Psychological Counseling Cannot Be Achieved Overnight: Automated Psychological Counseling Through Multi-Session Conversations

Anonymous ACL submission

Abstract

In recent years, Large Language Models (LLMs) have made significant progress in automated psychological counseling. However, current research focuses on single-session counseling, which fails to capture the ongoing nature of real-world counseling processes. Effective counseling is not a one-time interaction but an ongoing process, where each session builds upon the last to progressively address a client's issues. To overcome this limitation, we introduce a dataset for **Mu**lti-**S**ession **Psy**chological Counseling Conversation Dataset (**MusPsy-Dataset**). Due to the sensitive nature of counseling and inherent privacy concerns, we construct this synthetic dataset from authoritative, publicly available psychological case reports. The MusPsy-Dataset is designed to reflect the temporal continuity and shifting mental states of a single client across multiple sessions. Leveraging this resource, we propose the **MusPsy-Model**, which tracks client progress and dynamically adjusts counseling goals over time. Experiments show that our model performs better than baseline models across multiple sessions.

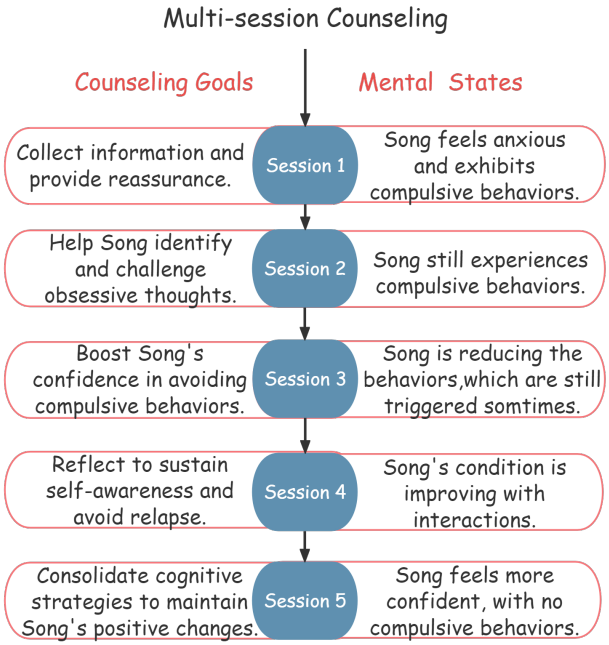

Figure 1: A real example of multi-session counseling involves a client demonstrating psychological improvement over five sessions. Each session targets specific goals, and the client's mental state gradually improves, an outcome that would be unachievable in a single-session setting.

1 Introduction

In today's society, individuals face increasing psychological pressure (Organization, 2024), and the demand for mental health support continues to surge in fast-paced modern life (Samji et al., 2022). Meanwhile, accessible mental health services remain scarce (Grant et al., 2018), motivating growing interest in leveraging large language models (LLMs) to deliver automated psychological counseling.

Previous research has advanced the modeling of psychological theory and the creation of various psychological counseling conversations, yielding positive results (Lee et al., 2024; Qiu et al., 2024; Na, 2024). However, most existing studies are limited to a single-session setting. Unlike casual chats, psychological counseling, such as Cognitive Behavioral Therapy (CBT), often spans multiple sessions (Craske, 2010). As shown in Figure 1, clients experience gradual psychological changes through counselor-guided multi-session counseling. During these sessions, counselors serve as navigators, dynamically recalibrating counseling goals based on the evolving mental states of their clients (Baur et al., 2024; Dobson and Dozois, 2021). This highlights the importance of developing automated models capable of sustaining coherent multi-session counseling conversations.

To address these limitations, we introduce a paradigm of multi-session automated psycho-

logical counseling. Unlike previous research focused on single sessions, our work tracks the progression across multiple sessions. This paradigm captures the dynamic mental states of clients and allows for continuous adjustments to counseling goals for each session. Multi-session counseling is not merely repeating the same counseling process in each session; instead, the goals evolve with the client's progress. Consequently, we can cumulatively build client rapport, effectively address clients' issues, and support long-term psychological improvement.

Overall, we develop **MusPsy**-Dataset, a large-scale **Mu**lti-**S**ession **Psy**chological counseling conversation dataset grounded in CBT, constructed with strong LLMs. We collect multi-session client profiles and session-wise counseling goals from publicly available psychological case reports as the foundation for generation. Since LLMs often fail to produce coherent multi-session dialogues in a single pass, we adopt a top-down generation pipeline. We first generate multiple short seed conversations to outline the cross-session flow, and then expand these seeds into complete, consistent, and goal-driven multi-session counseling conversations. During expansion, we incorporate counseling fragments from case reports and carefully designed prompts to improve fidelity and coherence. In addition, to facilitate training multi-session counseling models, we extract session-level memories from the generated conversations to serve as structured summaries of a client's state and progress.

Building on MusPsy-Dataset, we train **MusPsy-Model**, an automated counseling model with memory augmentation to conduct coherent multi-session counseling. We evaluate the model using multiple psychological metrics, including basic indicators, the Working Alliance Inventory–Short Form (WAI), and the Positive and Negative Affect Schedule (PANAS), demonstrating improved counseling quality and longitudinal consistency.

- We pioneer the multi-session paradigm in automated counseling to model the longitudinal counseling process.
- We introduce MusPsy-Dataset, the first large-scale, multi-session counseling dataset, constructed via a novel top-down generation methodology.
- We propose MusPsy-Model, a memory-augmented model that outperforms baselines and demonstrates improvement in counseling outcomes across multiple sessions.

2 Related Work

2.1 Cognitive Behavioral Therapy

Cognitive Behavioral Therapy (CBT) has long been recognized as an effective intervention for individuals struggling with depression and anxiety (Beck, 2020). Previous research describes how people with these disorders often develop negative, irrational thoughts that reinforce detrimental beliefs about themselves, others, and the world. To disrupt this cycle, CBT focuses on identifying and challenging these automatic thoughts and core beliefs (Longmore and Worrell, 2007). CBT is a structured, multi-session process that cannot be accomplished in a single session (Hayes and Hofmann, 2018). During CBT sessions, counselors first assist clients in recognizing unhelpful thoughts. They then guide clients in challenging and correcting these distortions using various CBT techniques, which ultimately help in reconstructing more positive automatic thoughts and beliefs over time (Fenn and Byrne, 2013). This counseling process is vital for promoting mental health and well-being.

2.2 Automated Psychological Counseling

Eliza, the pioneering system, used a rule-based approach for Rogerian counseling (Rogers, 1952; Weizenbaum, 1966). However, progress in data-driven research on Automated Psychological Counseling has been hindered by the limited availability of data. The creation of public, real, large-scale datasets for counseling remains impossible due to privacy concerns and the protection of vulnerable groups. However, the advent of LLMs has enabled the generation of synthetic data for counseling datasets. For instance, the SMILE dataset (Qiu et al., 2024), an improvement over PsyQA (Sun et al., 2021), includes counseling dialogues and is used by CBT-LLM (Qiu et al., 2024). CPsyCoun extracts counseling data from public psychology reports without requiring specific domain expertise (Zhang et al., 2024). Healme (Xiao et al., 2024) focuses on optimizing CBT guidance through LLMs, while Cactus (Lee et al., 2024) emphasizes the construction of more complete single-turn CBT processes.

Compared to these efforts, our work incorporates multi-session dynamics into automated psychological counseling. We introduce a novel task.

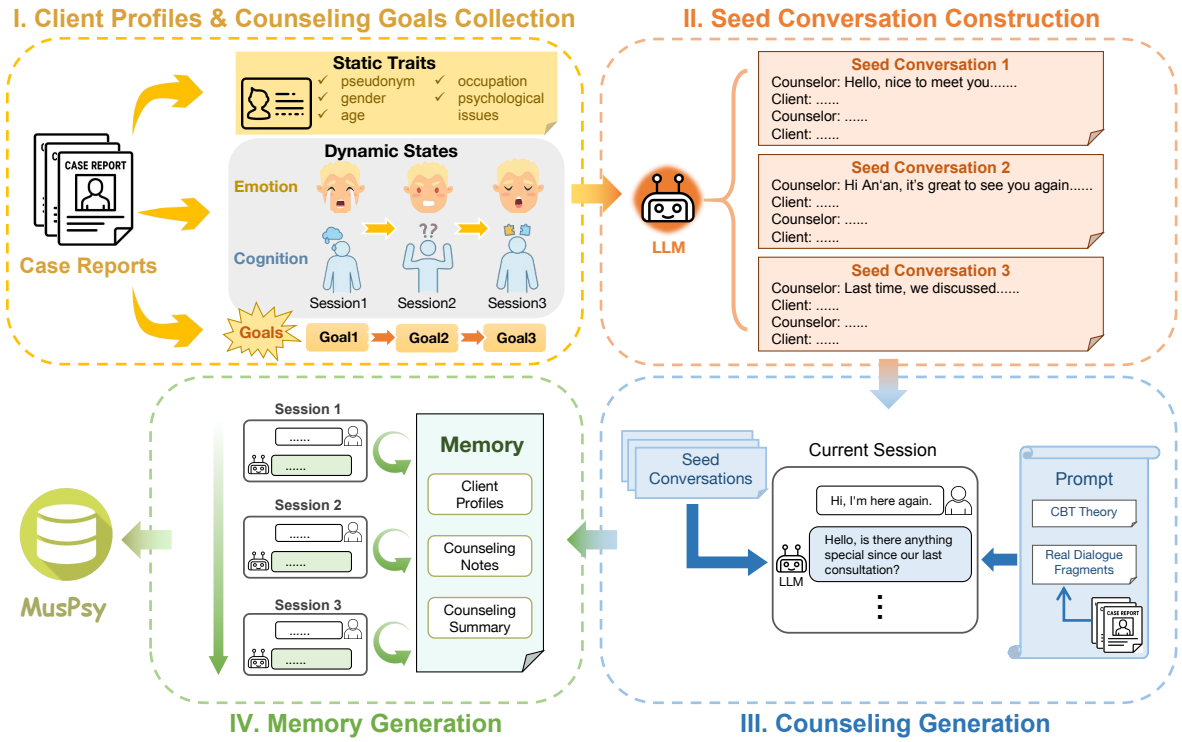

Figure 2: This figure illustrates the process of constructing the MusPsy-Dataset. It includes four parts: Client Profiles & Counseling Goals Collection, Seed Conversation Construction, Counseling Generation, and Memory Generation. Through these four steps, we obtained a high-quality MusPsy-Dataset.

3 MusPsy-Dataset: A Multi-Session Psychological Counseling Dataset

As illustrated in Figure 2, the construction of our dataset involves four key stages: **Client Profiles & Counseling Goals Collection, Seed Conversation Construction, Counseling Generation, and Memory Generation**.

3.1 Client Profiles & Counseling Goals Collection

To support multi-session automated counseling, constructing realistic and context-aware client profiles is essential. While previous studies often rely on synthetic client features focusing only on static traits—such as age, gender, and presenting problems—they typically lack the dynamic evolution of a client's mental state across multiple sessions, resulting in limited support for long-term counseling simulations.

In our work, we address this issue by gathering multi-session client profiles from credible psychological case reports published in academic journals and professional books. While these reports exclude full session transcripts for privacy reasons, they document the client's evolving state and the corresponding counseling goals for each session. Unlike prior methods that only capture a client's initial background, our data collection process captures both static traits and dynamic states, as well as the counselor's planned goals for each session.

- **Static Traits**: These include pseudonym, gender, age, occupation, and initial psychological issues reported during intake.
- **Dynamic States**: These capture session-by-session changes, including recent life events and the client's evolving emotional and cognitive state.
- **Counseling Goals**: These define the intended direction and focus for an upcoming counseling session.

Client profiles are composed of static traits and dynamic states. Counseling goals are not included in client profiles, as they represent the counselor's plan rather than the client's inherent state. Figure 3 illustrates typical multi-session CBT goals and methods, which guide the direction and content of each session.

We use GPT-4o to extract structured information from case reports, which is then manually validated.

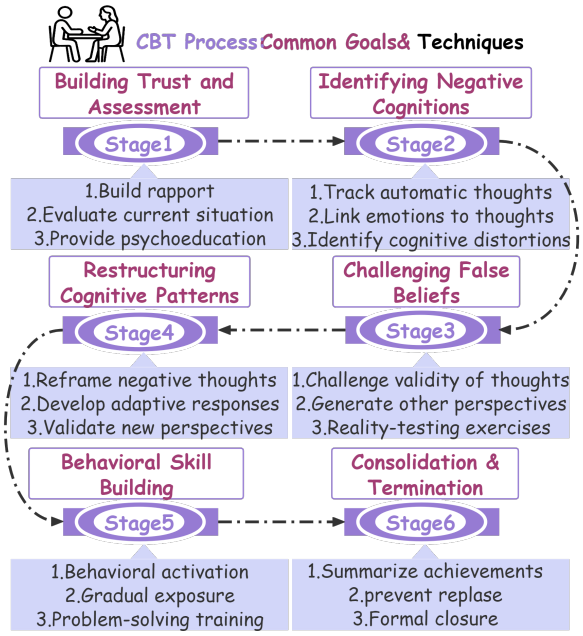

Figure 3: Multi-session CBT generally follows a paradigm in which the counselor typically adheres to this process, integrating it with the client's actual situation to set more specific and actionable counseling goals and use counseling techniques (Beck et al., 2011).

See Appendix A for details on client profile and counseling goal collection.

3.2 Seed Conversation Construction

Directly generating all complete sessions at once leads to limited depth due to LLM context and instruction-following constraints. Conversely, generating sessions separately often produces repetitive phrasing and weak continuity, as the LLM lacks awareness of counseling progression. This highlights the need for an alternative generation method that ensures both coherence and contextual richness across counseling sessions.

Generating Seed Conversations for Coherence: As shown in Figure 2, to ensure coherence across multi-session interactions, we introduce an intermediate step: constructing seed conversations before generating full counseling conversations. Seed conversations are short, concise multi-session conversations generated by the LLM in a single pass. Despite their brevity, they possess high coherence due to this concurrent generation process.

To ensure coherence, we input client profiles and their corresponding goals for each session simultaneously. We instruct the LLM to generate coherent seed conversations, each comprising only 3 to 4 turns and approximately 100–200 tokens, with the content aligned with the session's goal.

3.3 Counseling Generation

Given the generated seed conversations, we aim to generate complete, extended counseling sessions that are contextually coherent across sessions.

Session-Level Expansion with Contextual Constraints: During generation, to ensure contextual coherence, we prompt the LLM to generate each complete counseling session based on its corresponding seed conversation, while also providing all preceding seed conversations as context. We guide the model to create contextually linked conversations. This design allows us to preserve temporal dependencies and emotional progression across multiple sessions, ensuring a complex and coherent multi-session simulation.

Prompting with Few-Shot Guidance: We design prompts to guide the generation process by combining descriptions of core CBT techniques with the LLM's own understanding of CBT to expand the seed conversations.

We also extract anonymized, incomplete counseling fragments from a subset of case reports that include such content. We utilize these extracted fragments as few-shot examples, aligning each with the corresponding stage of the client's profile. If a corresponding counseling fragment exists in the source report for a given session, we use that specific fragment. If not, we retrieve a closely related fragment based on the session's counseling goals. Rather than strictly pre-defining the counseling techniques, we allow the LLM to leverage its inferential strengths for flexible application. However, to provide guidance, we prompt the LLM to integrate the sampled techniques into its generation, which results in more contextually appropriate and professional outputs.

Script-Based Generation: We adopt an efficient single-agent script-based approach, which has been shown in prior single-session research (Lee et al., 2024) to significantly reduce computational cost and inference time while maintaining counseling coherence and quality. In this approach, one LLM generates the entire multi-turn session in a single pass, much like a screenwriter writing a script, based on the client's profile and the corresponding seed conversation.

3.4 Memory Generation

Additionally, long-term multi-session counseling can easily exceed the available context window, presenting challenges in maintaining consistency. To support consistent and context-aware multi-session counseling, we construct structured memory representations for each session.

Unlike general chat memory, the design of counseling memory needs to reflect psychological requirements. To meet this challenge, we introduce an external memory module that captures key session-level information, enabling us to maintain coherence and continuity over extended counseling sessions. Inspired by documentation standards in CBT practice (Bemister and Dobson, 2011; Lawlor-Savage and Prentice, 2014), we design a memory structure specifically tailored for multi-session counseling.

- **Client Profile:** Contains the client's static and dynamic information, including personal background and evolving mental states.
- **Counseling Notes:** Captures key session-level information, including counselor observations, session goals, and counseling assignments during multiple counseling sessions.
- **Counseling Summary:** A brief summary of the entire series of sessions.

At the end of each counseling session, we summarize the client's conversation into this memory structure. This approach allows the model to update and retain crucial information without needing to store the complete counseling transcripts. As a result, subsequent sessions can maintain continuity and simulate a counselor's cross-session tracking and planning abilities. For more details on memory generation, please refer to Appendix D.

4 Data Statistics and Evaluation

4.1 Data Statistics

We provide an overview of the basic statistics of our dataset in Table 1. The MusPsy-Dataset consists of 1,400 clients' multi-turn counseling dialogues constructed based on CBT theory, with each session averaging 28.55 turns and 26.99 words per turn. On average, each client has 6.17 counseling sessions. This dataset scale is sufficient to support the construction of our new multi-session counseling model.

Item	Value
Client Profiles for Training	1,400
Client Profiles for Testing	100
Average Sessions Per Client	6.17
Average Turns Per Session	28.55
Average Words per Turn	26.99
Average Tokens Per Client	5693.27
Average Tokens Per Turn	33.23

Table 1: Data statistics of our MusPsy-Dataset

4.2 Competing Datasets

We introduce several previous works for comparison, including the SMILE dataset (Qiu et al., 2024), a multi-turn mental health conversation dataset based on PsyQA; and the Cactus dataset (Lee et al., 2024), another multi-turn CBT dataset for mental health. We also include the SimPsyDial (Qiu and Lan, 2024) and CPsyCoun (Zhang et al., 2024) datasets. We evaluate data quality and model performance as reported in these studies within our evaluation framework. These works are similar to ours in constructing mental health counseling datasets and providing mental support models.

4.3 Dataset Quality Evaluations

We randomly select 100 counseling examples per dataset. Two psychology experts perform manual evaluations, and GPT-4o does automated ones. We focus on four basic metrics and counselor-client working alliance, using it as a consistent measure across diverse counseling data (Horvath, 2001). Research indicates working alliance predicts counseling outcomes (Horvath and Symonds, 1991; Horvath et al., 2011). All metrics are rated 1-5 Likert scale.

We evaluate our dataset using basic metrics (Munder et al., 2010) across four dimensions: **Helpfulness:** the practical utility of counselor explanations and information; **Empathy:** the counselor's ability to understand and share client feelings; **Guidance:** the availability and specificity of practical advice; and **Coherence:** the logical consistency in the conversation. We also use the WAI (Munder et al., 2010), which assesses three dimensions: **Goal Agreement:** mutual agreement on counseling goals; **Task Agreement:** agreement on the methods used; and **Emotional Bond:** mutual trust, confidence, and liking. Each dimension in the WAI is assessed by four questions, and we report the average score for each dimension. For details

Evaluator		Hel.	Emp.	Gui.	Coh.	Avg.	Task.	Bond.	Goal.	Avg.
GPT-4o	SMILE	4.05	4.49	3.98	4.43	4.23	3.60	4.35	3.61	3.85
	Cactus	4.62	4.86	4.56	**4.99**	4.76	3.56	4.56	3.66	3.92
	SimPsyDial	4.58	4.90	4.51	4.99	4.75	3.77	4.31	3.56	3.97
	CPsyCoun	4.23	4.43	4.09	4.87	4.41	3.56	4.30	3.71	3.86
	MusPsy	**4.99**	**4.98**	**4.98**	**4.99**	**4.98**	**4.11**	**4.69**	**4.44**	**4.41**
Human	SMILE	3.84	3.90	3.57	3.77	3.77	3.13	4.22	3.14	3.49
	Cactus	4.49	4.59	4.18	4.65	4.47	3.60	4.65	3.49	3.92
	SimPsyDial	3.73	4.56	4.32	4.77	4.34	3.36	4.32	3.58	3.75
	CPsyCoun	3.60	3.83	4.17	3.73	3.83	3.08	4.02	3.43	3.51
	MusPsy	**4.89**	**4.82**	**4.70**	**4.88**	**4.82**	**4.20**	**4.72**	**4.14**	**4.35**

Table 2: Data Quality Evaluation. This involves the evaluation of four fundamental metrics and three dimensions of the WAI scale, completed through both GPT-4o and manual assessment. Evaluations are conducted in different languages based on the same prompt translated into English.

on the evaluations, see Appendix E.

4.3.1 Single-Session Evaluation

To compare data quality with other single-session datasets, we focus on individual sessions. For this evaluation, we treat each session from a client as a distinct entry and randomly sample one per client. As Table 2 shows, MusPsy-Dataset outperforms existing counseling datasets across most metrics.

While many current datasets perform well on basic measures, this highlights the importance of assessments grounded in psychological theory—an area where MusPsy-Dataset particularly excels. MusPsy-Dataset demonstrates particular strength in the WAI. It shows a strong Emotional Bond, which indicates the data captures realistic elements of trust and confidence. It also has high Task Agreement, suggesting the counseling reflects effective counseling methods. We believe this is because our dataset reflects a gradual progression and does not expect rapid improvement in a single session. This gradual approach also helps our data excel in Goal Agreement.

These findings confirm the overall high quality of our dataset. We attribute its success to a more accurate simulation of real-world counseling, where the entire counseling process is rarely completed in one session. The multi-session structure of MusPsy-Dataset captures the development of deeper understanding.

4.3.2 Multi-Session Evaluation

We conduct a session-by-session analysis of the MusPsy-Dataset. For each client, we analyze up to 6 sessions. As shown in Figure 4, our results

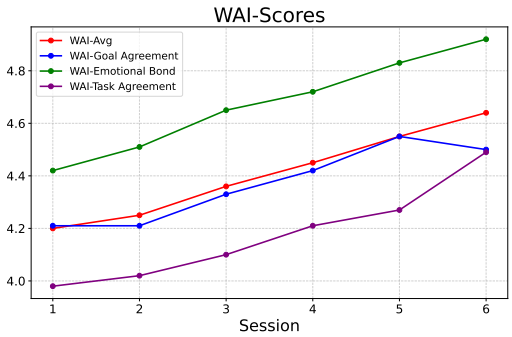

Figure 4: Evaluation of WAI scores for each counseling session in multi-session conversations. A general upward trend is observed.

show that as counseling progresses, the working alliance between clients and counselors typically strengthens. However, a slight decrease in goal agreement is observed. We attribute this to the fact that some clients in our dataset finish counseling around 5 sessions. This phenomenon occurs because clients who continue to a sixth session often need to work on deeper issues, such as core beliefs. These deeper goals are often harder for clients to fully accept.

Overall, this finding reinforces the idea that effective automated psychological counseling and support must be a process, rather than a one-time event.

5 Task Setting

As shown in Figure 5, we define three tasks that mirror the multi-session counseling workflow of a human counselor to evaluate our dataset's utility in training MusPsy-Model: Memory Extraction,

Goal Planning, and Counseling Generation. Our aim is for the counseling model to manage counseling progress and generate coherent counseling dialogues.

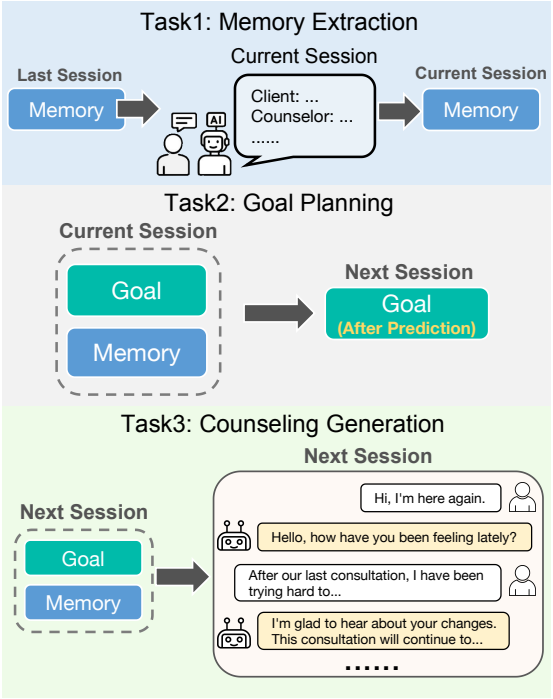

Figure 5: Overview of Our Three Tasks: Illustrating the three tasks of multi-session psychological counseling: Memory Extraction, Goal Planning, and Counseling Generation, where the extracted memory from a session informs the goal planning for the subsequent session, which guides the content generation during next counseling sessions.

- **Memory Extraction(Task 1):** MusPsy-Model extracts key information from the current session and updates its memory.
- **Goal Planning(Task 2):** Based on its memory, MusPsy-Model plans counseling goals for the next session, simulating pre-session counseling progress management.
- **Counseling Generation(Task 3):** MusPsy-Model engages in a counseling session with the client, informed by past memory and the planned goals.

We simultaneously optimize these tasks via supervised finetuning on annotated samples. During training, we employ specific prompts for each task to ensure the MusPsy-Model simultaneously develops proficiency in all three capabilities.

6 Experiment

6.1 Experiment Setup

Our experiments are conducted using the Meta-Llama-3-8B-Instruct model (Dubey et al., 2024). To maintain fairness and comparability, we train models the other datasets on Meta-Llama-3-8B-Instruct instead of using the official models. During the training phase, all prompts and hyperparameters are kept consistent with those specified in the original paper. For details on experiment setup, see Appendix F.

6.2 Evaluations and Results of Task 1&2

We evaluate the individual performance of Memory Extraction and Goal Planning using standard machine metrics. The results demonstrate our capability to effectively update its memory based on session content and to predict relevant counseling goals for the following session.

Metric	BLEU-1	BLEU-2	F1
Task 1	42.7	26.2	38.0
Task 2	44.1	34.0	35.4

Table 3: Evaluation results for Task 1 and Task 2.

These results suggest that our MusPsy-Model can learn to memorize key information discussed during a counseling session and utilize this understanding to anticipate the logical progression of counseling by proposing relevant next goals. This foundational capability is crucial for building more sophisticated counseling models that can engage in meaningful and progressive multi-session counseling.

6.3 Evaluations and Results of Task 3

As shown in Figure 5, for Counseling Generation, we provide the memory from the previous conversation along with the goal for the current session as input to the LLM. We emphasize the importance of this information in the instructions.

We employ an evaluation framework to assess model performance through multi-session counseling simulations, using 100 test client profiles. In these simulations, the LLM acts as a client interacting with the counselor model. In addition to direct metrics evaluating the counseling conversation, we measure changes in clients' emotional states using the Positive and Negative Affect Schedule (PANAS), which provides results entirely from the clients' perspective. After the session, we ask the

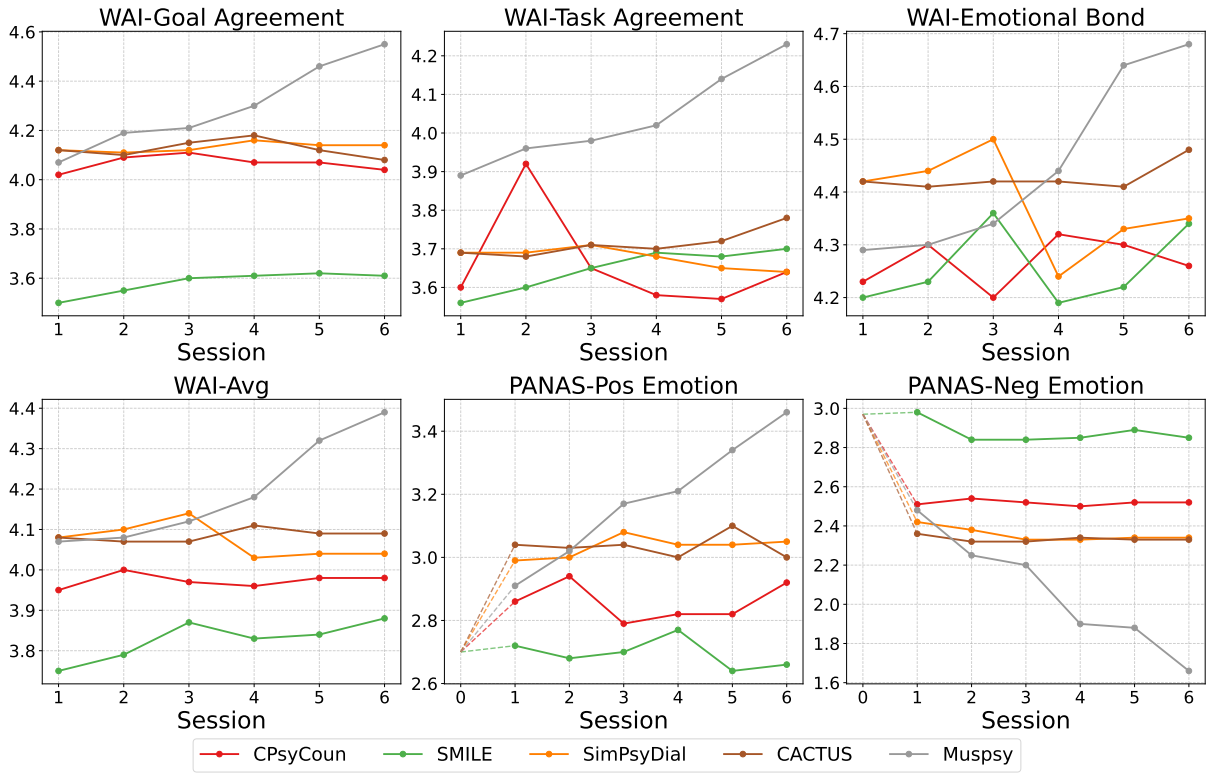

Figure 6: Emotional changes of the LLM client and performance changes of the LLM counselor across multiple sessions. It can be observed that for other models and approaches, there is essentially no significant change in the counselor's performance and the LLM's emotions within the first session. In contrast, the MusPsy-Model demonstrates a continuous improvement in the client's state.

LLM client to evaluate its PANAS score to analyze the counseling model's effectiveness. It updates its own state and prepares for the next counseling conversation.

We do not use PANAS to assess the dataset, as it measures emotion from the client's perspective, making comparisons impossible without the same client. For these evaluations, we utilize GPT-4o.

6.3.1 Evaluation Results

As shown in Figure 6, our MusPsy-Model effectively reduces negative emotions and enhances positive emotions in the long term. We test the LLM client for 6 sessions. We can observe that after two sessions, the other models are practically unable to continue addressing the client's issues, as they consistently generate counseling dialogues that are largely the same as the previous one. In contrast, MusPsy-Model's advantage is particularly evident; after six sessions, it significantly alleviates the client's negative emotions, creating a stark contrast with other approaches and highlighting our strengths.

For most counseling models, their LLM counselors also fail to benefit from multiple sessions. This implies that these models do not show significant improvement in their counseling quality as the simulated counseling progresses, which is counter to what is expected in human counseling. However, MusPsy-Model exhibits a long-term upward trend in its counseling performance, which we attribute to our modeling of multi-session dynamics. Across multiple turns, it continuously strengthens the emotional bond with the client and enhances the alignment of goals.

7 Conclusion

In conclusion, our work aims to build more comprehensive automated psychological counseling through the introduction of the MusPsy-Dataset. The dataset's focus on multi-session conversations and the explicit modeling of client progress offer a valuable resource for developing LLM-based systems that can better approximate the complexities of real-world counseling. Our initial results with the MusPsy-Model are promising, demonstrating its ability to track client state and adapt counseling goals over time, leading to positive outcomes.

Limitations

Despite what we believe to be many reasonable contributions, we acknowledge the limitations of our work. Many of these limitations stem from cost constraints and challenges inherent to the field, and are not entirely within our control.

- **Limited Counseling Theory:** While multiple rounds of counseling dialogues are conducted based on CBT theory, this approach may not be suitable for all clients. Our model lacks the ability to implement flexible, multi-modal counseling strategies.
- **Insufficient Attention to Resistant Clients:** Both in dataset synthesis and model evaluation, we do not give enough attention to clients who are resistant to treatment.
- **Realism concerns raised by LLM:** We acknowledge that despite our efforts to pursue realism by using authentic case reports, we cannot fully resolve the issue of realism. However, due to privacy concerns, fully authentic psychological counseling data is nearly impossible to obtain and share ethically. We believe our work represents a necessary compromise.
- Furthermore, we fully understand the potential biases introduced by LLM evaluation. While we incorporate human expert evaluation to mitigate this, cost and reproducibility considerations limit the availability of better solutions for LLM evaluation at this time. We consider this a limitation shared by this type of work.
- **Potential Privacy Risks:** Some studies have shown that LLMs may provide harmful advice and incorrect responses, which is unavoidable. Multi-turn memory also raises concerns about client privacy.

We acknowledge the limitations but emphasize the importance of pushing the field forward despite those limitations, given the real-world need for automated counseling tools. We are arguing for a pragmatic approach. We do not consider our work to be a completed and deployable mature study but rather an exploratory step. We do not intend and will not directly promote it as a commercial service to avoid harming vulnerable populations. Research in this area requires the joint efforts of researchers in other fields, such as safety, culture, privacy, and debiasing, to promote the widespread benefit of artificial intelligence technology to society. However, we believe our work can inspire subsequent researchers to introduce new solutions in psychological counseling research and consider multi-session scenarios. This is the significance and rationale behind our thinking and undertaking this work, and we hope it will bring more positive impact to society.

Ethical Statement

This study adheres to the Institutional Review Board (IRB) approval from our institution, ensuring that no psychological harm or burden is inflicted upon any participant. All case reports are publicly available and anonymized in their original sources. Therefore, we commit to publicly sharing all data after the paper is accepted. Participants for this study are recruited through advertisements targeting trained experts and human participants. Our experts hold a bachelor's degree and possess at least two years of relevant work experience. All human participants sign an informed consent form and are compensated at the average wage level of the local region. For the experts, compensation is based on the duration of their work, aligning with the average income level for their profession in the area. Participants are allowed to withdraw from the study at any time to ensure their rights are not violated. Furthermore, we are fully aware of the potential biases introduced by LLM evaluations, such as those performed by GPT-4o. While we incorporate human expert evaluations to mitigate these biases, cost and reproducibility considerations currently limit the scale of such human assessment. We consider this a shared limitation within the broader field of LLM-based conversational AI research.

References

Anaïs Baur, Anne Trösken, and Babette Renneberg. 2024. Content and attainment of individual treatment goals in cbt. *Psychotherapy Research*, 34(1):111–123.

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Judith S Beck, AT Beck, and JS Beck. 2011. Cognitive behavior therapy: basics and beyond. ed. *New York*.

Taryn B Bemister and Keith S Dobson. 2011. An updated account of the ethical and legal consid-

erations of record keeping. *Canadian Psychology/Psychologie canadienne*, 52(4):296.

Michelle G Craske. 2010. *Cognitive–behavioral therapy*. American Psychological Association.

Keith S Dobson and David JA Dozois. 2021. *Handbook of cognitive-behavioral therapies*. Guilford Publications.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kristina Fenn and Majella Byrne. 2013. The key principles of cognitive behavioural therapy. *InnovAiT*, 6(9):579–585.

Kiran L Grant, Magenta Bender Simmons, and Christopher G Davey. 2018. Three nontraditional approaches to improving the capacity, accessibility, and quality of mental health services: An overview. *Psychiatric Services*, 69(5):508–516.

Steven C Hayes and Stefan G Hofmann. 2018. *Process-based CBT: The science and core clinical competencies of cognitive behavioral therapy*. New Harbinger Publications.

Adam O Horvath. 2001. The alliance. *Psychotherapy: Theory, research, practice, training*, 38(4):365.

Adam O Horvath, AC Del Re, Christoph Flückiger, and Dianne Symonds. 2011. Alliance in individual psychotherapy. *Psychotherapy*, 48(1):9.

Adam O Horvath and B Dianne Symonds. 1991. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology*, 38(2):139.

Linette Lawlor-Savage and Jennifer L Prentice. 2014. Digital cognitive behaviour therapy (cbt) in canada: Ethical considerations. *Canadian Psychology/Psychologie canadienne*, 55(4):231.

Sueon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Kim, Seungbeen Lee, and 1 others. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274.

Richard J Longmore and Michael Worrell. 2007. Do we need to challenge thoughts in cognitive behavior therapy? *Clinical psychology review*, 27(2):173–187.

Thomas Munder, Fabian Wilmers, Rainer Leonhart, Hans Wolfgang Linster, and Jürgen Barth. 2010. Working alliance inventory-short revised (wai-sr): psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 17(3):231–239.

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.

World Health Organization. 2024. *Global tuberculosis report 2024*. World Health Organization.

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

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

Carl R Rogers. 1952. " client-centered" psychotherapy. *Scientific American*, 187(5):66–75.

Hasina Samji, Judy Wu, Amilya Ladak, Caralyn Vossen, Evelyn Stewart, Naomi Dove, David Long, and Gaelen Snell. 2022. Mental health impacts of the covid-19 pandemic on children and youth–a systematic review. *Child and adolescent mental health*, 27(2):173–189.

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.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.

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

A Client Profile and Counseling Goals Construction

We utilize GPT-4o to extract client profiles from case reports by employing a one-shot prompt. This ensures that the model extracts client profiles in a standardized format. The prompt format is shown in Figure 7. We also employed GPT-4o to generate counseling goals for each session.

Prompt

You now need to help me summarize the information of the psychological counseling client and the counselor's counseling goals based on the user profile text and the text of each consultation session.

Summary Content and Principles:

Client's Static Traits:
1.Basic information of the client, such as age, gender, environment, medical history, etc. Summarize in detail and keep this part consistent across all counseling sessions.
Client's Dynamic States:
1.Recent life events, These refer to the triggering events that have recently occurred for the client. These events will vary across counseling sessions.
2.Client's Emotional State: This indicates the client's current emotional state.
3.Client's Cognitive State: This encompasses the client's current cognitive state, particularly any maladaptive cognitions and beliefs.

Content and Principles of Counselor's Counseling Goal Summary:
1. Counselling goals for this session: i.e., what goals need to be achieved in this session.
2. The counselor is skilled in using Cognitive Behavioral Therapy (CBT) for counseling. The summarized counse
ling and plans should be established entirely based on CBT theory.

The summary format and examples are as follows. Note that spaces, indentation, and blank lines should be consistent with the

Example:
First Counseling Session
Client's Static Traits:
[Alias]: An'an
[Age]: High School, Grade 2
[Gender]: Male
[Family Status]: Father works away from home year-round; mother is a full-time housewife. The mother has been primarily responsible for An'an's studies since childhood and has extremely high expectations for his academic performance.
[Medical History]: Started experiencing emotional lability and visual hallucinations in the second semester of his first year of high school.
Client's Dynamic States:
[Recent Life Events]: An'an's dissatisfaction with his mother has intensified due to her refusal to allow him to participate in the school arts festival preparations. However, he feels that his mother is doing it for his own good and that he shouldn't be angry.
[Emotion State]: An'an feels irritable and anxious. He is also afraid of his own emotions, fearing that he will lose control of his anger and harm others.
[Cognitive State]: He suspects he has a mental health problem and experiences visual hallucinations while doing homework.

Counseling Goals:
Gather An'an's clinical information to understand his basic situation. The focus is on listening to An'an's narrative and establishing a trusting relationship with him through empathy and unconditional positive regard.

Figure 7: Prompt for extracting client profiles and counseling goals for each session from case reports.

B Seed Conversation Construction

We construct multiple seed conversation for each client. The creation of these seeds takes into account the previously established client profiles and the counseling goals for specific sessions. This preparatory stage is crucial for ensuring the coherence and relevance of our counseling. Figure 8 visually illustrates the prompt format used to create these foundational conversations, providing insight into the information and structure guiding their creation.

These brief seed counseling dialogues are generated simultaneously to ensure cross-session coherence. This concurrent generation helps the LLM to maintain a holistic view of the counseling arc, ensuring that each session logically flows from the previous one and sets the stage for the next. We use them to expand into complete counseling conversations.

C Counseling Generation

Given the multi-session seed conversations, the next step is to generate natural-sounding complete counseling sessions. This process aims to achieve

Prompt

Your task is to construct a set of multi-session seed psychological counseling conversations, realizing interaction with a professional counselor. The counseling must follow the guidelines below:

Basic Principles:
1.The counselor uses CBT (Cognitive Behavioral Therapy: CBT is a scientific psychotherapy method aimed at breaking the cycle of patients constantly reinforcing negative thoughts by identifying and challenging negative and irrational thinking beliefs.
2.The counseling includes a total of multi sessions. The continuity of each session must be guaranteed. In the first session, do not mention topics related to "the last time."
3.During the communication, the counselor should guide the entire counseling process.
4.The counseling should begin based on the preset [Counseling Goals]. The client has a given [Client Profile].
5.You need to generate a session for each goal and client profile, with each session outputting only 3–4 turns.

Basic Requirements for Counselor:
1.The counselor already knows the content and results of the previous counseling session. If it is the first consultation, do not output any content related to the previous consultation.
2.Use a friendly form of address for the user to create closeness.
3.Start with "Counselor:" to ensure the utterance follows the exact format and does not contain any control characters.
User's Speaking Guidelines:
Basic Requirements for Client:
1.The client seeks psychological counseling from the counselor and has a strong desire to talk about recent things that confuse them.
2.Fully express your feelings and reactions during the consultation process, including opinions on the counselor's questions and feelings about the discussed content, which can be questioning or agreeing.
3.Start with "Client:" to ensure the utterance follows the exact format and does not contain any control characters.

Client's Static Traits:
[Alias]: An'an
[Age]: High School, Grade 2
[Gender]: Male
[Family Status]: Father works away from home year-round; mother is a full-time housewife. The mother has been primarily responsible for An'an's studies since childhood and has extremely high expectations for his academic performance.
[Medical History]: Started experiencing emotional lability and visual hallucinations in the second semester of his first year of high school.

The first Counseling Session:
[Counseling Goals]: Gather An'an's clinical information to understand his basic situation. The focus is on listening to An'an's narrative and establishing a trusting relationship with him through empathy and unconditional positive regard.
Client Dynamic States:
[Recent Life Events]: An'an's dissatisfaction with his mother has intensified due to her refusal to allow him to participate in the school arts festival preparations. However, he feels that his mother is doing it for his own good and that he shouldn't be angry.
[Emotion State]: An'an feels irritable and anxious. He is also afraid of his own emotions, fearing that he will lose control of his anger and harm others.
[Cognitive State]: He suspects he has a mental health problem and experiences visual hallucinations while doing homework.
.....

Figure 8: Prompt for constructing seed conversations using counseling goals and client profiles.

Counselor: Hi An‘an, it’s great to see you again. Last time, we discussed having you reflect more on things that bring you joy. Have you identified any activities or sources of happiness besides music, particularly during stressful periods?

Client: Well, I've always enjoyed running. It helps clear my mind and gives me time to focus on other things.

Counselor: That‘s wonderful. Physical activity can be very effective for stress relief and mood improvement. Would you consider scheduling regular time for running?

Client: Yes, I've been wanting to start again but haven't found the time.

Figure 9: An example of a seed conversation, which is the client’s 6th session.

two key objectives: maintaining contextual coherence across all sessions and ensuring that the language and counseling techniques employed closely mirror those of a human counselor. The format of the prompts used to create complete counseling dialogues is visually illustrated in Figure 10, providing insight into the information and structure guiding their creation.

D Memory Generation

The purpose of the memory generation process is to summarize and synthesize the counselor’s record of the psychological counseling session, based on the provided counseling conversations. This involves extracting key information from the session and organizing it into a structured format. The memory generation process aims to create a structured representation of each counseling session, enabling

Prompt

Your task is to build a set of long-term psychological counseling conversations. Based on the following given seed conversations, generate more detailed complete counseling. The counseling must follow the guidelines below:

Basic Principles:
1.The counselor uses Cognitive Behavioral Therapy (CBT): CBT is a scientific psychotherapy method aimed at breaking the cycle of patients constantly reinforcing negative thoughts by identifying and challenging negative and irrational thinking beliefs.
2. The given seed conversation fragments represent a part of a continuous psychological counseling process. Your task is to expand upon these, generating more detailed session content to simulate the ongoing counseling process. Please maintain coherence in the expanded content, ensuring that subsequent sessions align logically and thematically with the previous seed conversations.
3.You only need to expand the current seed conversation, but the expanded content should connect with the previous seed conversations.
Basic Requirements for Counselor:
1.The counselor already knows the content and results of the previous counseling session
2.Use a friendly form of address for the user to create closeness.
3.Start with "Counselor:" to ensure the utterance follows the exact format and does not contain any control characters.
4.Refer to the techniques and tone of human counselors in the provided real counseling fragments when generating content.
Basic Requirements for Client:
1.The client seeks psychological counseling from the counselor and has a strong desire to talk about recent things that confuse them.
2.Fully express your feelings and reactions during the consultation process, including opinions on the counselor's questions and feelings about the discussed content, which can be questioning or agreeing.
3.Start with "Client:" to ensure the utterance follows the exact format and does not contain any control characters.

Real Counseling Fragments:
Client: I often blame myself for not being able to handle the pressure better. It's like there's this voice telling me that I'm not doing enough, and I'm failing my family.
Counselor : That's a tough place to be in emotionally. Do you find these thoughts intensify your stress or make you feel helpless at times?
Client: Absolutely. I feel trapped in this cycle of guilt and stress, making it hard to focus or sleep.
Counselor : Recognizing these patterns is a good first step. Can we explore the possibility that some of these thoughts might not be entirely accurate or fair to yourself?
Client: It's hard to imagine that at times, but I guess there could be exaggerations in what I'm thinking.

Preceding Seed Conversation:
The first Counseling Session:
Counselor: Hello Mr. Li, welcome to our session today. It's great that you reached out to discuss your current feelings. Could you share what recent events have been on your mind?
Client: Thank you. I've been feeling very anxious and overwhelmed lately, especially with my son's behavior and academic performance at school. I'm not really sure how to handle it all effectively.
Counselor: That sounds really tough. It seems like you're experiencing some significant stress regarding your son's situation. Is there a specific incident or event that's been worrying you recently?
Client: Yes, last weekend was quite difficult. My son refused to do his homework again, and I ended up slapping him in frustration. This led to a serious talk with his teacher, and now I'm feeling even worse about my response. He even ran away afterward, which really scared me.
Counselor: It sounds like you're dealing with a lot right now. Let's explore your thoughts and emotions around this situation. It can help us find a way to ease some of that burden you're feeling.
.....

Current Seed Conversation:
Counselor: Hello, Mr. Li. Welcome back. It's good to see you again. We talked about your plans for continuing to manage your anxiety and interact with your son. How have things been since our last session?
Client: It felt like a battle just to stay calm. I keep telling myself it's about progress, not perfection, but sometimes I still feel like I'm fighting against this overwhelming urge to just fix everything for him, or to get him to instantly understand the importance of his studies. It's that underlying anxiety about his future that still gets to me.
Counselor: That's a very common feeling for parents, Mr. Li — wanting the best for your child and feeling responsible for their path. Perhaps for this session, we can explore that specific anxiety: where it comes from, and how you might manage the pressure to "fix everything" while still guiding him effectively. How does that sound?
Client: That sounds very helpful. Yes, I'd really appreciate that. I think understanding that anxiety is key to moving forward without always feeling so burdened.

Figure 10: Prompt for constructing complete counseling dialogues using seed conversations.

the model to:

- Track client progress across multiple sessions.
- Maintain coherence and consistency in the counseling conversation.
- Simulate a counselor's ability to recall and utilize information from previous sessions.

The format of the prompts used to create these foundational conversations is visually illustrated in Figure 11, providing insight into the information and structure guiding their creation.

Prompt

Your task is to help me summarize and synthesize the counselor's memory of the psychological counseling session based on the provided counseling conversations.

Client Profiles: ……
Counseling Notes:
1. Counselor Observations: The counselor's observations and impressions of the client during multiple counseling sessions.
2. Counseling Assignments: Homework and tasks the client needs to complete after this session during multiple counseling sessions
3. Session Goal: The goal of this sessions.
Counseling Summary: A brief summary of the entire series of sessions.
Summary format and example are as follows. Please ensure that spaces, indentation, and blank lines match the example:

Example:
The First Counseling
Client Profile:
[Static Traits:] Star is a female in their late twenties working as a freelance graphic designer. They have been experiencing increased social anxiety and feelings of isolation over the past few months.
[Emotion and Cognitive State:] Star reported feeling anxious and avoidant in social situations, with recurring thoughts of being judged and misunderstood. They expressed a desire to build stronger connections with others but feel overwhelmed by the prospect.
Counseling Memory:
[Counselor Observations:] Star presented as articulate and self-aware but displayed noticeable discomfort when discussing social interactions. They expressed a strong internal conflict between their desire for connection and their fear of vulnerability.
[Counseling Assignments:] Star is encouraged to identify one small, manageable social interaction to engage in this week and to reflect on their thoughts and feelings before, during, and after the interaction. They are also asked to continue practicing mindfulness exercises to manage anxiety.
[Session Goal:] The goals for this session were to further explore the roots of Star's social anxiety, identify specific fear-based thoughts, and collaboratively develop a small, achievable behavioral experiment to challenge their avoidance patterns.
Counseling Summary:
Explored Star's social anxiety (desire for connection vs. fear of judgment). Agreed on a small social task to challenge avoidance. Mindfulness continues for anxiety management.

Figure 11: Prompt for constructing counselor's memory using counseling conversations.

E Data Evaluation

When conducting data quality assessments, we primarily use two types of metrics. The first category includes standard conversational evaluation metrics such as coherence, guidance, helpfulness, and empathy, with evaluation criteria shown in Figure 12. The second category adopts a psychological counseling perspective, employing the WAI to assess conversations, as illustrated in Figure 13. The instructions presented to human evaluators are largely identical to the prompts provided to the LLM evaluator.

- **Helpfulness** focuses on the applicability of explanations and suggestions provided by the **counselor**.
- **Coherence** evaluates the logical flow and structure of the conversation.
- **Empathy** assesses the counselor's ability to understand and respond to the client's feelings.
- **Guidance** evaluates the specificity and practicality of the counselor's suggestions.

The Working Alliance Inventory (WAI) is a tool designed to evaluate the quality of the therapeutic relationship between counselors and clients. It measures this relationship across three core dimensions: Goal Agreement, Task Agreement, and Bond.

- **Goal Agreement** focuses on whether the counselor and client share a mutual understanding of the counseling objectives and work together toward achieving them. This aspect is measured through items 4, 6, 8, and 11 on the scale.
- **Task Agreement** assesses the degree of cooperation between both parties in pursuing these goals. The scale items related to this are item 1, item 2, item 10, and item 12.
- **Emotional Bond** evaluates the level of emotional resonance and mutual understanding between the counselor and the client. This aspect is measured through items 3, 5, 7, and 9 on the scale.

To ensure fairness in the evaluation process, prompts are tailored to the specific language of the dataset being assessed. Both the English and

Chinese versions of the WAI used in this study are sourced from publicly available information on the official WAI website[1].

Prompt

Please evaluate the quality of the conversation between the counselor and the user based on the given criteria.

Criteria
- Helpfulness: Assesses the applicability of explanations and suggestions from a psychological counseling perspective. Rate from 1 to 5, where 1 means not helpful at all and 5 means highly helpful.
- Coherence: Evaluates the logical flow and structure of the conversation. Rate from 1 to 5, where 1 means completely incoherent and 5 means well-structured and logically sound.
- Empathy: Measures the counselor's ability to understand and respond to the user's feelings. Rate from 1 to 5, where 1 means no empathy at all and 5 means the same level of empathy as a human counselor.
- Guidance: Evaluates the specificity and practicality of the counselor's suggestions. Rate from 1 to 5, where 1 means the suggestions are not practical at all and 5 means the suggestions are very specific and feasible.

Format:
Just provide the scores without explanations, and the scores should be integers.
Helpfulness: [score]
Coherence: [score]
Empathy: [score]
Guidance: [score]

[Conversation Record]
{diag}

Figure 12: The prompt used to evaluate score.

Prompt

The following counseling session reflects a dialogue between the client and the counselor. Please evaluate how the client might respond to the questions below after the session by selecting the appropriate scores from the provided rating scale.

Note: Please provide only the question numbers and their respective scores in the specified format. Do not repeat the questions themselves or add unnecessary prefixes or control characters.

[Question List]
1. As a result of these sessions I am clearer as to how I might be able to change.
2. What I am doing in therapy gives me new ways of looking at my problem.
3. I believe counselor likes me.
4. counselor and I collaborate on setting goals for my therapy.
5. counselor and I respect each other.
6. counselor and I are working towards mutually agreed upon goals.
7. I feel that counselor appreciates me.
8. counselor and I agree on what is important for me to work on.
9. I feel counselor cares about me even when I do things that he/she does not approve of.
10. I feel that the things I do in therapy will help me to accomplish the changes that I want.
11. counselor and I have established a good understanding of the kind of changes that would be good for me.
12. I believe the way we are working with my problem is correct.

[Rating Scale]
1: Seldom
2: Sometimes
3: Fairly Often
4: Very Often
5: Always

[Response Format]
Question number: Score

[Below is the history of the counseling dialogue]
{diag}

Figure 13: The prompt used for evaluate WAI score.

F Experiment Details

To ensure a fair comparison across datasets and minimize biases arising from model selection, we finetuned the Meta-Llama-3-8B-Instruct on the SMILE, Cactus, and MusPsy-Dataset. During training, we used LoRA for finetuning, setting the low-rank matrix dimension to 32 and the alpha to 16. Training was conducted using the Llama-Factory library, with a learning rate of 2e-4. The model was trained for 2 epochs on the SMILE, Cactus, and MusPsy-Dataset.

When evaluating dialogues with large language models, we used GPT-4o and set the temperature sampling parameter to T=0.0. Additionally, for generating responses for the LLM client and counselor, the temperature sampling parameter was set to T=0.7. During finetuning, we used different prompts for the three tasks. The three prompts we finetuned are shown below.

[1]https://wai.profhorvath.com/downloads

Furthermore, the specific prompts utilized during the finetuning process for Task 1 (Figure 14), Task 2 (Figure 15), and Task 3 (Figure 16) are presented here. During the initial counseling session, the counselor's memory contains only the client's static traits, which we propose can be provided by the client prior to the interaction.

Prompt

You now need to help me summarize and consolidate the memory of this counseling session based on the given dialogue history. Note: All generated content must remain in English.
Client Background: Includes client profiles, personal traits

Counseling Memory:
1. Client's Mental State: The client's psychological state during this session.
2. Counselor Observations: The counselor's observations and impressions of the client.
3. Counseling Assignments: Homework and tasks the client needs to complete after this session
4. Session Goal: The goal of this sessions.

Conversation Summary: Please be concise and summarize the dialogue content so far.

Summary format and example are as follows. Please ensure that spaces, indentation, and blank lines match the example:

Figure 14: The prompt used for task 1 (Memory Extraction).

Prompt

The following is the memory from previous counseling and the client's background. Please generate an appropriate goal for the next consultation. Ensure that the goal takes into account their current state, challenges, and progress made in previous sessions.

By doing this, you will provide a goal that is personalized, actionable, and consistent with the principles of Cognitive Behavioral Therapy (CBT), tailored to the unique characteristics of the client.

Figure 15: The prompt used for task 2 (Goal Planning).

Prompt

You are a counselor who has conducted multiple counseling sessions with the client. Below are the given historical session memories and the goal for this consultation. Your role is to engage in a conversation that builds upon past discussions while acknowledging the user's unique experiences and ongoing challenges.

At the start of the session, if the previous session included specific tasks or reflections for the client, begin by checking in on their progress.

Throughout the session, incorporate details about the client's personal background. Acknowledge their past reflections, struggles, and achievements to create a supportive and personalized therapeutic space.Encourage the user to explore their thoughts and emotions in depth, guiding them toward actionable strategies that align with their unique circumstances.

The goals for this consultation have been provided. Please adhere to the set objectives and ensure the user achieves the defined goals by the end of the session.

Figure 16: The prompt used for task 3 (Counseling Generation).

G LLM Client Construction

To ensure a fair and direct comparison of the counseling models, we maintain a consistent simulated client from the initial session. Recognizing that the language environment of the model can influence its responses, we address potential discrepancies by translating some other research materials originally in Chinese into English, thus ensuring alignment with the model's primary linguistic context. Furthermore, to create a more realistic and continuous interaction, we instruct the model to update its dynamic state after the completion of

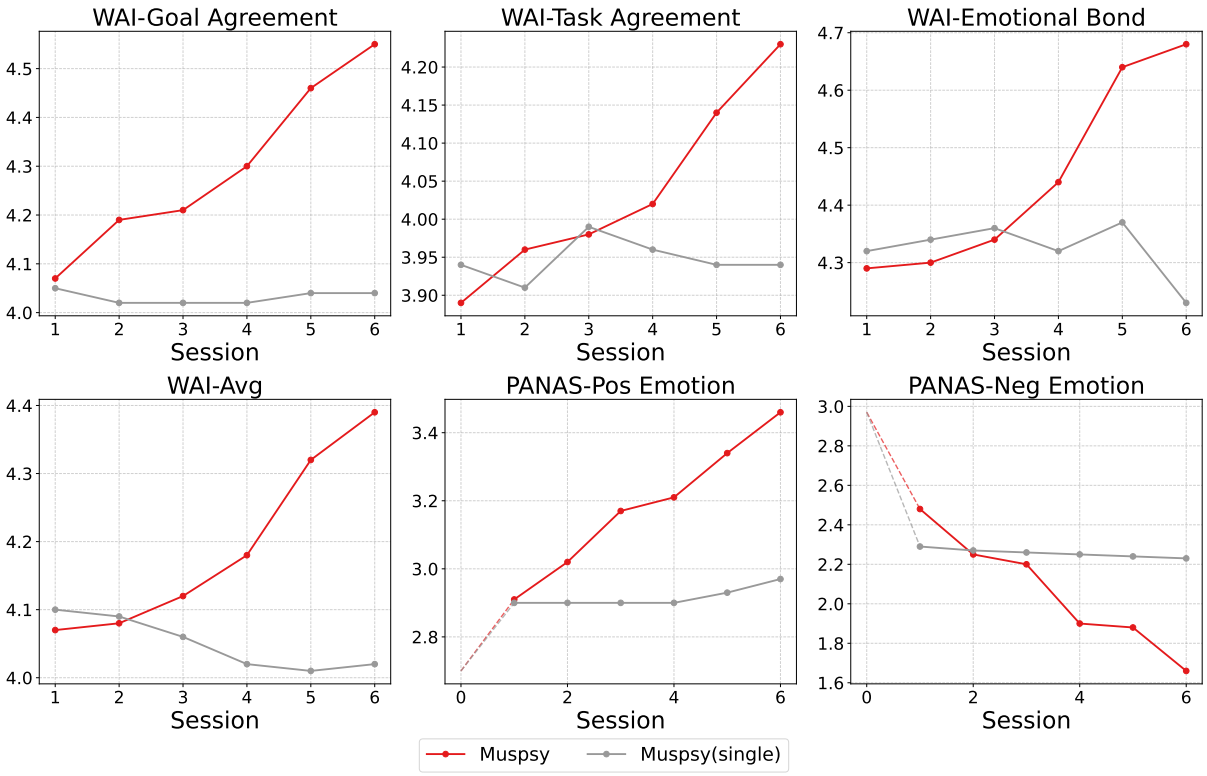

Figure 17: Emotional changes of the LLM client and performance changes of the LLM counselor across multiple sessions. The performance and trend of MusPsy-Model (single) are better than some baseline models; however, it cannot produce the same effects as multi-session models, which validates our emphasis on multi-session modeling.

each simulated session. This allows the model to retain information and context from the previous interaction, enabling a more coherent progression into the subsequent session, as visually represented in Figure 18.

Prompt

You need to play the role of the user in a multi-turn psychological counseling session with a professional counselor. The conversation should follow these rules:

Basic Guidelines:
You are the user receiving counseling, not the counselor. You should never take on the role of a counselor, advisor, or guide.
You must only respond as the user. You must not ask the counselor questions as if you were in their role.
Generate only one response per turn, keeping it natural and engaging.
If the session ends, mark it with [/END].

The Following is Your Static Traits and Dynamic States from last session, and you should play the role of the user who takes a long-term multi-turn psychological counseling:
Client Static Traits:
Client Dynamic States:

Figure 18: The prompt used for simulating client.

Following each session, we prompt the LLM to evaluate the PANAS score that a client undergoing such a session likely exhibits. The prompt we use for this evaluation is shown in Figure 19.

H Single Session Experiments

To differentiate between the contributions of the MusPsy-Dataset and the MusPsy-Model design, we present another set of experiments. These experiments compare using only Task 3 (similar to a single-session model) with the combined use of Task 1, Task 2, and Task 3. In these experiments, we only informed the model of the current counseling session number. The results show that Task 3 alone cannot track client information or dynamically adjust its counseling goals. While the model's performance significantly decreased, it still outperformed some baselines. We attribute this to the inherent design of the dataset itself; the MusPsy-Dataset naturally incorporates more advanced psychological counseling techniques and goals, making its internal content richer and its effects better

Prompt

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.

For each feeling, generate a score from 1 to 5 using the following scale:
1 - Very slightly or not at all
2 - A little
3 - Moderately
4 - Quite a bit
5 - Extremely

Additionally, please provide a brief explanation for each score. Output in the specified format without including any irrelevant control characters or prefixes.
Here is the text:
{user}

Here is the counseling session:
{history}

[Output Format]
Emotion: Score. Explanation

Figure 19: The prompt used to evaluate PANAS score.

than other datasets.

As shown in Figure 17, this demonstrates that our contributions are multifaceted, encompassing both the contribution of the dataset and the contribution of our design.

I Human Evaluation of Task3

To further validate the effectiveness of different counseling models in a multi-session setting, we conduct an additional human evaluation on counseling samples. In this setup, each counseling model interacts with eight identical virtual clients portrayed across six sequential sessions, resulting in a total of 48 distinct session transcripts per model.

These transcripts are then equally divided into four groups, where two trained professional annotators independently rate the model's performance as a counselor. The evaluation is based on the WAI, consisting of 12 items rated on a 1–5 Likert scale. These 12 items are designed to measure three key dimensions—*Task Agreement*, *Emotional Bond*, and *Goal Agreement*—with each dimension comprising four specific questions. We report the averaged scores in Table 4. Overall, MusPsy-Model achieves the best average WAI score, indicating stronger cooperation on counseling methods, improved counselor-client relationship, and better alignment on counseling objectives compared to baselines.

To measure annotation reliability, we compute inter-annotator agreement using Quadratic weighted κ on each WAI dimension. The agreement is moderate on Task and Goal, and substantial on Bond (Table 4), suggesting that the perceived relationship bonding is relatively consistent across annotators.

Model	**Task.**	**Bond.**	**Goal.**	**Avg.**
Cactus	3.89	3.51	3.91	3.77
SimPsyDial	3.83	3.75	3.79	3.79
CPsyCoun	3.78	3.46	3.75	3.66
MusPsy	**4.13**	**3.75**	**4.10**	**3.99**
Quadratic weighted κ	0.55	0.66	0.56	–

Table 4: Human evaluation results on 48 model-generated counseling samples using WAI dimensions. We also report inter-annotator agreement (quadratic weighted κ) for each dimension.

We further analyze MusPsy's session-wise WAI changes, as shown in Table 5. The scores exhibit a clear upward trend across sessions, especially in Task and Bond, which aligns with real-world counseling where rapport and cooperation are expected to strengthen over time. This trend provides additional evidence that modeling multi-session dynamics helps improve longitudinal counseling quality.

	Session						**Avg.**
Metric	1	2	3	4	5	6	
Task.	3.73	3.99	4.06	4.11	4.27	4.50	4.13
Bond.	3.55	3.53	3.59	3.67	3.89	4.27	3.75
Goal.	3.97	3.92	3.99	4.06	4.27	4.56	4.10

Table 5: Session-wise human evaluation of MusPsy using WAI dimensions. Scores generally increase over sessions, indicating strengthened working alliance.