Title

Knowledge-Informed Dynamic Strategy Adaptation Multi-Agent Framework for Psychotherapy

Permalink

https://escholarship.org/uc/item/57m9g54m

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Zhao, Lianxin
Guan, Hongjiao
Xu, Bing
et al.

Publication Date

2026

Copyright Information

Knowledge-Informed Dynamic Strategy Adaptation Multi-Agent Framework for Psychotherapy

Lianxin Zhao[1,3], Hongjiao Guan[1,3,*], Bing Xu[2], Wenpeng Lu[1,3]
[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), China
[2] Faculty of Computing, Harbin Institute of Technology, China
[3]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, China

Abstract

Psychotherapy is an inherently interactive and adaptive process, in which therapists continuously adjust intervention strategies according to clients' evolving emotional states. However, most existing LLM-based psychotherapy dialogue generation methods rely on single or fixed therapeutic techniques, failing to explicitly model the turn-level intervention decision-making. To overcome this drawback, we propose a knowledge-informed dynamic strategy adaptation multi-agent framework (K-DAF), which formulates psychotherapy as a turn-level emotion perception and adaptive strategy intervention process explicitly grounded in a psychotherapy knowledge base. K-DAF mainly contains the modules of client state tracking, knowledge-grounded strategy retrieval and recommendation, and strategy-conditioned response generation. Based on K-DAF, we construct a multi-turn psychotherapy dialogue dataset, PsyDAF, and fine-tune a psychotherapy-oriented model, DAF-Chat. Comparison experiments demonstrate the superiority of PsyDAF and DAF-Chat in terms of therapeutic alliance, empathic understanding, and counseling skills. The ablation study indicates the significance of the knowledge base and client state tracking.

Keywords: Computational Cognitive Modeling; Dynamic Strategy Adaptation; Multi-Agent; Psychotherapy; Mental Healthcare Application

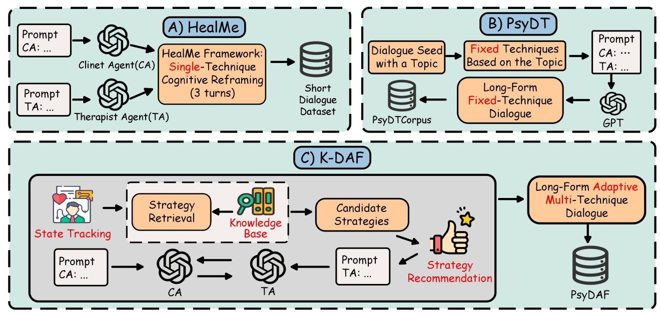

Figure 1: Comparison of different psychotherapy dialogue generation frameworks: single technique-based HealMe, fixed technique-based PsyDT, and our proposed K-DAF using multiple techniques and adaptive strategy intervention.

Introduction

High-quality psychotherapy resources have always been scarce and costly, while the population with mental health needs continues to grow. To address this challenge, recent years have seen increasing research exploring psychotherapy methods based on large language models (LLMs). These studies typically design psychotherapy dialogue generation frameworks to simulate professional therapeutic processes and construct psychotherapy dialogue datasets for training psychotherapy-oriented LLMs (Xie et al., 2025; Zhang et al., 2024).

Recent LLM-based psychotherapy approaches have demonstrated promising potential in producing fluent and empathetic psychotherapy-style responses (Kang et al., 2024). However, in the process of psychological counseling, many methods adopt single or fixed therapeutic techniques. HealMe (Xiao et al., 2024) uses a single technique of cognitive reframing to construct three-turn dialogues (Figure 1 (A)), which limits the flexibility in adapting intervention strategies to clients' evolving states. In the work of PsyDT (Xie et al., 2025), real conversations on 12 topics are collected, and the therapeutic technologies used for each topic are identified. Throughout the entire dialogue, PsyDT selects the fixed psychotherapy techniques according to the counseling topic (Figure 1 (B)), which does not explicitly model turn-level interaction dynamics. These approaches struggle to capture a fundamental characteristic of real-world psychotherapy that therapists continuously assess clients' evolving emotional or cognitive states and adapt their intervention strategies accordingly across dialogue turns (Norcross and Wampold, 2011).

From a clinical perspective, psychotherapy is inherently a highly interactive and adaptive process (Hayes and Hofmann, 2020; Newhill et al., 2003; Norcross, 2011). *Process-based psychotherapy* (Hayes and Hofmann, 2020) emphasizes continuous assessment of clients' moment-to-moment changes and flexible adjustment of interventions based on therapeutic progress. Similarly, *integrative psychotherapy* (Norcross and Goldfried, 2005) highlights the importance of dynamically selecting and combining different therapeutic techniques across the course of treatment, rather than adhering rigidly to a single technique. These perspectives suggest that effective psychotherapy cannot be reduced to a static strategy or fixed techniques, but instead requires ongoing decision-making grounded in the client's current state (Ethayarajh and Jurafsky, 2020).

To this end, we propose a knowledge-informed dynamic strategy adaptation multi-agent framework (K-DAF) for psychotherapy dialogue generation (Figure 1 (C)). Grounded in mature psychotherapy theories and developed under the guid-

* Corresponding author: guanhongjiao2008@163.com

ance of professional therapists, K-DAF integrates multiple psychotherapy techniques to build a knowledge base with a wealth of strategies, and guides the dialogue generation in a manner of dynamic emotion perception, continuous state tracking, and adaptive strategy response. Specifically, the *state tracking module* tracks the client's affective or cognitive shifts and deduces the intervention intent, based on which the *strategy retrieval module* retrieves candidate intervention strategies from the knowledge base. Then, the therapist agent generates natural language responses consistent with the therapeutic strategy selected by the *strategy recommendation module*. Based on K-DAF, we construct a psychotherapy dialogue dataset, PsyDAF, and fine-tune Llama-3.1-8B-Instruct (Grattafiori et al., 2024) on this dataset to obtain a psychotherapy-oriented model, DAF-Chat. Experimental results show that PsyDAF outperforms baseline datasets and DAF-Chat demonstrates stronger clinical consistency and adaptability compared with baseline models.

Our contributions are summarized as follows:

(1) We propose K-DAF, a psychotherapy dialogue generation framework, which models turn-level emotion perception and adaptive strategy intervention to enhance the flexibility and interpretability of psychological counseling. The ablation study indicates the significance of the knowledge base and state tracking module in K-DAF.

(2) Based on the K-DAF framework, we construct a psychotherapy dialogue dataset, PsyDAF. Experimental results show that PsyDAF consistently outperforms baseline datasets in terms of three scales (WAI, BLRI, and CCS-R).

(3) Using the PsyDAF dataset, we fine-tune a psychotherapy-oriented model, DAF-Chat. On the same three scales, DAF-Chat demonstrates stable and consistent improvements over baseline models, exhibiting stronger clinical coherence and adaptability.

Related Work

In recent years, researchers have begun to explore the application of LLMs to psychotherapy and counseling scenarios (Y. Chen et al., 2023; Qiu et al., 2024), typically by constructing large-scale synthetic psychotherapy dialogue datasets for model fine-tuning (Z. Chen et al., 2025; Xie et al., 2025; Zhang et al., 2024). This can be traced back to earlier work on emotional support conversations (ESC). For instance, ESConv (S. Liu et al., 2021) is a manually annotated multi-turn emotional support dialogue dataset, which provides foundational resources for modeling supportive interactions. Building on this, ESCoT (Zhang et al., 2024) constructs a new dataset, incorporating a broader set of counseling strategies.

Recent studies have proposed various LLM-based dialogue generation frameworks for psychotherapy and counseling. A common approach formulates dialogue generation as a prompt-conditioned narrative task, in which client profiles, therapeutic knowledge, and system instructions are compressed into a single prompt and fed into a single LLM to generate complete counseling conversations (Z. Chen et al., 2025; Xie et al., 2025). These methods blur the distinction between genuine therapeutic interactions and scripted dialogues, without explicitly modeling turn-level interaction dynamics. For example, PsyDT (Xie et al., 2025) selects fixed therapeutic techniques based on the counseling topic and fed them into the LLM prompt, along with the client's and therapist's instructions. In addition, HealMe (Xiao et al., 2024) adopts a fixed three-stage therapeutic technique of cognitive reframing to generate the dialogue dataset.

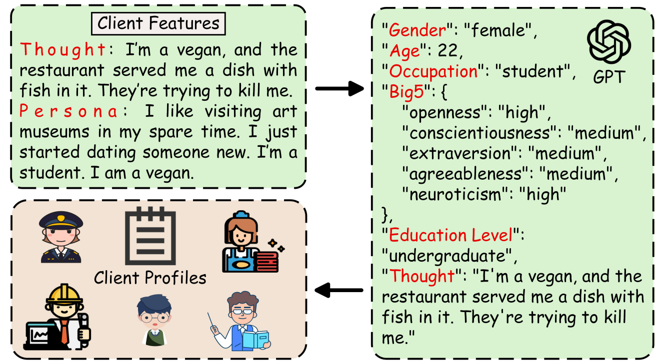

Figure 2: Client profile generation using GPT-3.5-turbo based on the extracted client features.

These methods mentioned above neglect the nature of high interactivity and intervention adaptability in real-world psychological counseling scenarios. Following the principle ideas of process-based psychotherapy and integrative psychotherapy, we propose the K-DAF framework, which integrates multiple psychotherapy techniques and models turn-level interaction dynamics.

Methodology

Client Profile Generation

To avoid the use of real personal data, we sample 2,000 instances from the manually annotated PATTERNREFRAME dataset (Maddela et al., 2023) and extract client features (e.g., core thoughts and persona cues) to construct client profiles using GPT-3.5-turbo, including demographic attributes and inferred personality traits. To improve the quality and diversity of the generated profiles, a professional therapist manually curate a small set of high-quality profiles, which are incorporated into the prompt for few-shot learning. The client profile generation process is illustrated in Figure 2.

The K-DAF Framework

The diagram of the proposed knowledge-informed dynamic strategy adaptation multi-agent framework (K-DAF) for psychotherapy dialogue generation is illustrated in Figure 3.

Client Simulation Module (CSM) As shown in Figure 3(A), we feed the generated client profiles into GPT-3.5-turbo to simulate a consistent client role. To enhance the generation quality, two constraints are used in the client agent.

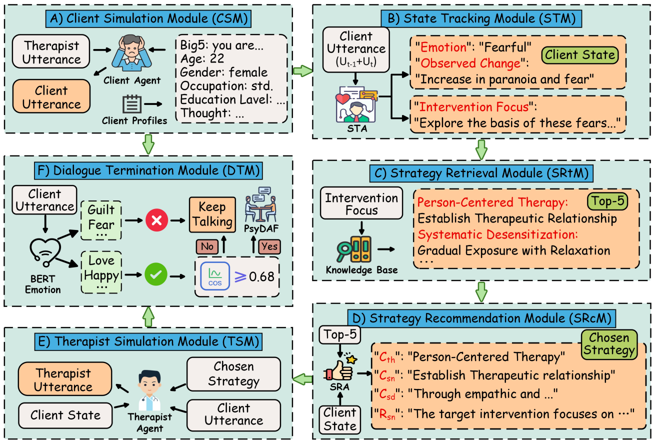

Figure 3: The K-DAF framework for psychotherapy dialogue generation. The client profile is fed into the client agent in the CSM (A). The STM tracks the client's affective or cognitive shifts and deduces the intervention intent (B), based on which the SRtM retrieves candidate intervention strategies from the knowledge base (C). Then, the therapist agent generates natural language responses (E) consistent with the recommended therapeutic strategy in the SRcM (D). This interaction process between the client agent and the therapist agent continues until the termination condition is met (F).

Contextualized Initialization: For the initial turn, we employ a specialized *client 1st round prompt* that mandates the client agent to present a concrete and scenario-based problem rather than a vague emotional statement. This ensures the subsequent therapeutic dialogue is anchored in specific scenarios.

Emotional Progression Control: Starting from the second turn, the module operates under a "slow-recovery" heuristic. The prompt constrains the client agent to maintain high levels of anxiety and resistance in the early phases, allowing for a gradual emotional shift only in response to effective interventions (Hayes et al., 2020). This prevents unrealistic "instant healing" and ensures the coherence and clinical depth of the multi-turn interaction.

State Tracking Module (STM) As shown in Figure 3(B), at each dialogue turn, the State Tracking Agent (STA) analyzes both the client's current utterance (u_t) and the preceding utterance (u_{t-1}). Rather than performing simple emotion classification, the STA infers three signals: (1) the client's current emotional state (Emotion), (2) cognitive or affective changes relative to the previous turn (Observed Change), and (3) an abstract intervention intent that guides the next response of the therapist agent (Intervention Focus). Unlike previous methods based on fixed psychotherapy techniques, the STA explicitly represents the client's affective state and turn-level changes, and specifies the intended intervention direction. Therefore, this module provides a foundation to support process-aware and interpretable clinical intervention (Zhao et al., 2025).

Strategy Retrieval Module (SRtM) As shown in Figure 3(C), the Intervention Focus inferred based on the Emotion and Observed Change is used to retrieve candidate intervention strategies from the psychotherapy knowledge base (KB).

Technique: Object-Relations Theory

Treatment Stages (Strategies):

1. Exploration of Internalized Relationships: By exploring early relational experiences, help the seeker become aware of internalized relationship patterns and how they are reenacted in current relationships.
2. Recognition of Transference: Through reflection on emotional reactions within the therapeutic relationship, help the seeker recognize transference linked to early attachment experiences.
3. Identification of Splitting and Projection: By identifying patterns of idealization, devaluation, and projection, help the seeker develop a more continuous and integrated view of self and others.
4. Integration of Self and Object: Through stable and corrective relational experiences, help the seeker form integrated, secure internal representations of self and others.

Figure 4: An example of specific psychotherapy technique with its treatment stages and detailed descriptions from the knowledge base.

Knowledge Base Construction: With reference to classical psychological literature (Corey, 2013; Norcross and Goldfried, 2005; Prochaska and Norcross, 2018) and under the guidance of professional therapists, we select 10 common psychotherapy techniques (e.g., Cognitive Behavioral Therapy, Object Relations Theory, and Person-Centered Therapy) and further decompose each technique into multiple treatment stages (Kazdin, 2007; Prochaska and DiClemente, 1983). Each treatment stage, combined with its belonging psychotherapy technique, is regarded as a candidate intervention strategy, and all these intervention strategies constitute the psychotherapy knowledge base (KB) in our framework. Figure 4 presents a psychotherapy technique of Object Relations Theory (Kernberg, 1995), along with detailed descriptions of its constituent treatment stages.

The Intervention Focus and KB strategies are encoded using the all-MiniLM-L6-v2 sentence embedding model (W. Wang et al., 2020) for vector-based retrieval. In each dialogue turn, the SRtM selects the top-5 candidate strategies most similar to the current Intervention Focus. These retrieved strategies are then passed to the Strategy Recommendation Agent (SRA) for final strategy recommendation.

Strategy Recommendation Module (SRcM) As shown in Figure 3(D), the Strategy Recommendation Agent (SRA) considers the client's state, including Emotion and Observed Change obtained in the STM, and selects the most appropriate intervention strategy from the top-5 candidate strategies returned by the SRtM. In addition, the SRA is required to generate a brief clinical rationale to explicate the reasoning behind the selected strategy, thereby enhancing the interpretability and clinical transparency of the decision-making process. The selected psychotherapy technique, stage name, stage description, and the corresponding rationale, denoted as C_{th}, C_{sn}, C_{sd}, and R_{sn} respectively, are then concatenated into a coherent natural-language statement and injected into the therapist agent's prompt to guide response generation.

To assess the rationality of the SRA, we randomly sampled 150 cases and invited two professional therapists to select the most appropriate strategy under the same conditions. The SRA achieved a high average agreement of 82% with human judgment. In cases of disagreement, the therapists judged the SRA's rationales to be clinically reasonable. Accordingly, the intervention strategies recommended by the SRA demonstrate high clinical credibility and strong interpretability.

Therapist Simulation Module (TSM) As shown in Figure 3(E), the Therapist Agent functions as an empathetic expert agent that synthesizes multiple data streams to generate clinically grounded therapist responses. To ensure high-fidelity and professional interventions, it adopts three core response principles as follows.

Knowledge-Informed Synthesis: The TSM integrates the following information into a specialized therapist prompt: the client's most recent utterance in the CSM, the client's state in the STM, the selected strategy (C_{th}, C_{sn}, and C_{sd}) and the clinical rationale (R_{sn}) in the SRcM. This design encourages the therapist agent's responses to remain consistent with the selected strategy.

Affective and Linguistic Alignment: The response is required to maintain an empathetic, warm, and supportive tone, incorporate therapeutic principles naturally, and refrain from using technical jargon.

Efficiency and Flow Control: Each response is limited to fewer than 40 words to prevent verbosity and preserve the natural pacing of real conversations.

Finally, the therapist agent's utterance is fed back to the CSM and then the next turn of dialogue begins. The interaction process will continue to capture the longitudinal evolution of the client's state until the termination criteria are met.

Dialogue Termination Module (DTM) To achieve natural and reliable dialogue termination, we adopt a dual-criterion mechanism that combines emotional state filtering and segment-level semantic matching (Figure 3(F)). A BERT-based emotion classifier (Devlin et al., 2019) is first used as a safety constraint to suppress conversation termination. Specifically, if the client's emotion is classified as strong negative affect (e.g., guilt, fear, sadness, anger, or shame), the conversation continues. Otherwise, the client's response is segmented into independent segments to match with predefined termination intent expressions (e.g., "I feel ready to end for today) using cosine similarity based on all-MiniLM-L6-v2 sentence embeddings. Dialogue termination is triggered when the maximum similarity exceeds 0.68, which is empirically determined through multiple rounds of testing to balance termination reasonableness and conversational naturalness. Note that response segmentation is to prevent long or compound sentences from diluting implicit termination signals (N. F. Liu et al., 2024; Niklaus et al., 2019).

Dataset Construction and Model Fine-Tuning

Based on the K-DAF framework, we construct a psychotherapy dialogue dataset named PsyDAF, consisting of 2,000 multi-turn conversations, which provides the basis for subsequent model training and evaluation.

Table 1: Basic Statistics of Psychotherapy Dialogue Datasets

Dataset	**#Samples**	**Avg. Turns**
ESConv (S. Liu et al., 2021)	1300	12.74
ESCoT-D (Zhang et al., 2024)	1195	11.76
PsyDTCorpus (Xie et al., 2025)	4760	37.16
PsyDAF (Ours)	2000	33.93

The PsyDAF dataset is split into training and validation sets at a ratio of 80% to 20%. Fine-tuning is performed on Llama-3.1-8B-Instruct using LlamaFactory (Y. Zheng et al., 2024) with LoRA (Hu et al., 2022). The model is trained for 4 epochs on two NVIDIA RTX 3090 GPUs, which takes approximately 5 hours.

Experiments

Evaluation Metrics

We adopt the evaluation framework proposed in (B. Wang et al., 2026), which transforms three psychotherapy scales into standardized metrics suitable for both human evaluation and automatic GPT-based evaluation. Specifically, the Working Alliance Inventory (WAI) (Horvath and Greenberg, 1989), scoring from 1 to 5, is used to measure therapeutic relationship quality, including goal agreement, task collaboration, and emotional bond between the therapist and the client. The Empathic Understanding subscale of the Barrett–Lennard Relationship Inventory (BLRI) (Barrett-Lennard, 1962), scoring from –3 to 3, assesses empathic understanding, capturing both cognitive and affective components of empathy. The Counselor Competencies Scale–Revised (CCS-R) (Lambie et al., 2018), scoring from 1 to 5, measures counseling skills, covering a broad range of professional dispositions related to therapeutic technique and interpersonal effectiveness.

Dataset Evaluation

The proposed PsyDAF dataset is compared with three previous multi-turn long-form psychotherapy dialogue datasets, including ESConv (S. Liu et al., 2021), the ESCoT dataset (ESCoT-D) (Zhang et al., 2024), and PsyDTCorpus (Xie et al., 2025). The basic statistics of these datasets are summarized in Table 1. Following common practice in prior work (B. Wang et al., 2026; L. Zheng et al., 2023), we randomly sample 150 dialogue instances from each dataset and conduct automatic evaluation using GPT-4o, aiming to achieve a representative and cost-efficient comparison.

Results and Analysis Table 2 reports the evaluation results of the four datasets across three psychotherapy dimensions: Therapeutic Relationship (WAI), Empathic Understanding (BLRI), and Counseling Skills (CCS-R). On the WAI metric, PsyDAF attains the highest average WAI score (4.18), ranking first on both Task (3.95) and Bond (4.76), indicating stronger alliance maintenance and trust building. Although PsyDTCorpus scores highest on Goal (3.92), its overall WAI

Table 2: Evaluation Scores of Different Datasets across Three Psychotherapy Metrics. **Bold** denotes the best result, and † the second best.

Dataset	WAI				BLRI					CCS-R				
	Goal	Task	Bond	Avg.	Cogn.	Affect.	Differ.	Inner	Avg.	Probe	Environ	Reflect	Change	Avg.
ESConv	2.36	2.76	3.30	2.81	1.39	1.44	1.47	1.52	1.46	2.79	2.56	2.21	1.78	2.33
ESCoT-D	3.27	3.21	3.92	3.47	2.22	2.32†	2.22	2.35†	2.28†	3.80	3.50	3.29	2.77	3.34
PsyDTCorpus	**3.92**	3.93†	4.40†	4.08†	2.24†	2.24	2.25†	2.25	2.25	4.25†	4.23†	4.08†	3.58†	4.04†
PsyDAF	3.84†	**3.95**	**4.76**	**4.18**	**2.49**	**2.66**	**2.66**	**2.67**	**2.62**	**4.53**	**4.56**	**4.42**	**3.67**	**4.29**

Table 3: Human Evaluation Results of Datasets

Dataset	WAI	BLRI	CCS-R
ESConv	2.85	1.60	2.41
ESCoT-D	3.34	2.23	3.54
PsyDTCorpus	4.01	2.41	3.98
PsyDAF	**4.09**	**2.53**	**4.21**

Table 4: Ablation Results of Framework Components

Variant	WAI	BLRI	CCS-R
K-DAF	**4.21**	**2.70**	**4.18**
w/o KB	4.08	2.66	3.91
w/o STM	4.08	2.70	3.98
w/o SRcM	4.10	2.70	3.98

is slightly lower. For BLRI, PsyDAF leads all BLRI sub-dimensions with the highest overall average (2.62), reflecting stronger cognitive or affective empathy and more accurate understanding of clients' internal experiences. On CCS-R, PsyDAF also achieves the highest scores, with an overall average of 4.29. PsyDTCorpus ranks second, while ESConv and ESCoT-D exhibit weaker and less stable counseling skills.

Overall, PsyDAF consistently outperforms baseline datasets across therapeutic relationship, empathic understanding, and counseling skills, demonstrating the effectiveness of the K-DAF framework in generating more adaptive and high-quality psychotherapy dialogues.

Human Evaluation To assess the reliability of GPT-4o for automatic evaluation, we invite two expert evaluators to independently assess the same 150 samples evaluated by GPT-4o for each psychotherapy dialogue dataset. The agreement between GPT-4o and human evaluators is measured using Spearman's rank correlation coefficient (Poria et al., 2017), yielding an average correlation of 0.64, which indicates a strong positive correlation (Dancey and Reidy, 2007). Note that the correlation coefficient in the interval of 0.61–0.80 indicates a strong correlation. These results indicate that GPT-4o has a high correlation with human evaluators and can serve as a reliable evaluator. Human evaluation results are shown in Table 3, where PsyDAF outperforms all baseline datasets in terms of the three scales. Due to space limitations, only the average scores of the main dimensions are reported.

Ablation Study

To analyze the contribution of core components, we conduct ablation studies by removing the knowledge base (KB), the state tracking module (STM), and the strategy recommendation module (SRcM), respectively, with GPT-4o used as the automatic evaluator. As shown in Table 4, removing the KB leads to the largest performance drop, particularly on CCS-R, highlighting the importance of external psychotherapy knowledge for effective interventions. Removing STM results in moderate degradation on WAI and CCS-R, indicating the value of explicit client state modeling for improving dialogue quality. Removing SRcM also causes slight performance declines, suggesting that reasoning over candidate strategies improves intervention appropriateness. In contrast, BLRI remains substantially stable across all variants, suggesting that empathic expression and language fluency mainly depend on the base language model and carefully designed prompts (Rashkin et al., 2019). To sum up, these results demonstrate that each module contributes to dialogue quality, with the KB and STM playing especially important roles in professional and adaptive psychotherapy.

Model Evaluation

In this section, the proposed DAF-Chat and previous psychotherapy baseline models are evaluated by simulating the interaction process between clients and therapists to assess the quality of counseling dialogues. To this end, 150 client profiles are generated using the method described in the section of Client Profile Generation and fed into GPT-3.5-turbo, which serves as the client. Note that the 150 client profiles are different from those used to generate 2000 dialogues. DAF-Chat and the baseline models serve as the therapists. The simulated dialogues are evaluated by GPT-4o.

Baseline Models SoulChat2.0 (Xie et al., 2025) is an open-source model, obtained by fine-tuning Llama-3.1-8B-Instruct on PsyDTCorpus. Zhang et al. (Zhang et al., 2024) proposed the ESCoT dataset (ESCoT-D); however, they did not release the corresponding trained model. To ensure fair comparison, we fine-tune Llama-3.1-8B-Instruct using ESCoT-D and obtain the model called ESCoT-M. Note that Llama-3.1-8B-Instruct itself is also included as a baseline model. All models are fine-tuned using LlamaFactory with LoRA.

Table 5: Evaluation Scores of Different Models across Three Psychotherapy Metrics. **Bold** denotes the best result, and † the second best.

Model	WAI				BLRI					CCS-R				
	Goal	Task	Bond	Avg.	Cogn.	Affect.	Differ.	Inner	Avg.	Probe	Environ	Reflect	Change	Avg.
Llama-3.1-8B (Grattafiori et al., 2024)	3.92	3.96	4.40	4.09	2.51	2.65	2.67	2.70	2.63	4.10	4.32	4.07	3.49	4.00
SoulChat2.0 (Xie et al., 2025)	4.06	4.18	4.70†	4.31	2.57	2.77	2.74†	2.80	2.72	4.42	4.47	4.34	**3.80**	4.26
ESCoT-M (Zhang et al., 2024)	**4.28**	4.36†	4.70†	4.45†	2.63†	2.81†	2.72	2.84†	2.75†	4.48†	4.52†	4.43†	3.75	4.30†
DAF-Chat (Ours)	4.08†	**4.46**	**4.83**	**4.46**	**2.74**	**2.96**	**2.93**	**2.97**	**2.90**	**4.58**	**4.74**	**4.55**	3.79†	**4.40**

Table 6: Human Evaluation Results of Models

Model	WAI	BLRI	CCS-R
Llama-3.1-8B	3.58	2.62	3.63
SoulChat2.0	4.33	2.7	4.23
ESCoT-M	4.39	2.68	4.32
DAF-Chat	**4.44**	**2.78**	**4.51**

Results and Analysis Table 5 reports the results of different models across three psychotherapy-related metrics (WAI, BLRI, and CCS-R), which are similar to those of evaluating psychotherapy dialogue datasets in Table 2. DAF-Chat achieves the highest average scores on the three metrics, which indicates that DAF-Chat performs the best in terms of therapeutic alliance, empathic understanding, and counseling skills, compared to the baseline models. This can be attributed to the effectiveness of modeling psychotherapy as a turn-level strategy-adaptive decision process, which enables the generation of clinically coherent and adaptive responses.

An interesting phenomenon is that PsyDTCorpus achieves higher scores than ESCoT-D, whereas their corresponding model performance shows the reverse results (ESCoT-M outperforms SoulChat2.0). This indicates that dataset quality alone does not necessarily transfer into better model performance. The modeling paradigm of the psychotherapy dialogue generation framework may be more crucial (Peng et al., 2017; Zelikman et al., 2022). The ESCoT framework simulates the thinking process of human emotional support and constructs a reasoning chain. However, ESCoT integrates the emotional support and reasoning chain into a prompt, without explicitly separating the client and the therapist. In contrast, both PsyDAF and DAF-Chat perform best, which can be attributed to the K-DAF framework's turn-level state tracking of the client and strategy-adaptive intervention of the therapist. Human evaluation results (Table 6) show that DAF-Chat outperforms all baseline models.

Limitations and Ethics Considerations

Limitations: Large-scale real human–model interaction experiments are not included in this study. Although small-scale human studies could be conducted, limited sample sizes often lack sufficient statistical power to reliably assess dialogue quality and therapeutic effectiveness. In practice, large-scale human evaluation in psychotherapy settings is difficult due to participant recruitment challenges, high financial and time costs, and ethical constraints. Several recent studies, such as SocialSim (Z. Chen et al., 2025) and PsyDT (Xie et al., 2025), similarly exclude real human subjects, reflecting a broader challenge in this domain. Future work will explore ethically approved and scalable human-in-the-loop evaluation protocols to validate real-world applicability.

Ethics Considerations: This study is conducted using publicly available datasets and does not involve any personal or sensitive information. All participants involved in this study participated voluntarily and were fully informed that they could withdraw from the study at any time without any negative consequences. To mitigate the risk of generating potentially harmful or unsafe content in psychotherapy dialogues, we apply strict keyword- and rule-based filtering mechanisms during dataset construction. These filters are designed to identify and remove content related to high-risk topics, including self-harm, suicidal ideation and violence, thereby reducing the likelihood of exposing users to harmful material. In real-world deployment scenarios, additional high-risk content detection mechanisms (e.g., continuous keyword- and rule-based monitoring) should be employed to actively monitor both model outputs and user utterances. When potential risks are detected, the corresponding cases should be promptly escalated and referred to qualified mental health professionals for appropriate assessment and intervention, in order to ensure user safety and minimize potential harm.

Conclusion

In this paper, we propose a knowledge-informed dynamic strategy adaptation multi-agent framework (K-DAF) to overcome the shortcoming of previous methods using single or fixed psychotherapy techniques in the process of psychological counseling. K-DAF models psychotherapy as a turn-level emotion perception and adaptive strategy intervention process. Based on this framework, we construct the PsyDAF dataset and fine-tune the DAF-Chat psychotherapy-oriented dialogue model. Extensive automatic and human evaluation results over psychotherapy datasets and models demonstrate PsyDAF's and DAF-Chat's strong performance in therapeutic alliance, empathy, and counseling skills, indicating the potential of K-DAF for generating clinically coherent and adaptive psychotherapy dialogues in real-world settings. The ablation study indicates that the knowledge base with multiple psychotherapy techniques and the client's state tracking are critically significant for the effectiveness of the K-DAF framework.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (No.72501151, No.62376130), the Taishan Scholars Program (No.TSPD20240814), the Program of New Twenty Policies for Universities of Jinan (No.202333008), the Pilot Project for Integrated Innovation of Science, Education, Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2025ZDZX01), the Shandong Talent Introduction Program (No.WSR2025005), and the Program of Innovation Improvement for Small and Medium-sized Enterprises of Shandong (2024TSGC0094).

References

Barrett-Lennard, G. T. (1962). Dimensions of therapist response as causal factors in therapeutic change. *Psychological Monographs: General and Applied*, *76*(43), 1.

Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q., & Xu, X. (2023). SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1170–1183.

Chen, Z., Cao, Y., Bi, G., Wu, J., Zhou, J., Xiao, X., Chen, S., Wang, H., & Huang, M. (2025). SocialSim: Towards socialized simulation of emotional support conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(2), 1274–1282.

Corey, G. (2013). *Theory and practice of counseling and psychotherapy*. Brooks/Cole Cengage Learning.

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson Education.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Ethayarajh, K., & Jurafsky, D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4846–4853.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hayes, S. C., & Hofmann, S. G. (2020). *Beyond the DSM: Toward a process-based alternative for diagnosis and mental health treatment*. New Harbinger Publications.

Hayes, S. C., Hofmann, S. G., & Ciarrochi, J. (2020). A process-based approach to psychological diagnosis and treatment: The conceptual and treatment utility of an extended evolutionary meta model. *Clinical Psychology Review*, *82*, 101908.

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of Counseling Psychology*, *36*(2), 223.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, *1*(2), 3.

Kang, D., Kim, S., Kwon, T., Moon, S., Cho, H., Yu, Y., Lee, D., & Yeo, J. (2024). Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 15232–15261.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology*, *3*(1), 1–27.

Kernberg, O. F. (1995). *Object relations theory and clinical psychoanalysis*. Jason Aronson.

Lambie, G. W., Mullen, P. R., Swank, J. M., & Blount, A. (2018). The counseling competencies scale: Validation and refinement. *Measurement and Evaluation in Counseling and Development*, *51*(1), 1–15.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173.

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., & Huang, M. (2021). Towards emotional support dialog systems. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 3469–3483.

Maddela, M., Ung, M., Xu, J., Madotto, A., Foran, H., & Boureau, Y.-L. (2023). Training models to generate, recognize, and reframe unhelpful thoughts. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13641–13660.

Newhill, C. E., Safran, J. D., & Muran, J. C. (2003). *Negotiating the therapeutic alliance: A relational treatment guide*. Guilford Press.

Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2019). Transforming complex sentences into a semantic hierarchy. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3415–3427.

Norcross, J. C. (2011). *Psychotherapy relationships that work: Evidence-based responsiveness*. Oxford University Press.

Norcross, J. C., & Goldfried, M. R. (2005). *Handbook of psychotherapy integration*. Oxford University Press.

Norcross, J. C., & Wampold, B. E. (2011). Evidence-based therapy relationships: Research conclusions and clinical practices. *Psychotherapy*, *48*(1), 98.

Peng, B., Li, X., Li, L., Gao, J., Celikyilmaz, A., Lee, S., & Wong, K.-F. (2017). Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2231–2240.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125.

Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: Toward an integrative model of change. *Journal of Consulting and Clinical Psychology*, *51*(3), 390.

Prochaska, J. O., & Norcross, J. C. (2018). *Systems of psychotherapy: A transtheoretical analysis*. Oxford University Press.

Qiu, H., He, H., Zhang, S., Li, A., & Lan, Z. (2024). Smile: Single-turn to multi-turn inclusive language expansion via ChatGpt for mental health support. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 615–636.

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381.

Wang, B., Sun, Y., Wang, J., Yang, H., Fu, X., Zhao, Y., Wei, S., Wang, S., & Qin, B. (2026). Care-bench: A benchmark of diverse client simulations guided by expert principles for evaluating llms in psychological counseling. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(46), 39378–39386.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, *33*, 5776–5788.

Xiao, M., Xie, Q., Kuang, Z., Liu, Z., Yang, K., Peng, M., Han, W., & Huang, J. (2024). HealMe: Harnessing cognitive reframing in large language models for psychotherapy. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1707–1725.

Xie, H., Chen, Y., Xing, X., Lin, J., & Xu, X. (2025). PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 1081–1115.

Zelikman, E., Wu, Y., Mu, J., & Goodman, N. (2022). STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, *35*, 15476–15488.

Zhang, T., Zhang, X., Zhao, J., Zhou, L., & Jin, Q. (2024). ESCoT: Towards interpretable emotional support dialogue systems. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 13395–13412.

Zhao, X., Lu, W., Zheng, C., Zhang, W., Si, J., & Zhou, D. (2025). Plan dynamically, express rhetorically: A debate-driven rhetorical framework for argumentative writing. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 9551–9573.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging LLM-as-a-Judge with MT-bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, *36*, 46595–46623.

Zheng, Y., Zhang, R., Zhang, J., Ye, Y., & Luo, Z. (2024). Llamafactory: Unified efficient fine-tuning of 100+ language models. *Proceedings of the 62nd annual meeting of the association for computational linguistics*, 400–410.