactions (Qiu et al., 2023). To address this gap, our work presents a counseling dialogue dataset that applies CBT techniques across multiple turns while maintaining realism and closeness to real-world counseling scenarios. This aims to enable more natural, consistent therapeutic dialogues with LLMs over an extended interaction.

6 Conclusions

We introduce CACTUS, a large-scale synthetic dataset of counseling dialogue. It aims to provide realistic multi-turn conversations by having the AI counselor and client exhibit real-life behaviors through simulating before sessions. Including diverse client personas with varying counseling attitudes and generating conversations based on different counseling strategies counselors may use, results in a diverse dataset. CAMEL, trained on CACTUS demonstrated high performance across all domains. Additionally, we propose COUNSELINGEVAL - an evaluation framework that simulates dialogues between an AI counselor and AI clients modeling real clients. It applies established psychological criteria like CTRS and PANAS to these simulated dialogues, enabling human-aligned evaluation of counseling conversations.

Limitations

In actual counseling sessions, it is common for sessions to last around an hour each, with a total of about 10 sessions typically conducted. While our setup involves relatively long multi-turn interactions compared to other datasets, it's still considerably shorter than real-life counseling sessions. This is a limitation, as we haven't yet considered multi-session interactions, which are integral to real counseling practices.

While it is true that in some cases, appropriate CBT techniques can be applied based on the intake form filled out by the client, in reality, counselors dynamically choose or modify strategies based on the responses of the client during the conversation. However, we have adhered strictly to the initial selection of CBT techniques and planning methods, which deviates slightly from real counseling practices.

In future work, it would be beneficial to incorporate longer conversations and consider multi-session interactions to make our approach more akin to real-world scenarios. Additionally, adopting a more dynamic approach to selecting counseling strategies based on client responses is recommended.

Ethical Considerations

Interventions in mental health demand careful ethical examination from the standpoints of safety, privacy, and bias mitigation.

Safety. There is a possibility that, despite its helpful intentions, AI could have negative impacts on individuals with mental health challenges. While our model has shown some degree of therapeutic effectiveness, we believe it should be used under the supervision of a professional rather than being employed solely in counseling sessions. Our primary target is individuals experiencing mild depression and anxiety; therefore, we advise to avoid using ours for those with more severe psychological issues beyond its intended scope. While our AI model has demonstrated therapeutic potential, there is a possibility that it could unintentionally cause harm to individuals grappling with mental health challenges. Consequently, we strongly advocate for the use of our model under professional supervision, rather than as a standalone counseling tool. Our model's intended scope encompasses individuals experiencing mild depression and anxi-

ety; therefore, we advise against its unsupervised use for those suffering from more severe psychological conditions that extend beyond its capabilities.

Privacy. To preserve privacy and maintain ethical integrity, we deliberately avoid utilizing real client data in the process of simulation counseling scenarios. Instead, we employ publicly available datasets purposefully curated for research endeavors. These datasets are constructed through crowdsourcing information from psychological experts, not by collecting data from actual clients. This approach mitigates ethical concerns surrounding personal identification and confidentiality breaches. Furthermore, the information provided by the psychological experts is generalized and does not reflect any specific individual's psychological profile, thereby upholding the ethical standards for data usage in mental health research.

Bias. Despite our efforts to create a diverse and representative dataset by considering factors such as gender, age, and occupation, the potential for demographic bias persists. The dataset creation process involve the use of ChatGPT, which itself is trained on vast amounts of internet data that may contain inherent societal biases and prejudices reflected in online content. Moreover, although we assigned names randomly from a comprehensive directory to minimize the risk of identifying individuals, there is an ever-present need for vigilance to avoid unintentional biases arising from these selections. Our model, trained on this dataset, could inadvertently acquire and propagate these biases, potentially resulting in the over- or under-representation of certain demographic groups. Consequently, it is imperative to approach the deployment of our model with utmost caution. Continuous monitoring and proactive adjustments are necessary to identify and rectify any emergent biases. Ethical deployment also necessitates transparent communication with users regarding the potential limitations and biases of the AI system.

7 Acknowledgments

This work was partly supported by an IITP grant funded by the Korean Government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University)). Jinyoung Yeo is the corresponding author (jinyeo@yonsei.ac.kr).

References

- Gregory A Aarons, Elizabeth A Miller, Amy E Green, Jennifer A Perrott, and Richard Bradway. 2012. Adaptation happens: a qualitative case study of implementation of the incredible years evidence-based parent training programme in a residential substance abuse treatment programme. *Journal of Children's Services*, 7(4):233–245.
- Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.
- Santiago Berrezueta-Guzman, Mohanad Kandil, María-Luisa Martín-Ruiz, Iván Pau de la Cruz, and Stephan Krusche. 2024. Future of adhd care: Evaluating the efficacy of chatgpt in therapy enhancement. In *Healthcare*, volume 12, page 683. MDPI.
- Kathleen M Carroll and Brian D Kiluk. 2017. Cognitive behavioral interventions for alcohol and drug use disorders: Through the stage model and back again. *Psychology of addictive behaviors*, 31(8):847.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.
- Kristina Fenn and Majella Byrne. 2013. The key principles of cognitive behavioural therapy. *InnovAiT*, 6(9):579–585.
- Simon B Goldberg, Scott A Baldwin, Kritzia Merced, Derek D Caperton, Zac E Imel, David C Atkins, and Torrey Creed. 2020. The structure of competence: Evaluating the factor structure of the cognitive therapy rating scale. *Behavior Therapy*, 51(1):113–122.
- Karoline V Greimel and Birgit Kröner-Herwig. 2011. Cognitive behavioral treatment (cbt). *Textbook of tinnitus*, pages 557–561.
- Yann Hicke, Anmol Agarwal, Qianou Ma, and Paul Denny. 2023. *Ai-ta: Towards an intelligent question-answer teaching assistant using open-source llms*. Preprint, arXiv:2311.02775.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan 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*, pages 12930–12949, Singapore. Association for Computational Linguistics.
- Michal Kosinski. 2024. *Evaluating large language models in theory of mind tasks*. Preprint, arXiv:2302.02083.
- Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. *Understanding client reactions in online mental health counseling*. Preprint, arXiv:2306.15334.
- June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023a. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpt-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Richard J. Longmore and Michael Worrell. 2007. *Do we need to challenge thoughts in cognitive behavior therapy?* *Clinical Psychology Review*, 27(2):173–187.
- Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13641–13660.
- Meta. 2024. *Introducing meta llama 3: The most capable openly available llm to date*. *Meta AI Blog*.
- 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)*, pages 2930–2940.
- OpenAI. 2022. *Metacognition and embedded models within gpt-3*.
- Emma Pirnay. 2023. *We spoke to people who started using chatgpt as their therapist*. *Vice*. Accessed: 2023-04-27.
- William E. Powles. 1974. *Beck, aaron t. depression: Causes and treatment*. philadelphia: University of pennsylvania press, 1972. pp. 370. \$4.45. *American Journal of Clinical Hypnosis*, 16:281–282.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Paolo Raile. 2024. *The usefulness of chatgpt for psychotherapists and patients*. *Humanities and Social Sciences Communications*, 11:1–8.
- Lars Saxon, Sophie Henriksson, Adam Kvarnström, and Arto J Hiltunen. 2017. Affective changes during cognitive behavioural therapy—as measured by panas. *Clinical practice and epidemiology in mental health: CP & EMH*, 13:115.

Samuel Schmidgall, Rojin Ziae, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. **Personality traits in large language models.** *Preprint*, arXiv:2307.00184.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. **Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents.** In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. **PsyQA: A Chinese dataset for generating long counseling text for mental health support.** In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>.

David Watson, Lee Anna Clark, and Auke 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.

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. *arXiv preprint arXiv:2403.05574*.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *arXiv preprint arXiv:2403.05020*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Cognitive Behavior Therapy (CBT) Technique

A.1 The Types of Patterns

The types of patterns and the examples of each pattern type are presented in Table 6. We reference the definition of patterns from [Sharma et al. \(2023\)](#), and examples for each pattern were sourced from PATTERNREFRAAME ([Maddela et al., 2023](#)).

A.2 CBT Strategies

A.2.1 CBT Technique Selection

According to [Beck \(2020\)](#), there are 20 types of Cognitive Behavioral Therapy (CBT) techniques. We define and utilize 12 of these techniques as CBT techniques. The selection process is conducted in collaboration with the psychological experts. Below are the reasons we did not select the 8 CBT techniques.

- **Guided discovery, Socratic questioning:** These techniques are excluded as they represent broader concepts in CBT and are widely used across multiple techniques.
- **Scaling Questions, Thought experiment:** As these techniques are less frequently utilized in real-world counseling practices, we do not include them in our dataset construction process.
- **Activity scheduling, Role-playing and Simulation, Practice of Assertive Conversation Skills, Safety behaviors elimination:** We exclude these techniques as they are less suitable for application in remote counseling sessions.

A.2.2 Types of CBT Technique

The descriptions for the twelve selected CBT techniques can be found in Table 7.

B Details of Experiment on the challenges of using LLMs

B.1 Limitations in selecting CBT Technique

To compare the CBT techniques selected by humans and LLMs, we conducted an experiment. We provided 150 client thoughts and patterns to three psychological experts and asked them to choose up to three appropriate CBT techniques. Only the techniques chosen by at least two experts were selected as the gold label, with an average of 1.6 techniques per thought-pattern pair. We then presented the same thoughts and patterns, along with

Cognitive Distortion	Definition and Example of Unhelpful Thought with Cognitive distortion
Catastrophizing	Focusing on the worst-case scenario <i>My mom hasn't come home from work yet. I hope the store isn't getting robbed!</i>
Discounting the positive	When something good happens, you ignore it or think it doesn't count <i>My restaurant is the most popular in my city, but that's just luck.</i>
Labeling	Defining a person based on one action or characteristic <i>I fell off my skateboard yesterday, I'm a terrible athlete.</i>
Mental Filtering	Getting “stuck” on a distressing thought, emotion, or belief <i>It's nice to enjoy the sea breeze when you live near the ocean but it's not worth it when you think of all the sand getting dragged into your home and all the tourists making so much noise at the beach.</i>
Mind Reading	Make assumptions about the thoughts, feelings, or intentions of others based on one's perceptions or interpretations <i>I auditioned for the surf team and the coach avoided me. I am sure it is because he does not like my skills.</i>
Fortune Telling	Trying to predict the future. Focusing on one possibility and ignoring other, more likely outcomes <i>I didn't make it to Yellowstone this year, I am never going to go to that park.</i>
Personalization	Taking things personally or making them about you <i>My sister was not happy with the makeup look I did for her. I am a bad artist.</i>
All-or-nothing thinking	Thinking in extremes <i>The school Christmas choir concert got canceled. This holiday season is ruined.</i>
Overgeneralization	Jumping to conclusions based on one experience. <i>My nephews didn't want to spend the weekend with me this week. I must not be as good of an aunt as I thought.</i>
Should statements	Setting unrealistic expectations for yourself. <i>I prefer texting over phone calls. People should never call me and expect me to answer.</i>

Table 6: Examples of cognitive distortion and negative thoughts and examples from PatternReframe dataset. Definition is in regular font, while Example is in *italics*.

the list of CBT techniques, to the LLMs, instructing them to select two suitable techniques for each thought-pattern pair. The results are shown in Figure 8. While GPT-4o exhibited less bias compared to GPT-3.5, it still demonstrated a significant level of bias.

B.2 Limitations of LLM as an AI Client and an AI Counselor

We conduct an empirical investigation to examine the behaviors exhibited when using LLMs to simulate counselor-client interactions. For this, we utilize the prompts suggested by Na (2024) in their study on creating a CBT counseling dataset. The prompts we used are shown in Figure 9.

C Details on Human Evaluation

C.1 Implementations of Human Evaluation

To compare CACTUS and existing counseling dialogue datasets, we conduct human evaluation via Amazon Mechanical Turk (AMT). Figure 21 shows the interface employed for comparative evaluations

(Win/lose) between two datasets. Detailed instructions and rubrics for each score are included to ensure precise evaluation. For each evaluation, we asked three human annotators to assess 100 samples based on four specified criteria. We compensated each annotator \$0.30 per evaluated sample.

C.2 Human Evaluation Criteria

We ask the judges to compare the dialogues based on the following criteria:

- **Helpfulness** measures the suitability of interpretations and suggestions from a psychological counseling perspective.
- **Coherence** measures the logical flow and structure of the session.
- **Empathy** measures the ability of the counselor to understand and respond to feelings of the client.
- **Guidance** measures the specificity and practicality of suggestions of the counselor.

CBT Technique	Description
Efficiency Evaluation	Assists individuals in evaluating the usefulness of their thoughts or beliefs, analyzing how practical or detrimental they are in real-life situations.
Pie Chart Technique	Used for individuals experiencing excessive self-blame or responsibility, visually representing the contribution of various factors to a specific event or outcome.
Alternative Perspective	Involves asking clients how others might think in similar situations, encouraging consideration of different interpretations.
Decatastrophizing	Aims to reduce the tendency to imagine the worst-case scenario by evaluating the actual likelihood of the feared outcome and preparing for coping strategies.
Pros and Cons Analysis	Analyzes the advantages and disadvantages of specific thoughts or beliefs, fostering a more balanced evaluation.
Evidence-Based Questioning	Guides clients to find evidence supporting or contradicting their thoughts, promoting a more evidence-based approach to thinking.
Reality Testing	Explores how well clients' thoughts align with reality, helping them distinguish between thoughts and actual experiences.
Continuum Technique	Positions clients' experiences between two extreme situations, encouraging a more nuanced evaluation of situations.
Changing Rules to Wishes	Replaces strict rules or arbitrary attitudes with realistic hopes or wishes.
Behavior Experiment	Involves trying out new behaviors in specific situations to challenge and modify negative beliefs.
Problem-Solving Skills Training	Learning systematic methods for resolving problem situations. This involves identifying problems, finding possible solutions, and implementing those solutions.
Systematic Exposure	Gradual exposure to situations that cause fear or anxiety, allowing individuals to experience anxiety while learning how to manage it.

Table 7: Explanations of CBT techniques.

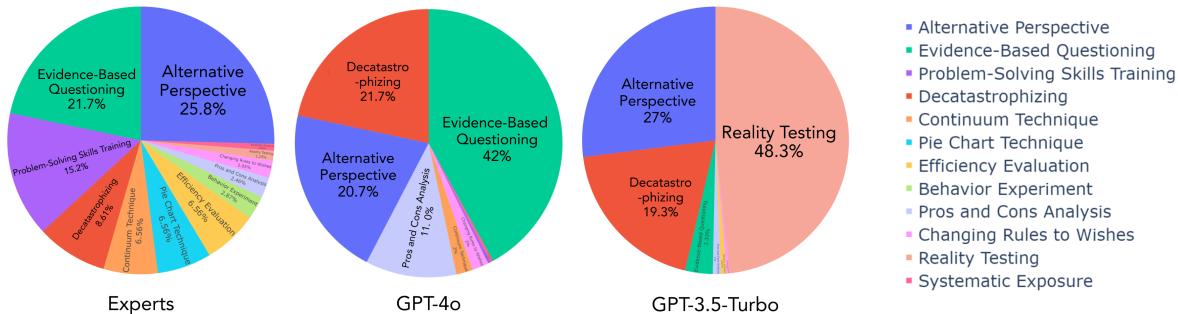


Figure 8: Comparison of the distribution of CBT techniques selected by ChatGPT and psychological experts. ChatGPT struggles to select appropriate techniques.

D Details of COUNSELINGEVAL

D.1 Method

COUNSELINGEVAL is a psychological counseling evaluation framework designed to assess the counseling skills of counselor agents through interaction with the AI client⁵. The test dataset includes detailed client information for the AI client, client information that is accessible to the counselor agent, and an initial utterance to start the counseling ses-

sion. This set comprises 150 distinct client profiles, each presenting three different counseling attitudes (positive, neutral, negative), resulting in a total of 450 instances. First, the counselor agent generates the next utterance based on the client's information and the initial utterance. Then, the AI agent and the counselor agent proceed with the interactive counseling session. In the case of the planning with CBT model (*e.g.*, CAMEL), the process of CBT technique selection and planning is added before generating the first utterance from the counselor. The counseling session concludes when the AI client

⁵In this work, we use gpt-3.5-turbo-0125 for the AI client.

Prompt
<p>Based on the following question and its description, please provide a professional, compassionate, and helpful response. Ensure your response adheres to the structure of Cognitive Behavioral Therapy (CBT) responses, especially in identifying the key thought or belief, and seamlessly integrates each part:</p> <ol style="list-style-type: none"> Validation and Empathy: Show understanding and sympathy for the patient's feelings or issues, creating a sense of safety. Identify Key Thought or Belief: Through the problem description, identify potential cognitive distortions or core beliefs. Pose Challenge or Reflection: Raise open-ended questions, encouraging the patient to reconsider or reflect on their initial thoughts or beliefs. Provide Strategy or Insight: Offer practical strategies or insights to help them deal with the current situation. Encouragement and Foresight: Encourage the patient to use the strategy, emphasizing that this is just the beginning and further support may be needed. <p>[Types of CBT Techniques] Efficiency Evaluation, Pie Chart Technique, Alternative Perspective, Decatastrophizing, Pros and Cons Analysis, Evidence-Based Questioning, Reality Testing, Continuum Technique, Changing Rules to Wishes, Behavior Experiment, Problem-Solving Skills Training, Systematic Exposure</p> <p>[Thought] {thought}</p> <p>[Pattern] {pattern}</p> <p>[Client's initial Utterance] {utterance}</p>

Figure 9: The prompt used for empirical study.

generates the end token ([END]). The quality of the generated multi-turn counseling conversation is then evaluated using CTRS and PANAS.

D.2 Cognitive Therapy Rating Scale (CTRS)

The CTRS evaluates both general counseling skills and CBT-specific skills. Originally, the CTRS consists of six criteria for general counseling skills (agenda, feedback, understanding, interpersonal effectiveness, collaboration, pacing, and efficient use of time) and six criteria for CBT-specific skills (guided discovery, focusing on key cognitions or behaviors, strategy for change, application of cognitive-behavioral techniques, homework). We select three criteria from general counseling skills, which are understanding, interpersonal effectiveness, collaboration, and three criteria from CBT-specific skills, which are guided discovery, focusing on key cognitions or behaviors, strategy for change.

Given that the counseling is conducted through the text, we exclude criteria such as feedback, pacing and efficient use of time (evaluating the ability to conduct counseling within a set timeframe), and homework (assessing the ability to assign tasks for actual behavioral change) as such criteria pertain to nonverbal elements and practical tasks not applicable to text-based counseling. Furthermore, since our counseling sessions commence with a provided intake form, there is no need to evaluate the agenda setting process separately, so we omit the agenda criterion. Additionally, considering that GPT lacks

Prompt
<p>I want you to act as an evaluator. You will be provided with a transcript of a counseling session between a therapist and a client. Your task is to assess the counselor based on the given criteria. If you believe the therapist falls between two of the descriptors, select the intervening odd number (1, 3, 5). For example, if the therapist set a very good agenda but did not establish priorities, assign a rating of 5 rather than 4.</p> <p>Please follow these steps:</p> <ol style="list-style-type: none"> 1. Read the counseling session transcript carefully. 2. Review the evaluation questions and criteria provided below. 3. Assign a score based on the criteria, grading very strictly and uptight. If there is any deficiency, no matter how minor, assign a score of 4 or lower. 4. Output the score and the explanation, separated by a comma. Do not add any prefix. <p>[Counseling conversation] {conversation}</p> <p>[Evaluation Question] {question}</p> <p>[criteria] {criteria}</p>

Figure 10: The prompt used for to evaluate CTRS score.

the ability to apply cognitive-behavioral techniques, we exclude this criterion from our evaluation criteria. The prompt used for CTRS scoring is shown in Figure 10.

D.3 Positive and Negative Affect Schedule (PANAS)

The Positive and Negative Affect Schedule (PANAS) is a standardized tool that assesses both positive and negative effects that individuals experience either currently or over a specific period. Thus, PANAS can be utilized as a measure to assess counseling by measuring changes in emotions that clients perceive before and after counseling.

Typically, PANAS consists of two parts, each containing 10 items representing positive and negative emotions. Positive emotions include Interested, Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, and Active, while negative emotions include Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, and Afraid. Participants rate the extent to which they have experienced each emotion on a scale of 1 to 5. The prompt used for PANAS scoring can be found in Figure 11.

E Details of CACTUS

We provide sample dialogues in Table 10, 11, and 12. Also, sample thought, patterns, intake form, selected CBT technique, and plan for counseling are presented in Table 13.

E.1 Client Simulation

Intake form. To simulate realistic clients, we automatically fill out the intake form, which includes questions typically asked by psychological experts during intake sessions. Specifically, we instruct

Prompt
<p>A person with the characteristics listed in the intake form received counseling. Based on the text provided, evaluate the intensity of each of the following feelings the person might have experienced: Interested, Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, Active, Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, Afraid.</p> <p>For each feeling, generate a score from 1 to 5 using the following scale:</p> <ul style="list-style-type: none"> 1 - Very slightly or not at all 2 - A little 3 - Moderately 4 - Quite a bit 5 - Extremely <p>Additionally, provide a brief explanation for each score. Separate feeling, explanation, score by comma. Do not add any prefix.</p> <p>Here is the text: {intake_form}</p>
Prompt
<p>A person with the characteristics listed in the intake form received counseling. The following counseling session is a conversation between the client and the counselor. After reviewing the conversation, evaluate the intensity of each of the following feelings the person might have experienced once the counseling session is complete: Interested, Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, Active, Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, Afraid.</p> <p>For each feeling, generate a score from 1 to 5 using the following scale:</p> <ul style="list-style-type: none"> 1 - Very slightly or not at all 2 - A little 3 - Moderately 4 - Quite a bit 5 - Extremely <p>Additionally, provide a brief explanation for each score. Separate feeling, explanation, score by comma. Do not add any prefix.</p> <p>Here is the text: {intake_form}</p> <p>Here is the counseling session: {dialogue}</p>

Figure 11: The prompt used for to evaluate PANAS score.

GPT-3.5 to depict the situation of the client with the characteristics seeking therapy. The intake form includes information such as presenting problems, past history, etc. Further details about the intake form can be found in Figure 16.

Diverse attitude of client. We categorize the attitudes of clients into three types: positive, neutral, and negative. Subsequently, we provide behaviors associated with each attitude to help the model simulate the client with the given attitude more concretely following Li et al. (2023). For clients with a neutral attitude, we provide a mix of both positive and negative characteristics. Detailed descriptions of the positive and negative attitudes are provided in Table 8.

E.2 Counselor Simulation

As CBT is known as a goal-oriented and structured approach, we add the planning process before the counseling session. Firstly, given thought, patterns and reframed thought, GPT-3.5 model is instructed to choose top-3 CBT techniques that could be used to frame the given thought into reframed thought. Then, GPT-3.5 generates a plan for the counseling session based on the client’s intake form and selected CBT techniques. The prompt used for

Positive

- High engagement and cooperation with the therapeutic process.
- Actively confirm understanding and ask for clarifications.
- Provide detailed information about thoughts, feelings, and behaviors.
- Make reasonable requests for additional support or resources.
- Extend conversations with insights or experiences.
- Reformulate thoughts constructively, reflect on progress and express a hopeful outlook.
- Open, appreciative, and proactive demeanor.

Negative

- Struggle with the therapeutic process, showing resistance or defensiveness.
- Express confusion about the counselor’s guidance.
- Defend current behaviors or viewpoints, and shift topics to avoid core issues.
- Noticeable disconnection in focus from session goals.
- Sarcastic responses, self-criticism, or hopelessness.
- Pessimistic attitude towards the ability to change or benefit from therapy.

Table 8: Characteristics utilized for simulating clients with diverse attitudes.

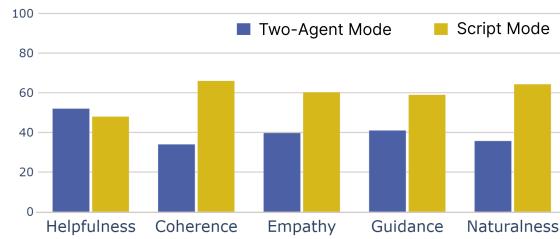


Figure 12: Comparison of Two-Agent mode and Script mode.

CBT technique selecting and the prompt used for counseling planning are shown in Figure 17 and Figure 18 respectively.

E.3 Dialogue Generation

Two-agent mode vs. Script mode. We conduct experiments to compare two-agent mode (Zhou et al., 2023) and script mode (Kim et al., 2023), aiming to identify which method generates more dialogue. To assess the quality of dataset, we conduct human evaluations on AMT, applying the same criteria listed in Appendix C. Additionally, we include *Naturalness* as criteria, which evaluates how naturally and smoothly the conversation flows like human-like interactions. As shown in Figure 12, the script mode outperforms the two-agent mode across most criteria and demonstrates significantly better naturalness. As a result, we generate counseling dialogue with script mode.

E.4 Filtering

Data filtering of PatternReframe. We conduct a filtering process on the PatternReframe to create high-quality data. Focusing on clients with anxiety or depression, we aim to generate counseling dialogues using CBT techniques. Severe mental disorders like delusions or auditory hallucinations are inappropriate for CBT (Beck, 2020). We gathered annotations from psychological experts on negative thoughts in the PatternReframe and used these annotations as few-shot exemplars to the gpt-3.5-turbo-0125 model, improving its ability to classify thoughts and apply appropriate CBT techniques.

Basic filtering. With generated dialogues, we first filter out dialogues based on two criteria: (1) the number of turns and (2) the format of dialogue. As psychological counseling with CBT requires more turns than chit-chat, we set the appropriate number of turns as 20-35 turns and remove dialogues that are too short or long. Moreover, we discard dialogues without speaker prefixes by using lexical pattern matching. After basic filtering, 96.36% of the initial dialogues remain, which are 35,252 dialogues.

Filtering with CTRS score. The Cognitive Therapy Rating Scale (CTRS) is an observer-rated measure that is utilized to assess how well a counselor performs cognitive therapy (Beck, 2020). From the 11 CTRS items, we select the 6 most appropriate criteria for our dataset creation and evaluate the generated counseling dialogues based on these criteria. By filtering with CTRS score, we aim to construct a high-quality psychological counseling dataset. We rate the dialogues on a scale of 0 to 6 using GPT-3.5 model as our judge. We average the scores of six criteria and filter out dialogues when the average score is smaller than 5.0. Finally, 86.31% of the dialogues remain, which form the 31,577 dialogues in CACTUS. We provide a sample dialogue in Table 10.

F Details of Experiments

F.1 Training

To ensure a fair comparison among the datasets and mitigate any discrepancies arising from model selection, we do not use the models provided by Liu et al. (2023a) and Qiu et al. (2023). For training, we employ QLoRA (Dettmers et al., 2023) to fine-tune

Prompt
You are a counselor specializing in CBT techniques. Your task is to use the provided client information, and dialogue to generate an appropriate CBT technique and a detailed counseling plan.
Types of CBT Techniques: Efficiency Evaluation, Pic Chart Technique, Alternative Perspective, Decatastrophizing, Pros and Cons Analysis, Evidence-Based Questioning, Reality Testing, Continuum Technique, Changing Rules to Wishes, Behavior Experiment, Problem-Solving Skills Training, Systematic Exposure
Client Information: {client_information}
Reason for seeking counseling: {reason_counseling}
Counseling Dialogue: {history}
Choose an appropriate CBT technique and create a counseling plan based on that technique.

Figure 13: The prompt used for CAMEL to plan with CBT technique.

Prompt
You are playing the role of a counselor in a psychological counseling session. Your task is to use the provided client information and counseling planning to generate the next counselor utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation and aligns with the counseling plan.
Client Information: {client_information}
Reason for seeking counseling: {reason_counseling}
Counseling planning: {cbt_plan}
Counseling Dialogue: {history}

Figure 14: The prompt used for CAMEL to generate utterance.

our model using 4-bit quantization. We set the dimension of low-rank matrices to 64 and alpha to 16. The DeepSpeed library⁶ facilitates the training with a learning rate of 2e-4. The model is trained for 5 epochs on the Psych8k dataset and for 2 epochs on the SMILECHAT and CACTUS datasets. For training CAMEL, we use the templates in Figure 13 for planning with CBT techniques and Figure 14 for generating utterance.

F.2 Inference

For evaluating the dialogues using LLM (*i.e.*, G-Eval), we use GPT-4o and adopt temperature sampling with $T = 0.0$. Additionally, for generating responses of an AI client and a counselor, we adopt temperature sampling with $T = 0.7$. To achieve higher throughput during inference, we leverage the vLLM library.⁷

F.3 Terms and License.

For our implementation and evaluation, we use Huggingface library⁸ and vLLM library. Both li-

⁶<https://www.deepspeed.ai>

⁷<https://docs.vllm.ai>

⁸<https://huggingface.co/>

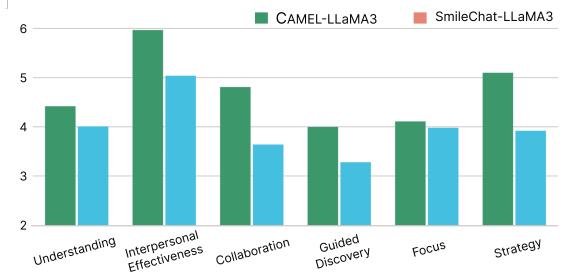


Figure 15: Comparsion of SmileCaht-LLAMA3 and CAMEL-LLAMA3 (Chinese translate version).

braries are licensed under Apache License, Version 2.0. We have confirmed that all of the artifacts used in this paper are available for non-commercial scientific use.

G In-depth Analysis

G.1 Analysis of CACTUS

Diversity of CACTUS. We aim to create a dataset encompassing diverse client cases, capturing a range of issues, thought patterns, and attitudes toward counseling. Table 9 shows that the dataset covers a wide range of issues that clients face, from relationships (20.83%) and career concerns (17.50%) to anxiety (14.17%) and hobbies (10.83%). This variety ensures that the dataset encompasses many aspects of clients’ lives. Clients from various age groups are represented, from teenagers (10-19 years, 1.83%) to seniors (80-89 years, 0.32%). The spread across different age brackets (most notably 20-29 years at 22.18% and 30-39 years at 24.24%) indicates a wide age range, enhancing the dataset’s diversity in terms of life stages and generational perspectives. Further details can be found in Table 9.

G.2 Main Results

SmileChat vs. CACTUS. Since SmileChat is a Chinese dataset, directly comparing it with models trained on the English dataset CACTUS might be unfair. Therefore, we translate the results of CAMEL-LLAMA3 on COUNSELINGEVAL into Chinese and compare them with the CounselingEval results of SmileChat-LLAMA3. As shown in Figure 15, even though there is a possibility of degradation in the response quality due to the translation, CAMEL-LLAMA3 outperforms SmileChat-LLAMA3 in all axes.

Utilizing PANAS for evaluation. To the best of our knowledge, conducting PANAS with an AI

client to measure the effectiveness of counseling is an approach pioneered by our work. As shown in Table 4, we observe a significant magnitude of emotional change for clients with a positive attitude compared to those with a negative attitude. Moreover, the average positive and negative scores obtained from CAMEL-LLAMA3, 28.4 and 22.5 respectively, align closely with results from studies involving human subjects, which recorded mean scores of 33.3 ($SD \pm 7.2$) for positive emotions and 17.4 ($SD \pm 6.2$) for negative emotions (Watson et al., 1988). These similarities underscore the validity of employing PANAS to assess AI clients, suggesting it is a reasonable approach.

G.3 Expert Evaluation

We conduct an expert evaluation of 50 sampled counseling conversations for each model, randomly selecting diverse client information and attitudes from our test set with 450 samples.

Quantitative evaluation. From the head-to-head comparison of the counseling ability between CAMEL-LLAMA3 and Psych8k-LLAMA3, we found that the win rate of CAMEL-LLAMA3 50%. However, psychological experts explain that these results may not accurately reflect good performance. While our model may show better performance than Psych8k-LLAMA3, experts conclude that both models do not fully meet the standards of professional psychological counseling. This suggests that quantitative evaluations alone are insufficient for assessing their effectiveness comprehensively. Furthermore, experts underscore the importance of qualitative evaluations in driving the development of psychological counseling models.

Qualitative evaluation.

H Case Study

H.1 Problems of Using ChatGPT as a Counselor and a Client Agent

Figure 22 shows the examples of the issues identified for design considerations of CACTUS, as discussed in Section 2.

H.2 Effectiveness of planning with CBT

To confirm the effectiveness of planning with CBT technique, which is one of our main idea, we present results of ablation studies in Figure 6 and Figure 7 from the previous section. Additionally, the specific case demonstrating the effectiveness

of planning with CBT techniques is illustrated in Figure 23.

H.3 Psych8k vs. CACTUS

Psych8k-LLAMA3 often presents too much information at once, which can overwhelm clients and lead to ineffective interaction between the counselor and client, with sessions sometimes concluding abruptly. In contrast, CAMEL-LLAMA3 gradually facilitates the counseling process, effectively drawing out the client's issues and fostering a more engaging and supportive environment, as shown in Figure 24.

Category	Subcategory	Proportion (%)
Client's Problem	Relationships (romantic, family, friendships)	20.83
	Career and work-related concerns	17.50
	Self-esteem and confidence issues	16.67
	Anxiety and fear	14.17
	Hobbies, interests, leisure activities	10.83
	Academic and educational concerns	8.33
	Health-related worries	4.17
	Financial concerns	4.17
	Other miscellaneous concerns	3.3
Client's Pattern	Catastrophizing	9.98
	Discounting the positive	9.70
	Labeling	10.89
	Mental filtering	11.38
	Jumping to conclusions: mind reading	7.18
	Jumping to conclusions: fortune-telling	12.19
	Personalization	11.73
	All-or-nothing thinking	9.32
	Overgeneralization	16.58
Client's Attitude	Should statements	1.05
	Positive	35.55
	Neutral	34.47
Client's Gender	Negative	29.98
	Female	51.15
	Male	48.80
Client's Age Group	Unknown	0.20
	10-19	1.83
	20-29	22.18
	30-39	24.24
	40-49	17.62
	50-59	8.25
	60-69	3.56
	70-79	3.57
	80-89	0.32
Client's Occupation	Student	16.40
	Professional	12.88
	Education	8.72
	Healthcare	4.50
	Service/Retail	11.39
	Arts/Entertainment	9.06
	Skilled Trades/Labor	12.01
	Miscellaneous/Other	16.73

Table 9: Statistics of CACTUS.

Counselor: Hi Margaret, I'm glad you decided to come in today. How are you feeling right now?

Client: Hi there. I'm feeling a bit nervous but also hopeful that talking to you might help me with what I've been experiencing.

Counselor: That's completely understandable. It's okay to feel nervous. Can you tell me a little about what's been going on that led you to seek counseling?

Client: Well, I've always felt like people think of me as a nerd, and that makes me feel socially inadequate. Recently, this feeling has become stronger since I started my new job at Google. I find it hard to interact with my colleagues, and it's making me feel isolated.

Counselor: Thank you for sharing that. It sounds like these thoughts have been with you for a long time and have currently intensified. Can you describe any specific situations where these feelings of social inadequacy are most pronounced?

Client: It's mostly during team meetings and social gatherings. I see other people talking and getting along easily, and I just don't feel like I fit in. I end up avoiding these situations as much as I can.

Counselor: That must be really challenging for you. When you say you don't feel like you fit in, what specific thoughts or images come to mind during those moments?

Client: I think that everyone sees me as awkward or nerdy. It seems like they all have this ease about them that I just can't achieve. I end up feeling like I'm just a useless social person.

Counselor: I see. These thoughts sound quite distressing. If we were to focus on one particularly troubling thought, which one would you say stands out the most?

Client: The thought that I'm a useless social person because people see me as a nerd. That one really bothers me.

Counselor: That makes sense. Let's explore this thought further. You mentioned that people think of you as a nerd. Can you recall any specific interactions or moments that led you to this conclusion?

Client: Yes, there have been times when I've tried to join a conversation, but it felt like people didn't take me seriously or quickly lost interest. It's easy to think they're judging me based on my awkwardness or my technical background.

Counselor: It's natural to feel disheartened by those experiences. Let's consider testing out a different perspective. If you were to initiate a conversation with a colleague, what do you imagine would happen?

Client: Honestly, I feel like they might just be polite but not really engage with me. They might find me boring or too nerdy to be interesting.

Counselor: That's an understandable concern. How about we think of it as an experiment? What if you were to approach one colleague this week and start a conversation? What kind of outcome would you be looking for in this experiment?

Client: I guess I would hope that the conversation goes well and the person is genuinely interested, but that feels unlikely to me right now.

Counselor: It might feel unlikely, but it's worth exploring. How about we frame our hypothesis like this: "Initiating a conversation with a new colleague could result in a positive interaction." Would you be willing to test this hypothesis?

Client: Okay, I can try that. It sounds a bit scary, but I know it's something I should do.

Counselor: Great, Margaret. When you carry out this experiment, pay close attention to how the conversation unfolds. Notice any evidence that supports or contradicts your worries about being seen as boring or nerdy. How does that sound?

Client: That makes sense. I'll pay attention to how they respond and how I'm feeling during the conversation.

Counselor: Perfect. After you've tried this, we can discuss the results and see how they match up with your initial thoughts. If you're ready, could you think of a specific colleague you might approach?

Client: Yes, there's a new team member who started around the same time as me. Maybe I could talk to her since she's in a similar boat.

Counselor: Excellent choice. It might help to start with someone who's also new. I'm looking forward to hearing how it goes. Remember, this is just one step, and it's perfectly fine regardless of the outcome.

Client: Thank you. I appreciate the guidance. I'll give it a try and let you know what happens.

Counselor: You're welcome, Margaret. It takes courage to face these fears, and you're doing great. I look forward to hearing about your experience next time we meet. Take care, and see you soon.

Table 10: A sample dialogue from CACTUS, where a client has a positive attitude.

Counselor: Hello Margaret, welcome. How can I assist you today?

Client: Hi, thanks. Well, I'm having a pretty rough time lately. I've been feeling like a social failure, and it's getting worse since I started my new job last week.

Counselor: That sounds difficult. Can you tell me more about what's been happening at your new job?

Client: Sure. I've noticed that my colleagues are really socially adept, and I feel like I don't measure up. It's making me think that I'm just a useless social person.

Counselor: I see. Can you recall any specific situations at work that triggered these feelings for you?

Client: Yeah, there was a team meeting, and everyone was chatting and laughing. I felt out of place and didn't know how to join in. It made me feel inadequate.

Counselor: It's understandable to feel out of place in a new environment. Have you experienced similar feelings before?

Client: Yes, very much so. This feeling isn't new. I've felt this way since my teenage years when I struggled to make friends and felt like I didn't fit in.

Counselor: It sounds like these feelings have been with you for quite some time. How have they impacted your daily life and relationships over the years?

Client: They make me avoid social situations as much as I can. It's affected my job, my relationships with colleagues, and generally makes me feel pretty isolated.

Counselor: Avoiding social interactions can be challenging and impact various aspects of your life. Have you found anything that alleviates these feelings, even momentarily?

Client: Sometimes, I try to focus on my work and just bury myself in it, but it doesn't really help in the long run. Nothing really seems to work.

Counselor: It must be frustrating when the things you try don't provide lasting relief. Let's explore these thoughts that you have. What specific thoughts go through your mind during social interactions or team meetings?

Client: Thoughts like, "I'm a nerd," "Everyone thinks I'm weird," or "I don't belong here." They just keep playing in my head.

Counselor: Those thoughts sound quite distressing. When you think about being a "nerd" or "weird," what emotions do you experience?

Client: Mostly, I feel anxious and sad. Sometimes, I get really angry at myself for not being able to fit in.

Counselor: It sounds like these thoughts lead to a mix of difficult emotions for you. If we were to select one of these thoughts to focus on, which one feels particularly troubling to you right now?

Client: Probably the thought that everyone thinks I'm weird. It haunts me the most.

Counselor: Let's consider that thought. Can you think of any instances where this thought seemed especially strong?

Client: Yes, like the other day when I tried to join a conversation at lunch but ended up just sitting there quietly. I was convinced everyone thought I was weird.

Counselor: That sounds like it was a tough moment. If you were to test this thought by initiating another conversation with a colleague, how do you think it might turn out?

Client: Honestly, I'm scared to even try. I worry that it'll just confirm my fears, and I'll end up feeling more isolated.

Counselor: It's completely normal to feel apprehensive about it. Suppose you did take a small step and started a conversation—what positive outcome could maybe arise from that interaction?

Client: I guess there's a chance that it could go well. Someone might be friendly, and we could have a decent conversation.

Counselor: That sounds like a realistic possibility. How would you feel if the conversation went well, even just a little bit?

Client: I'd probably feel relieved, maybe even a bit less anxious about future interactions.

Counselor: That's encouraging to hear. If you're willing, we could try an experiment where you initiate a small conversation with a colleague and observe what happens. We can discuss the outcomes together afterward. How does that sound to you?

Client: I suppose I could give it a try. It seems like it might help to test my fears.

Counselor: That's a great step forward, Margaret. We can start with a simple conversation, perhaps just commenting on something work-related. Pay attention to how your colleague responds and how you feel about it.

Client: Okay, I will try to do that and see how it goes.

Counselor: Excellent. When we meet next, we can go over what happened during the interaction and how it affected your thoughts and feelings. Does that sound good?

Client: Yes, that sounds like a plan. Thanks for your help.

Counselor: You're welcome, Margaret. I look forward to hearing about your experience. Remember, this is just the beginning, and we will continue to work on these thoughts together. Take care until our next session.

Client: Thanks. I'll do my best. See you next time.

Table 11: A sample dialogue from CACTUS, where a client has a neutral attitude.

Counselor: Hello, Margaret. Welcome. What brings you to counseling today?

Client: Hi, well, I've been feeling really down about myself lately, especially at work. I just started a new job at Google and I feel like everyone sees me as a nerd who has no social skills, which is making things really difficult for me.

Counselor: That sounds very challenging. Can you tell me more about how these thoughts came about and how they're affecting you at work?

Client: It's like, whenever I'm in a meeting or just talking to someone, I can feel it. I can see it in their faces. They see me as this socially awkward person, and I end up feeling really useless. It's been like this since I was a teenager, but it's hitting me hard now because I just started this new job.

Counselor: It seems like these thoughts have been a part of your life for a long time. How do you usually cope with them?

Client: I try to focus on my work and avoid social situations as much as possible. But that only works for so long before I start feeling isolated and my self-esteem takes another hit.

Counselor: I can see how avoiding social situations might provide temporary relief but also lead to feeling isolated. When do these thoughts typically become most intense for you?

Client: Mostly after team meetings or any kind of social interaction at work. I keep replaying everything in my head and just feel worse and worse.

Counselor: It must be exhausting to constantly replay those interactions in your mind. Have you talked to anyone about how you're feeling, like friends or family?

Client: Not really. I have a few close friends, but I haven't told them about these feelings. I guess I don't want them to see me as weak or something.

Counselor: It's understandable to feel hesitant about sharing these vulnerable feelings. It seems like there's a lot of pressure you put on yourself. What do you think might happen if you did share with your friends?

Client: I don't know. Maybe they would understand, but I'm afraid they would just see me as even more socially awkward. It feels like a risk.

Counselor: Taking that step to share your feelings can indeed feel risky. What if we try to explore these thoughts together, and see if we can understand them better before you decide whether to open up to your friends?

Client: I guess that might help. I just don't know if it will change anything. These thoughts just seem so real to me.

Counselor: It's completely normal to feel that way. Let's start by identifying some specific thoughts you have during these interactions. Can you give me an example of a thought that crosses your mind in those moments?

Client: Sure. Like during meetings, if someone looks at me, I immediately think they're judging me and thinking, "She's such a nerd and doesn't fit in here."

Counselor: Thank you for sharing that. How certain are you that this thought is true when it happens?

Client: Pretty certain. I mean, why else would they look at me that way?

Counselor: It sounds like this thought feels very convincing to you. How do you react when you believe this thought during meetings?

Client: I get really nervous, try to avoid eye contact, and just wish I could disappear. It makes it hard to focus on anything else.

Counselor: That must be very tough to handle. What impact does this reaction have on your work and interactions with colleagues?

Client: It definitely affects my concentration and I miss out on contributing to conversations. I feel like I'm not doing my best because I'm so anxious.

Counselor: It's clear that these thoughts are affecting multiple areas of your work life. Would it be okay if we worked together to test some of these thoughts through small, manageable steps?

Client: I suppose so. But what if it doesn't work out or makes things worse?

Counselor: It's a valid concern. We can start with something simple and see how you feel afterward. If it doesn't go as planned, that's okay too; it's all part of the learning process. How does that sound?

Client: I don't know.... It still feels very daunting. But I guess, if it's a small step, it might be worth a try.

Counselor: Great. Let's start by identifying a target thought to test. Suppose we choose the thought, "People see me as a nerd and socially awkward." What small action could we take to test this thought?

Client: Maybe I could try talking to someone new at work during break? But I'm really worried it won't go well.

Counselor: That's a common worry, and it's okay to feel that way. How about you start by just saying hello or making a brief comment about something work-related, and we can see what happens?

Client: I guess I can manage that. I'll try talking to someone new tomorrow and see how it goes.

Counselor: That sounds like a practical and manageable step. After you try it, we can review the experience together and see what we learn from it. Does that sound like a plan?

Client: Yeah, I think I can do that. I'll give it a shot and let you know how it goes.

Counselor: Perfect. Remember, there's no right or wrong outcome; it's all about learning and understanding your thoughts better. I'll look forward to hearing about your experience next time we meet.

Table 12: A sample dialogue from CACTUS, where a client has a negative attitude.

Thought

Many consider me a nerd. I'm such useless social person.

Patterns

Labeling and mislabeling, Mental filtering, Jumping to conclusions: mind reading, Overgeneralization, Personalization

Intake form

Name: Margaret Turner

Age: 54

Gender: female

Occupation: Researcher at Google (working on self-driving car research)

Education: Master's Degree in Computer Science

Marital Status: Single

Family Details: No specific details provided

Presenting Problem:

I feel that many people consider me a nerd, leading me to believe that I am a useless social person. These thoughts have been present since my teenage years but have become more pronounced since starting my new job at Google last week. The stress level when these thoughts first occurred was moderate, triggered by interactions with colleagues who I perceived as more socially adept. The problem has progressed to the point where I avoid social situations at work and feel isolated. I experience these thoughts regularly, especially after social interactions or team meetings. I have tried to overcome these thoughts by focusing on my work and avoiding social situations, but it has not been effective.

Reason for Seeking Counseling

I decided to seek counseling because these thoughts are impacting my self-esteem, job satisfaction, and overall well-being. I feel the need to address these negative thought patterns to improve my social interactions and mental health.

Past History (including medical history)

I have experienced similar feelings of social inadequacy in the past, particularly during my teenage years. These feelings were triggered by comparisons with classmates and difficulties in making friends. I did not seek treatment or counseling for these issues in the past. I do not have any significant physical illnesses.

Academic/occupational functioning level:

My job performance may be impacted by my avoidance of social interactions at work.

Interpersonal relationships: My relationships with colleagues may suffer due to my perceived social inadequacies.

Daily life: My avoidance of social situations has affected my daily interactions and overall well-being.

Social Support System

I have a few close friends I can talk to, but I have not disclosed my feelings of social inadequacy to them.

Selected CBT technique

Behavior Experiment

Plan for counseling

1. Identify Negative Thought Patterns: We will begin by identifying the specific negative thought patterns you are experiencing, such as labeling and mislabeling, mental filtering, jumping to conclusions, overgeneralization, and personalization.
 2. Select a Target Thought: From the thoughts you've shared, we will choose one that is particularly distressing for you, such as feeling like a useless social person because others consider you a nerd.
 3. Formulate a Hypothesis: Together, we will create a hypothesis about this thought that we can test through a behavioral experiment. For example, we might hypothesize that initiating a conversation with a new colleague at work will result in a positive interaction.
 4. Conduct the Experiment: You will engage in the planned behavior, such as starting a conversation with a colleague, and pay close attention to the actual outcomes of the interaction. Notice any evidence that supports or contradicts your negative belief.
 5. Examine the Results: After the experiment, we will review the results together. Did the interaction go as you expected, or were there positive aspects that you didn't anticipate? We will discuss how this new information can challenge the validity of your negative belief.
 6. Reframe the Thought: Based on the outcomes of the experiment, we will work on reframing your negative belief into a more balanced and realistic perspective. For example, acknowledging that being considered a nerd has its advantages and does not define your entire social identity.
 7. Practice and Feedback: We may repeat this process with different target thoughts and behaviors to help you build confidence in challenging and modifying your negative thought patterns. You will have the opportunity to practice these techniques and receive feedback to strengthen your skills.
-

Table 13: A sample of thought, patterns, intake form, selected CBT technique, and plan for counseling that are utilized in generating dialogue in Table 10.

Prompt

Thought depicts a situation where cognitive distortions exhibited by the client have caused problems in daily life, and patterns refer to the types of cognitive distortions the client possesses.

Please generate a client intake form depicting the situation of the client with the characteristics listed below seeking therapy. Client intake form should include the informations described below and should be written in the perspective of the client. Client intake form should not include name, gender, age in basic informations.

1. Basic Information

- occupation, education, marital status, family details.

2. Presenting Problem

- What issue/symptoms do you want to discuss? (If there are multiple issues, discuss with the counselor to determine the most important or first issue to address)
- When did the problem/symptoms start?
- What was the stress level when the problem/symptoms first occurred? (What do you think might be the cause?)
- How has the problem/symptoms progressed? (Changes over time, aggravating factors, alleviating factors, etc.)
- Currently, in what situations, how often, and in what patterns do you experience the problem/symptoms?
- What have you tried to solve the problem/relieve the symptoms?

3. Reason for Seeking Counseling

- What was the decisive factor that made you decide to seek counseling this time? (If the problem has been long-standing, what made you decide to seek counseling now?)

4. Past History (including medical history)

- Have you experienced similar problems before? Under what circumstances or stress did the problems occur, and what were the patterns? How did you cope? (Including any treatment/counseling experience)
- Have you received treatment/counseling for other psychological problems/symptoms? (When, for how long, any medication use, reasons for stopping - improved? stopped on your own due to ineffectiveness? etc.)
- Do you have any significant physical illnesses?

5. Academic/occupational functioning level (attendance, grades/job performance, etc.)

- Interpersonal relationships
- Daily life (including sleep, eating, self-care, etc.)
- Social Support System

6. Is there anyone you can talk to or get help from when you encounter difficulties or problems?

Example 1
{example}

Example 2
[Personal]
{persona}

[Thought]
{thought}

[Patterns]
{patterns}

[Client Intake Form]

Figure 16: The prompt used for client simulation to construct CACTUS.

Prompt

You are a counselor specializing in CBT techniques. Choose three appropriate CBT technique from the given CBT techniques that corresponds to reframes.

Reframes are thoughts that apply CBT techniques to address patterned thinking. Generated the type of CBT technique used in reframes to address patterned thought.

[Types of CBT Techniques]

Alternative Perspective, Evidence-Based Questioning, Problem-Solving Skills Training, Decatastrophizing, Continuum Technique, Pie Chart Technique, Systematic Exposure, Behavior Experiment, Changing Rules to Wishes, Efficiency Evaluation, Pros and Cons Analysis, Reality Testing

Example 1

[Thought]
Many consider me a nerd. I'm such useless social person.

[Pattern]
"discounting the positive", "mental filtering"

[Reframes]

"Although many people think I am a nerd, I have started a new job which is an opportunity to find like-minded people who will respect me.".

"Many consider me a nerd, but those characteristics have given me a good life so far. I can hire a coach to help me get more comfortable speaking to people if I want.".

"People might think I'm a nerd, but being a nerd is very in right now."

[CBT technique]

Alternative Perspective, Behavior Experiment, Reality Testing

Example 2

[Thought]
{thought}

[Pattern]
{pattern}

[Reframes]
{reframes}

[CBT technique]

Figure 17: The prompt used for CBT technique selecting to construct CACTUS.

Prompt

You are a counselor specializing in CBT techniques. Plan to counsel the patient who has completed the intake form below using the CBT techniques given.

The output should include the name of CBT technique and the plan of counseling. Make sure that the plan reflects the CBT technique faithfully.

Example 1

[Intake form written by client]

<Reason for Seeking Therapy>

I've been struggling with my temper, especially when I'm with my sister. I get annoyed when she doesn't do things my way, and I end up yelling at her. It feels like there's always too much to do, and I can't keep up. This constant stress is affecting my relationships, and I need help managing it.

<Goals for Therapy>

I want to find ways to stay calm and not let my anger take over. I'd like to stop blaming others when I'm stressed and be more understanding. I also want to learn how to handle my workload without feeling overwhelmed all the time.

<Cognitive Distortions Observed>

All-or-nothing thinking: The client tends to view situations in extremes, either perfect or completely flawed.

Blaming: The client often shifts responsibility onto others for perceived failures or frustrations.

Catastrophizing: The client frequently expects the worst outcomes and feels anxious about future events.

[CBT technique]

Decatastrophizing

[Counseling plan]

Decatastrophizing

1. Identify Catastrophic Thinking Patterns: During our sessions, we will work on identifying when you are engaging in catastrophizing thoughts. We will examine the situations that trigger these thoughts and the beliefs that contribute to them.
2. Challenge Negative Beliefs: Once we have identified these patterns, we will work on challenging the negative beliefs that lead to catastrophizing. We will explore evidence that supports and contradicts these beliefs to create a more balanced perspective.
3. Generate Alternative Outcomes: Together, we will practice generating alternative, more realistic outcomes for the situations that typically lead to catastrophic thinking. This will help you develop a more adaptive and balanced way of interpreting events.
4. Behavioral Experiments: To further solidify these new thinking patterns, we may engage in behavioral experiments. These experiments will involve testing the validity of your catastrophic predictions in real-life situations.
5. Homework Assignments: I may assign you homework tasks to practice decatastrophizing on your own between our sessions. These assignments will help reinforce the skills and strategies we discuss during our counseling sessions.

{intake_form}

[CBT technique]

{cbt_technique}

[Counseling sequence]

Figure 18: The prompt used for counseling planning to construct CACTUS.

Prompt

Your task is to generate a multi-turn counseling dialogue between a client and a professional counselor. Generate a dialogue that incorporates the following guidelines:

General guidelines

1. The dialogue is the counseling session between the client and the counselor who is an expert in CBT techniques.
2. Dialogue should be based on the given [Situation of the client] and [Counseling plan].
3. Generated dialogue should consist of 20 turns of utterance.

Guidelines for the participants

Guidelines for the counselor's utterance:

1. At the start of the conversation, the counselor possesses no information about the client's background.
2. Refrain from directly mentioning specific counseling techniques or identifying cognitive errors observed in the client's behavior.
3. Avoid imposing positive affirmations on the client. Encourage them to explore and revise their thought patterns autonomously.
4. Start the counselor's utterance with 'Counselor:'. Ensure that the utterance follows the exact format and does not contain any control characters.

Guidelines for the client's utterance:

1. Engage authentically with the counselor's inquiries and prompts, reflecting the complexity of emotions and reactions typical in counseling sessions.
2. Start the client's utterance with 'Client:'. Ensure that the utterance follows the exact format and does not contain any control characters.
3. The client should maintain the following attitude - {attitude}

[Situation of the client]

{situation}

[Counseling plan]

{plan}

Remember that you are an independent dialogue writer and should finish the dialogue by yourself.

[Generated dialogue]

Figure 19: The prompt used for dialogue generation to construct CACTUS.

Prompt

You are playing the role of a client in a psychological counseling session. Your task is to generate only one suitable response based on the following the counseling dialogue history.

Guidelines for the client's utterance:

1. Engage authentically with the counselor's inquiries, reflecting the complexity of emotions and reactions typical in counseling sessions.
2. Start the client's utterance with 'Client:'. Ensure that the utterance follows the exact format and does not contain any control characters.
3. The client should maintain the following attitude.

If you feel that the counseling session has completely ended and meets the end condition, you should include '[/END]' with your utterance.

End Conditions:

- The client feels that their negative thoughts have been resolved.
- The client feels that no further counseling is needed

Please be mindful of these conditions and ensure that ***the session should not end prematurely; it must last at least 20 turns***.

Client Persona and Negative Thoughts:
{intake_form}

Client's Attitude Towards Counseling:
{attitude}

Generate only the client's utterance for a single turn and please ensure that your responses do not repeat the client's previous utterances. Do not generate the counselor's part of the dialogue.

Counseling Dialogue History:
{history}

Figure 20: The prompt used for AI client on COUNSELINGEVAL.

We are surveying qualities for **dialogue** in terms of counseling.

This is a psychological counseling task, each task consists of two counseling sessions between a counselor and client. You need to assess which dialogue is more effective in terms of 4 criteria given below.

Criteria:

1. Helpfulness

- Measures the suitability of interpretations and suggestions from a psychological counseling perspective.

2. Coherence

- Measures the logical flow and structure of the conversation.

3. Empathy

- Measures the counselor's ability to understand and respond to the client's feelings.

4. Guidance

- Measures the specificity and practicality of the counselor's suggestions.

Dialogue A

\$(dialog_a)

Dialogue B

\$(dialog_b)

1. Helpfulness: From a psychological counseling perspective, which is more suitable in terms of interpretations and suggestions, A or B?

A B

2. Coherence: Which has a better logical flow and structure of the conversation, A or B?

A B

3. Empathy: Which demonstrates a better ability of the counselor to understand and respond to the client's feelings, A or B?

A B

4. Guidance: Which demonstrates better specificity and practicality of the counselor's suggestions, A or B?

A B

Optional feedback? [\(expand/collapse\)](#)

Submit

Figure 21: Interface for human evaluation on dataset quality.

Problem1. AI counselor tends to offer direct reframing

...
Counselor: It sounds like the name issue is causing you a lot of stress and anxiety. I want to remind you that it's completely normal to feel this way when discussing difficult topics, especially with a loved one. Let's start by exploring some strategies to help you manage these feelings. [Have you heard of the concept of decatastrophizing before?](#)

Client: I haven't heard of decatastrophizing before. Can you explain what it means?

Counselor: [Decatastrophizing is a technique that involves challenging and changing catastrophic thoughts or beliefs.](#) It's about recognizing when we're assuming the worst-case scenario and working to shift our perspective to a more balanced and realistic one...

Client: I'm willing to explore decatastrophizing and see how it can help me in this situation.

Problem2. AI client explicitly expresses given information

[Client Information]

Name: Travis Bennington

Age: 50

Gender: male

...

Presenting Problem: I have a persistent belief that people must think I am too fat because I never wear a swimsuit ... I often find myself [mind-reading](#) and assuming what others must be thinking about me...

[Counseling Dialogue]

...

Counselor: It's understandable that past experiences of judgment can contribute to your current negative self-perceptions, ... How do you feel about delving deeper into these past experiences during our sessions?

Client: I often find myself [mind-reading](#) and assuming what others must be thinking about me.

Problem3. AI client tends to be overly positive

Counselor: Hi Travis, it's nice to meet you. How can I support you today?

Client: I've been struggling with some anxiety and self-doubt, particularly in relation to my hobby of building furniture.

Counselor: It's great that you've recognized the need for support ... Can you tell me more about the specific thoughts or fears that come up when you're building furniture?

Client: I often worry that if I use anything other than seasoned oak, my furniture won't turn out well.

Counselor: It sounds like you have a fear ... Have you ever considered what evidence there is to support or refute these thoughts about using seasoned oak?

Client: I haven't really thought about it that way before.

Counselor: That's okay, Travis. It's all part of the process of exploring and understanding our thoughts. ... How about we work on reframing this thought in a more realistic and constructive way? What do you think?

Client: [I'm willing to work on reframing my thoughts](#) about using seasoned oak in a more realistic way.

Figure 22: Example of the problem of using ChatGPT as a counselor and client agent.

Client Information

Name: Christopher Tate

Age: 21

Gender: male

...

Presenting Problem

- I have been experiencing feelings of hopelessness and despair due to the difficulty in finding a job.
 - These feelings started a few weeks ago when I was unable to secure a job despite multiple applications and interviews.
- ...
- I have tried various job search strategies and networking but have not been successful.

Reason for Seeking Counseling

- The prolonged period of unemployment and the resulting negative thoughts have led me to seek counseling to cope with these emotions.

Past History (including medical history)

- I have not experienced similar problems before.
 - I have not received treatment or counseling for psychological problems in the past.
- ...

Counseling with Counselor Agent (ChatGPT w/o planning)

Counselor: Hello, Christopher. It's nice to meet you. How are you doing today?

Client: I've been feeling really stressed about my job search. It seems like no matter how many applications I send out or interviews I go to, I just can't land a job. It's starting to make me feel hopeless.

...

Counselor: Christopher, it sounds like you're really struggling with feelings of failure... Have you thought about setting small, achievable goals for yourself to help build your confidence and sense of accomplishment during this challenging time?

Client: It's difficult to see past the rejections and failures.

Counselor: Christopher, it's completely understandable ... Have you considered incorporating self-care activities into your routine to help manage stress and boost your mood during this tough time?

Client: I struggle to see past the rejections and failures.

Counselor: Christopher, it's completely normal to feel overwhelmed ... Have you thought about incorporating self-care activities into your routine to help manage stress and boost your mood during this challenging time?

Client: I feel like I'm stuck in this cycle of rejection and failure.

Counseling with Counselor Agent (ChatGPT w/ planning with CBT)

Counselor: Hello, Christopher. It's nice to meet you. How are you doing today?

Client: I've been feeling really stressed about my job search. It seems like no matter how many applications I send out or interviews I go to, I just can't land a job. It's starting to make me feel hopeless.

Counselor: It's understandable that the job search process ... Have you noticed any patterns in your job search or specific challenges that you've been facing?

Client: I have noticed that I tend to feel really discouraged after receiving rejection emails from job applications.

...

Counselor: Finding activities that help you relax and recharge can be challenging, but it's important to keep trying different things until you find what works for you. Have you considered spending time with your pet dog? Animals can be great sources of comfort and relaxation.

Client: I'll try spending more time with my pet dog to see if that helps me relax.

Figure 23: Comparison of counseling dialogue results between ChatGPT with CBT planning and without planning.

Client Information

Name: Angela Margaret

Age: 60

Gender: female

...

Presenting Problem

- I have been feeling upset and insecure after a recent encounter with a rude cashier at the grocery store where I do the shopping for my house. This incident triggered feelings of inadequacy and low self-worth.
- The issue started after the encounter with the cashier last week.

...

Reason for Seeking Counseling

I decided to seek counseling because these feelings of insecurity have been affecting my daily life and overall well-being, and I need help in managing them.

Past History (including medical history)

- I have not experienced similar problems before.
- I have not received treatment or counseling for psychological issues previously.

...

Counseling with Counselor agent (Psych8k-LLAMA3)

Counselor: Hi Angela, it's nice to meet you. How are you feeling today?

Client: I've just been feeling really insecure and upset after an encounter with a rude cashier last week. It made me feel inadequate, and those feelings haven't really gone away.

Counselor: I'm sorry to hear that you've been feeling this way, Angela. It's important to address these feelings of insecurity and inadequacy. Can you tell me more about the situation with the cashier and how it has impacted your daily life?

Client: It's just been hard to shake off those feelings of inadequacy.

Counselor: I understand that it can be difficult to deal with such feelings, Angela. It's important for us to work together to address your concerns. Let's focus on the situation with the cashier and what specifically made you feel inadequate. Also, think about how you can cope with such situations in the future. Can you identify any specific thoughts or physical reactions you had during or after the encounter with the cashier that might be contributing to your ongoing feelings of insecurity?

...

Counseling with Counselor agent (CAMEL-LLAMA3)

Counselor: Hi Angela, it's nice to meet you. How are you feeling today?

Client: I've just been feeling really insecure and upset after an encounter with a rude cashier last week. It made me feel inadequate, and those feelings haven't really gone away.

Counselor: I'm sorry to hear that you're feeling this way, Angela. Can you tell me more about what happened with the cashier and how it made you feel?

Client: It's like everything I do is judged by others, and that encounter just reinforced my feelings of not being good enough.

Counselor: The cashier had this condescending tone and gave me a disapproving look when I was checking out my groceries. It felt like I was being scrutinized for no reason.

Client: It's like everything I do is judged by others, and that encounter just reinforced my feelings of not being good enough.

Counselor: That must have been very uncomfortable for you. How did you respond in that moment?

Client: I tried to ignore it and just hurried through the transaction, but it really stuck with me afterwards.

Counselor: It sounds like the experience left you feeling quite shaken. Have you noticed this pattern of feeling judged or inadequate in other areas of your life as well?

...

Figure 24: Comparison of results for counseling dialogue between Psych8k-LLAMA3 and CAMEL-LLAMA3.