MAGneT: Coordinated Multi-Agent Generation of Synthetic Multi-Turn Mental Health Counseling Sessions

Anonymous ACL submission

Abstract

The growing demand for scalable psychological counseling highlights the need for high-quality, privacy-compliant data, yet such data remains scarce. Here we introduce MAGneT, a novel multi-agent framework for synthetic psychological counseling session generation that decomposes counselor response generation into coordinated sub-tasks handled by specialized LLM agents, each modeling a key psychological technique. Unlike prior single-agent approaches, MAGneT better captures the structure and nuance of real counseling. We further propose a unified evaluation framework that consolidates diverse automatic metrics and expands expert assessment from four to nine counseling dimensions, thus addressing inconsistencies in prior evaluation protocols. Empirically, MAGneT substantially outperforms existing methods: experts prefer MAGneT-generated sessions in 77.2% of cases, and sessions generated by MAGneT yield 3.2% higher general counseling skills and 4.3% higher CBT-specific skills on cognitive therapy rating scale (CTRS). A open source Llama3-8B-Instruct model fine-tuned on MAGneT-generated data also outperforms models fine-tuned using baseline synthetic datasets by 6.9% on average on CTRS.[1]

1 Introduction

Mental health issues are increasingly prevalent, affecting 1 in 7 people worldwide [2]. This indicates an urgent need for scalable and accessible mental health counseling solutions. However, the growing demand for psychological support far outpaces the availability of trained professionals, leaving many without access to necessary care (Kazdin, 2021).

Recently, there has been a growing interest in using Large Language Models (LLMs) for mental health counseling (hereafter, referred to as counseling). While closed-source models like ChatGPT show promising conversational and psychological capabilities (Raile, 2024; Moëll, 2024), their practical use is limited by privacy concerns and weaker performance on counseling-specific tasks (Lee et al., 2024; Chiu et al., 2024; Demszky et al., 2023). Open-source LLMs fine-tuned on counseling data offer an alternative, but such data is scarce due to privacy constraints. Solutions such as manual de-identification or automatic pseudonymization (Tang et al., 2019; Yue and Zhou, 2020) remain limited in scalability and robustness.

Synthetic data generation offers a scalable, privacy-preserving solution for fine-tuning LLMs for counseling. Early work focused on single-turn therapeutic responses (Sharma et al., 2023; Sun et al., 2021; Liu et al., 2023), and subsequent works extended to multi-turn interactions using single-turn Q&A datasets (Chen et al., 2023; Qiu et al., 2024) or role-playing LLMs (Qiu and Lan, 2024; De Duro et al., 2025). However, these systems lack grounding in established psychology theory. To address this gap, some prior works leverage structured psychological inputs (Zhang et al., 2024; Vu et al., 2025; Anonymous, 2026a)[3], while CACTUS (Lee et al., 2024) incorporates a Cognitive Behavioral Therapy (CBT) based planning agent. Yet, these approaches rely on a single agent to generate the counselor's response, which is insufficient for modeling complex therapeutic strategies such as reflection, questioning, solution provision, normalization, and psycho-education (Chiu et al., 2024). While multi-agent systems have been explored for counseling tasks, existing works focus on generating single-turn supportive responses (Chen and Liu, 2025), or structured diagnostic interviews (Ozgun et al., 2025), failing to capture the open-ended, dynamic nature of counseling sessions, or address session-level planning without turn-level psychological grounding (Chen et al., 2025).

[1] Our dataset and code are attached in the appendix

[2] WHO (2025)

[3] Anonymous papers can be found in the Data Appendix

Method	CBT	Multi-Agent	Evaluation Framework				
			Diversity	CTRS	WAI	PANAS	Expert Evaluation
SMILE (Qiu et al., 2024)	✗	✗	✓	✗	✗	✗	1 Aspect
Psych8k (Liu et al., 2023)	✗	✗	✗	✗	✗	✗	✗
CPsyCoun (Zhang et al., 2024)	✗	✗	✓	✗	✗	✗	✗
Qiu and Lan (2024)	✗	✗	✓	✗	✓	✗	1 Aspect
CACTUS (Lee et al., 2024)	✓	✗	✓	✓	✗	✓	4 Aspects
MAGneT	✓	✓	✓	✓	✓	✓	9 Aspects

Table 1: MAGneT and our unified evaluation compared to prior works in synthetic counseling session generation.

To address these limitations, we introduce MAGneT, a multi-agent framework that decomposes counselor response generation into coordinated subtasks. The framework includes five specialized response agents, each aligned with core therapeutic strategies described in prior literature (Chiu et al., 2024; Lee et al., 2019; Cao et al., 2019): reflection, questioning, solution provision, normalization, and psycho-education. Their outputs are integrated by a final response generation agent responsible for producing a coherent, contextually appropriate counselor utterance. The final response generation agent is guided by two controllers: a turn-level technique selector agent and a session-level CBT-based planning agent. On the client side, we simulate realistic behavior via detailed profiles and attitude modeling. MAGneT enables scalable generation of multi-turn, psychologically grounded synthetic counseling sessions via client-counselor role-play.

Another persistent challenge in this domain is the lack of standardized evaluation. Prior works use inconsistent evaluation metrics – CACTUS (Lee et al., 2024) uses Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988) and Cognitive Therapy Rating Scale (CTRS) (Goldberg et al., 2020), while other works (Qiu and Lan, 2024) use Working Alliance Inventory (WAI) (Horvath and Greenberg, 1989), making it difficult to compare the effectiveness of synthetic data generated using different methods. A similar pattern is observed in expert evaluation as well. We address this gap by proposing a unified evaluation framework that consolidates these metrics and expands expert assessment from four to nine counseling aspects, enabling a more rigorous and comprehensive evaluation. Through this evaluation, we demonstrate that experts prefer MAGneT-generated sessions in 77.2% of cases across the nine aspects compared to those produced by the strongest baseline. On the automatic metrics, MAGneT-generated sessions outperforms those produced by the current state-of-the-art methods (Lee et al., 2024; Liu et al., 2023) by 3.2% on general counseling skills and 4.3% on CBT-specific skills on average on CTRS. Furthermore, a Llama3-8B-Instruct model (Meta, 2024) fine-tuned on MAGneT-generated data outperforms those fine-tuned on existing synthetic datasets by 6.8% on average on CTRS. Table 1 summarizes the novelty of our work relative to prior work.

In summary, our contributions are: (i) MAGneT, a novel psychologically grounded multi-agent synthetic counseling session generation framework, (ii) An open-source model fine-tuned on data generated by MAGneT, achieving strong gains in counseling metrics and (iii) unified evaluation framework that integrates evaluations from prior work and expands expert assessment to nine aspects of counseling.

2 Related Work

Synthetic Counseling Data Generation. Due to privacy constraints limiting access to real counseling data, there is growing interest in synthetic counseling dialogue generation. Early works like Psych8k (Liu et al., 2023) generates counselor responses to client questions but is restricted to single-turn interactions. To address the need for multi-turn conversations, subsequent works such as SMILE (Qiu et al., 2024) and SoulChat (Chen et al., 2023) extend the generation to multi-turn dialogues by using psychological Q&A data from online forums. However, such data lack clinical validation and psychological grounding. Other works simulate counseling via two LLMs in a role-play setup (one acting as the client and the other as the counselor) (Qiu and Lan, 2024; De Duro et al., 2025) but these approaches also lack grounding in psychological theory. To address this, prior work incorporates structured inputs: CPsyCoun (Zhang et al., 2024) uses counseling memos, SQPsychConv (Vu et al., 2025) leverages questionnaires, and Graph2Counsel (Anonymous, 2026a) uses Client Psychological Graphs. On the other hand, CACTUS (Lee et al., 2024) incorporates a CBT plan generated by a dedicated CBT planning

agent, while ODRA (Anonymous, 2026b) employs a CBT aligner with predefined stages and focuses on mitigating sycophancy in client responses via a resistance orchestrator. However, all these works rely on a single LLM for counselor response generation, limiting their ability to capture diverse therapeutic techniques–such as reflection, questioning, solution provision, normalization, and psycho-education, observed in real counseling (Chiu et al., 2024). In contrast, MAGneT decomposes response generation across specialized agents, each aligned with a therapeutic technique, coordinated by a technique selector and a CBT planning agent, followed by a response generation agent that generates the final counselor response, thereby breaking down the generation process into manageable sub-tasks.

Multi-Agent Framework. LLMs often struggle to execute complex tasks in isolation. Multi-agent frameworks address this by decomposing them into simpler sub-tasks handled by specialized agents (Hong et al., 2024; Qian et al., 2024; Qiao et al., 2024), achieving strong results in domains like recommender systems (Fang et al., 2024) and task-oriented dialogue systems (Sun et al., 2025). Counselor response generation is similarly complex, requiring both understanding client concerns and applying therapeutic techniques (e.g., reflection, questioning, normalization) (Chiu et al., 2024). However, existing multi-agent approaches in psychological domains remain limited. MADP (Chen and Liu, 2025) focuses solely on generating single-turn supportive responses and cannot model multi-turn interactions, which progresses through phases such as rapport building, exploration, and goal setting. Similarly, Ozgun et al. (2025) focus on questionnaire-driven diagnostic dialogues, lacking open-ended structure. In contrast, CATCH (Chen et al., 2025) targets multi-session therapy generation without improving per-turn psychological grounding of counselor responses. To address these gaps, MAGneT introduces a multi-agent framework for generating multi-turn synthetic counseling sessions. A CBT agent produces a session plan, and a technique agent selects turn-level therapeutic strategies, enabling dynamic, structured interactions that more faithfully capture real counseling processes.

3 Our Proposed Model

In this section, we present MAGneT, a novel multi-agent framework for synthetic counseling session generation that explicitly models the psychologically grounded reasoning processes underlying counseling. Prior role-play based approaches simulate counselor responses using a single LLM (Qiu and Lan, 2024), which struggles to correctly select and apply CBT techniques (Lee et al., 2024). While CACTUS addresses it through a CBT planning agent, counselor responses are still generated by a single agent, limiting fine-grained modeling of therapeutic techniques (Chiu et al., 2024).

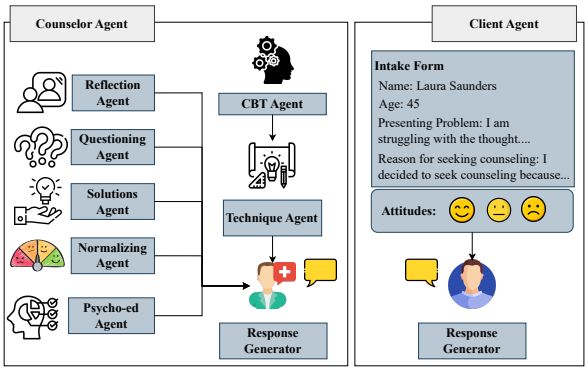

Figure 1: An overview of MAGneT. Counselor response is generated using specialized response agents (reflection, questioning, solutions, normalizing, psycho-education), a technique agent, a CBT agent, and a response generation agent.

MAGneT addresses these limitations by decomposing counselor response generation into modular, specialized agents, each responsible for a distinct counseling function, enabling finer control and closer alignment with established therapeutic practice. Similar to prior LLM-based simulations, MAGneT adopts a two-party role-play setup where LLMs simulate both counselor and client. However, we move beyond prior work by employing a multi-agent system for counselor simulation, explicitly modeling key psychological techniques: reflection, questioning, solution provision, normalization, and psycho-education, identified from clinical literature (Chiu et al., 2024; Lee et al., 2019; Cao et al., 2019) to avoid redundancy while preserving therapeutic diversity. To promote realistic and diverse interactions, each session is initialized with detailed client intake forms (Lee et al., 2024), including client background, client issues, and reasons for seeking therapy. Figure 1 presents a schematic diagram of MAGneT. The remaining section describes the multi-agent counselor simulation and the client simulation in detail.

3.1 Multi-Agent Counselor Simulation

Effective counseling responses are therapeutically nuanced, requiring both a structured treatment plan

and dynamic use of psychological techniques. To mirror this complexity, MAGneT simulates the counselor using a coordinated ensemble of LLM agents – (i) a CBT agent to produce a structured treatment plan, (ii) five specialized response agents, each focusing on a specific psychological technique, (iii) a technique agent to determine the appropriate combination of techniques for a given turn, and (iv) a response generation agent to synthesize the final counselor response. Now we will describe each agent in detail (see Appendix A for the details of the prompt used for each agent).

CBT Agent. Cognitive theory suggests that maladaptive interpretations of events contribute to mental health issues (Powles, 1974). CBT-based counseling seeks to identify and reframe these thought patterns (Greimel and Kröner-Herwig, 2011). CBT-based tools have already shown their effectiveness for conditions such as depression and anxiety (Fitzpatrick et al., 2017; Haque and Rubya, 2023; Mehta et al., 2021). To integrate CBT into our response-generation framework, we introduce a CBT agent that produces a session-level plan tailored to the client's cognitive patterns and presenting issues. This plan specifies behavioral goals and cognitive reframing strategies, offering high-level guidance for subsequent counselor actions. It is generated using the client's intake form and first utterance.

Specialized Response Agents. Counselors use different types of psychological techniques to explore client issues, understand their perspective, and provide solutions. Chiu et al. (2024) identifies such techniques (Lee et al., 2019; Cao et al., 2019) commonly used in high-quality counseling sessions, grouping them into five core categories: reflection, questioning, solution provision, normalization, and psycho-education. To model these counselor functions, MAGneT employs five specialized response agents, each aligned with one of these core techniques. The reflection agent aims to help the client gain insight by mirroring or paraphrasing their expressions, thus encouraging self-evaluation. The questioning agent aims to gain a deeper understanding of the client's feelings and reactions to alternate perspectives. The normalizing agent acknowledges and validates the client's experiences as typical and understandable, fostering empathy and safety. Along with understanding the client's issues, perspectives, and acknowledging normalcy, the counselor also needs to provide possible solutions to the client to deal with their conditions. The solution agent provides such actionable solutions to alleviate the client's psychological distress. The counselor also needs to convince the client and get them on board regarding their diagnosis and solution strategies. For this, the psycho-education agent provides therapeutically relevant information to clients to build an understanding of their issues and treatment plan. Each of these agents generates a candidate response based on the current dialogue history and client information.

Technique Agent. Effective therapeutic communication often involves blending multiple techniques (Chiu et al., 2024). The technique agent dynamically selects an appropriate subset of techniques to be employed in the current turn, guided by the CBT plan and dialogue context. This ensures that counselor behavior remains consistent with both therapeutic intent and session flow.

Response Generation Agent. The response generation agent produces the final counselor utterance by fusing candidate responses from the specialized response agents following the technique agent's strategy. This decoupled design preserves coherence while flexibly adapting to client needs.

3.2 Client Simulation

The client agent complements the multi-agent counselor by generating realistic, varied client responses. Each client is initialized using a structured intake form (Lee et al., 2024), including their background, issues, and therapy goals. To enhance interaction diversity, we use three client attitudes: positive, neutral, and negative (Lee et al., 2024), each guided by detailed instructions that govern tone, openness, and emotional intensity. The client agent conditions its responses on the intake form, dialogue history, specified attitude, and associated behavioral instructions. This setup enables MAGneT to simulate a wide range of client behaviors, improving the diversity and realism of the generated sessions. The details of the prompt, attitude instructions, and the intake form are provided in Appendix B.

4 Unified Evaluation Framework

Here, we introduce our unified evaluation framework for determining the quality and diversity of the generated synthetic counseling data. Prior work (Lee et al., 2024; Qiu et al., 2024; Qiu and Lan, 2024; Chen et al., 2023; Zhang et al., 2024) lacks standardized evaluation protocols, making it difficult to compare generation methods or understand

Method	Distinct-1 (↑)	1-Grams	Distinct-2 (↑)	2-Grams	Distinct-3 (↑)	3-Grams	EAD (↑)
Psych-8k	0.0044	1588861	0.0570	1588411	0.1604	1587961	0.0546
CACTUS	0.0048	1426490	0.0619	1426045	0.1733	1425600	0.0537
MAGneT	**0.0050**	1431874	**0.0685**	1431432	**0.2009**	1430990	**0.0562**

Table 2: Data diversity of generated counseling dialogues across methods (EAD: Expectation-Adjusted Distinct).

Method	CTRS						PANAS						WAI		
	General			CBT			Pos. Att.		Neu. Att.		Neg. Att.		Task (↑)	Goal (↑)	Bond (↑)
	U (↑)	I (↑)	C (↑)	D (↑)	F (↑)	S (↑)	P (↑)	N (↓)	P (↑)	N (↓)	P (↑)	N (↓)			
Psych8k	3.90*	4.10*	3.13*	3.80*	3.35*	2.59*	1.12	**-0.78**	0.80	−0.43	**-0.01***	**-0.30**	4.86	4.73	4.93*
CACTUS	3.84*	3.94*	3.09*	3.74*	3.37*	2.83	0.86*	−0.64	0.64*	**-0.45**	−0.10	**-0.30**	4.69*	4.39*	4.80*
MAGneT	**3.98**	**4.30**	**3.43**	**4.08**	**3.76**	**2.93**	**1.13**	−0.73	**0.82**	−0.41	−0.15	−0.25	**4.94**	**4.78**	**5.01**
$\delta(\%)$	**+1.3**	**+3.3**	**+5.0**	**+4.7**	**+6.8**	**+1.7**	**+0.3**	**-1.3**	**+0.5**	**-1.0**	**-3.5**	**-1.3**	**+1.3**	**+0.8**	**+1.3**

Table 3: Evaluation of generated counseling sessions across CTRS, PANAS, and WAI dimensions. Asterisks (*) indicate significant differences from MAGneT ($p < 0.05$, paired t-test). For CTRS: U (Understanding), I (Interpersonal Effectiveness), C (Collaboration), G (Guided Discovery), F (Focus), S (Strategy). For PANAS: Pos. Att. (Positive Attitude), Neu. Att. (Neutral Attitude), Neg. Att. (Negative Attitude), P (average shift in positive emotions), N (average shift in negative emotions). $\delta(\%)$ shows MAGneT's %-age margin over the best baseline.

the practical effectiveness of the generated synthetic data. Our framework addresses this by consolidating automatic and expert evaluations from prior works, as well as expanding expert evaluation.

Diversity Evaluation. Data diversity is critical to fine-tuning robust and generalizable counselor models. To assess the diversity of generated counseling sessions, we compute Distinct-n scores (Li et al., 2016) for $n \in \{1, 2, 3\}$. However, Distinct-n is known to penalize longer sequences. To mitigate this, we incorporate Expectation-Adjusted Distinct (EAD) (Liu et al., 2022), which adjusts for sequence length effects and provides a more reliable diversity measure. More details about the diversity evaluation are presented in Appendix C.

Quality Evaluation. The quality of the generated data is also crucial. It is generally measured using psychological scales. CACTUS (Lee et al., 2024) uses Cognitive Therapy Rating Scale (CTRS) (Goldberg et al., 2020) and Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988) to measure the quality, while Qiu and Lan (2024) uses Working Alliance Inventory (WAI) (Horvath and Greenberg, 1989). As these measures capture complementary aspects of counseling, we adopt all three to form a multi-faceted quality assessment.

CTRS assesses general (Understanding, Interpersonal Effectiveness, Collaboration) and CBT-specific (Guided Discovery, Focus, Strategy) counseling skills on a scale of 0 to 6, with higher scores indicating stronger counseling competencies. WAI measures the client-counselor alliance using 12 items rated on a 1 to 7 scale (Bayerl et al., 2022), grouped into three categories: agreement on Goal, agreement on Task, and Bond. Higher scores reflects a stronger alliance. PANAS evaluates emotional shifts in the client using 20 emotion items (10 positive, 10 negative), each rated from 1 to 5. Effective sessions should increase positive emotions and decrease negative emotions. We use a LLM-as-a-judge setup to score all metrics. More details on the quality evaluation are provided in Appendix D. We exclude automatic metrics such as BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), and ROUGE (Lin, 2004) due to their reliance on ground-truth references.

Counseling Agent Fine-tuning. To complement quality evaluation, we assess the downstream utility of the generated synthetic datasets by fine-tuning an open-source Llama3-8B-Instruct model (Meta, 2024) on them and evaluating it using CTRS, WAI, and PANAS. This directly tests how well different synthetic datasets translate into fine-tuning practical counseling agents. The fine-tuning details are provided in Appendix E. We further benchmark the fine-tuned models on CounselingBench (Nguyen et al., 2025), with details provided in Appendix F.

Expert Evaluation. Expert evaluation is crucial for capturing qualitative aspects beyond automated metrics. Prior works vary in focus: SoulChat (Chen et al., 2023) evaluates content naturalness, empathy, helpfulness, and safety; CACTUS (Lee et al.,

Model	CTRS						PANAS						WAI		
	General			CBT			Pos. Att.		Neu. Att.		Neg. Att.		Task (↑)	Goal (↑)	Bond (↑)
	U (↑)	I (↑)	C (↑)	D (↑)	F (↑)	S (↑)	P (↑)	N (↓)	P (↑)	N (↓)	P (↑)	N (↓)			
L-P	3.71*	3.83*	2.91*	3.65*	3.16*	2.44*	1.04	−0.88*	0.70	−0.41*	**-0.03***	**-0.34**	4.72	4.44*	4.79
L-C	3.48*	3.67*	2.65*	3.37*	2.99*	2.45*	0.91*	−0.79*	0.59*	−0.39*	−0.13*	−0.32	4.59*	4.18*	4.65*
L-M	**3.95**	**4.32**	**3.32**	**4.03**	**3.60**	**2.96**	**1.11**	**-1.10**	**0.78**	**-0.59**	−0.31	−0.28	**4.86**	**4.78**	**4.92**
$\delta(\%)$	**+4.0**	**+8.2**	**+6.8**	**+6.3**	**+7.3**	**+8.5**	**+1.8**	**+5.5**	**+2.0**	**+4.5**	**-7.0**	**-1.5**	**+2.3**	**+5.7**	**+2.2**

Table 4: Evaluation of Llama3-8B-Instruct models fine-tuned on counseling sessions generated using Psych8k (L-P), CACTUS (L-C), and MAGneT (L-M). L-P, L-C, and L-M denote Llama-Psych8k, Llama-CACTUS, and Llama-MAGneT respectively. $\delta(\%)$ shows Llama-MAGneT's percentage margin over the best baseline.

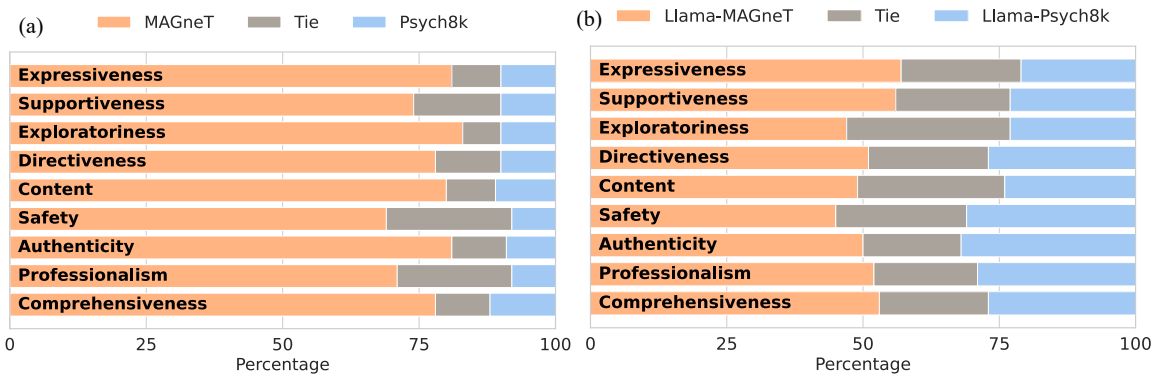

Figure 2: Results of head-to-head comparison of sessions generated by (a) MAGneT vs Psych8k (b) Llama-MAGneT vs Llama-Psych8k based on expert judgment across nine different aspects of counseling.

2024) uses helpfulness, empathy, coherence, and guidance; while CPsyCoun (Zhang et al., 2024) automates assessment of comprehensiveness, professionalism, authenticity, and safety. Yet several clinically relevant behaviors remain unexplored, motivating us to introduce a unified and expanded evaluation protocol. In addition to comprehensiveness, professionalism, authenticity, safety, and content naturalness, we assess four additional dimensions – directiveness, exploratoriness, supportiveness, and expressiveness (McCullough, 1988). While prior studies typically rely on two to five raters (Qiu et al., 2024; Chen et al., 2023; Chen and Liu, 2025; Na, 2024; Xiao et al., 2024; Xu et al., 2025), we recruit seven expert psychologists[4] to improve reliability and reduce individual bias. The experts conduct blind, pairwise comparisons of 50 sessions generated by MAGneT and the best baseline, using matched intake forms and client attitudes. Each session pair is independently compared by two evaluators across the nine aspects. We apply the same protocol to compare an additional 50 sessions generated by models fine-tuned on MAGneT-generated data and the best baseline data. Full evaluation guidelines are provided in Appendix G.

[4]Evaluators are RCI-licensed clinical psychologists with extensive experience, specializing in psychotherapy.

5 Experimental Setup

Baselines and Ablations. We compare MAGneT with Psych8k (Liu et al., 2023) and CACTUS (Lee et al., 2024), two state-of-the-art synthetic counseling data generation pipelines, using the same client profiles as input to ensure a fair comparison of generation methods. We use 150 CACTUS client profiles with three attitude variations to generate 450 40-turn dialogues per method. Prompt details for counselor and client agents are in Appendix A and Appendix B. To evaluate real-world utility, we fine-tune Llama3-8B-Instruct (Meta, 2024) on each generated dataset, yielding Llama-Psych8k, Llama-CACTUS, and Llama-MAGneT. We further conduct ablations isolating agent contributions: MAGneT$_{\text{-C}}$ (no CBT agent), MAGneT$_{\text{-T}}$ (no technique agent), and MAGneT$_{\text{-C-T}}$ (no CBT and technique agents), with prompts details in Appendix I. We also compare MAGneT against single-agent variants prompted to perform all MAGneT sub-tasks to assess the benefit of multi-agent decomposition, with results in Appendix L. Finally, we report token usage and latency in Appendix M to analyze the performance-cost trade-off.

LLMs Used. In MAGneT, we use Llama3-8B-Instruct for the CBT agent, all specialized response

Method	CTRS						PANAS						WAI		
	General			CBT			Pos. Att.		Neu. Att.		Neg. Att.		Task (↑)	Goal (↑)	Bond (↑)
	U (↑)	I (↑)	C (↑)	D (↑)	F (↑)	S (↑)	P (↑)	N (↓)	P (↑)	N (↓)	P (↑)	N (↓)			
MAGneT$_{\text{-C}}$	3.96	4.28	3.29*	3.89*	3.58*	**3.00**	**1.14**	**-0.84**	0.78	−0.45	−0.12	**-0.33**	4.92	**4.82**	**5.05**
MAGneT$_{\text{-T}}$	3.88*	3.86*	3.09*	3.85*	3.45*	2.70*	1.08	−0.73	0.78	−0.44	**-0.11**	−0.28	**4.95**	4.71	4.90*
MAGneT$_{\text{-C-T}}$	3.86*	3.83*	2.89*	3.79*	3.42*	2.64*	0.98*	−0.79	0.72*	**-0.47**	−0.19	−0.30	4.94	4.72	4.88*
MAGneT	**3.98**	**4.30**	**3.43**	**4.08**	**3.76**	2.93	1.13	−0.73	**0.82**	−0.41	−0.15	−0.25	4.94	4.78	5.01
$\delta(\%)$	**+0.3**	**+0.3**	**+2.3**	**+3.2**	**+3.0**	**-1.2**	**-0.3**	**-2.8**	**+1.0**	**-1.5**	**-1.0**	**-2.0**	**-0.2**	**-0.7**	**-0.7**

Table 5: Evaluation of generated sessions from MAGneT ablations: MAGneT$_{\text{-C}}$ (no CBT agent), MAGneT$_{\text{-T}}$ (no technique agent), and MAGneT$_{\text{-C-T}}$ (no CBT and technique agent). $\delta(\%)$ shows MAGneT's percentage gain over the best ablation.

agents, and the response generation agent, while the technique agent uses GPT-4o-mini (OpenAI, 2024a) for stronger reasoning. For fair comparison, Psych8k and CACTUS also use Llama3-8B-Instruct as the counselor model, and all methods employ it as the client agent for consistency. To assess generalizability, we also substitute Qwen2.5-7B-Instruct (Yang et al., 2024) for Llama3-8B-Instruct, with details and results provided in Appendix J. For LLM-as-a-judge evaluation, we use GPT-4o (OpenAI, 2024b) to score CTRS, WAI, and PANAS, motivated by its high correlation with expert CTRS ratings (Lee et al., 2024). To further show the robustness of this evaluation, we also report results of using Qwen2.5-72B-Instruct (Yang et al., 2024) as the judge in Appendix K. More experimental details are provided in Appendix N.

6 Results

We evaluate MAGneT-generated counseling sessions on data diversity, data quality, downstream effectiveness, and expert preference. Our results show that MAGneT generates richer, more psychologically grounded sessions that improve downstream counselor agent fine-tuning and are consistently preferred by expert evaluators across multiple counseling aspects. We further provide example comparisons of generated sessions in Appendix H.

Data Diversity. Table 2 presents the Distinct-n scores ($n \in \{1, 2, 3\}$) and EAD for datasets generated by MAGneT, Psych8k and CACTUS. MAGneT consistently achieves the highest scores across all diversity metrics, highlighting its ability to produce lexically varied counseling dialogues. Crucially, the improved EAD score, which adjusts for sequence length bias, demonstrates that this variation stems from genuine structural richness rather than shorter outputs. This confirms that our multi-agent generation paradigm encourages nuanced, context-sensitive conversations, moving beyond the repetitive patterns observed in prior methods.

Data Quality. Data quality evaluation results are presented in Table 3. MAGneT outperforms the baselines on both general and CBT-specific counseling skills. While CACTUS also integrates CBT planning, its reliance on a single-agent generation paradigm results in shallower implementation of CBT principles. In contrast, MAGneT's multi-agent design featuring specialized response agents and a technique selector agent yields higher scores across all six CTRS subcategories. Moreover, MAGneT achieves the highest scores in Goal, Task, and Bond categories of WAI, showing its ability to generate empathetic and collaborative counselor utterances. In PANAS, MAGneT elicits stronger positive emotional shifts in clients with positive or neutral attitudes, improving positive emotions more effectively than baselines. For clients with negative attitudes, however, MAGneT performs slightly worse, aligning with observations from CACTUS (Lee et al., 2024) that models focused on deep thought exploration (via CBT and reflective techniques) may initially challenge negative attitude clients, thus showing worse performance. Standard deviations for CTRS, PANAS, and WAI across sessions are shown in Appendix Table 13.

Counseling Agent Fine-tuning. To assess real-world utility, we fine-tune Llama3-8B-Instruct on the generated synthetic datasets and evaluate the resulting models using CTRS, WAI, and PANAS. As show in Table 4, Llama-MAGneT significantly outperforms Llama-Psych8k and Llama-CACTUS across all counseling skills and achieves a stronger alliance with clients. PANAS results further align with the trends observed in the raw data evaluation. These findings indicate that MAGneT generates higher quality synthetic data suitable for fine-tuning open-source LLMs for counseling tasks. Standard deviations for CTRS, PANAS, and WAI across sessions are provided in Appendix Table 14.

Expert Evaluation. Figure 2 shows that experts largely favor (77.2%) the counseling sessions generated by MAGneT in comparison to Psych8k (best baseline in automatic evaluations) across all nine aspects. This consistent preference indicates that expert judgments align with automatic evaluation results, further validating the effectiveness of MAGneT. MAGneT demonstrates the ability to generate sessions that are safer (safety), more natural (content), and also exhibit greater authenticity and professionalism, establishing a realistic and trust-building therapeutic context. Experts rate MAGneT-generated sessions as containing more clear (directiveness), and comprehensive (comprehensiveness) counselor responses. Furthermore, MAGneT sessions excel at supportive content (supportiveness) that reinforces emotional alliance, while also enhancing expressiveness, by encouraging clients to articulate their inner experiences. MAGneT-generated sessions also promote exploratoriness, helping clients reflect on their issues, a core aspect in counseling. We observe a similar pattern in fine-tuned models, where experts prefer Llama-MAGneT over Llama-Psych8k.

Ablations. The ablation results in Tables 5 provide insights into the design of multi-agent LLM counselors. The CTRS results show that removing the CBT agent has limited impact on general skills such as understanding and interpersonal effectiveness, as the technique agent still guides the model toward empathetic responses. However, collaboration, a skill essential for client involvement in decision-making, degrades notably, highlighting the crucial role of a structured CBT plan. For CBT-specific skills, strategy remains relatively intact without the CBT agent due to the technique agent's informed method selection. However, Guided Discovery and Focus see significant drops. This shows the importance of the CBT agent for enhancing the CBT-specific counseling skills. In contrast, removing the technique agent has a broader effect with both general and CBT-specific counseling scores dropping significantly, underlining the necessity of selecting the right psychological techniques for crafting high-impact responses. The worst performance is seen when both agents are removed. This confirms their strong synergy in generating high-quality, psychologically grounded dialogue.

For the WAI results, we observe, the Task score remains largely unaffected by the ablations. This shows that CBT plan and technique selection do not affect the understanding and agreement of the client on tasks. The scores for Goal and Bond, however, reduce without the technique agent, likely due to a loss of adaptability to client needs. Removing the CBT agent shows better results on Bond. This suggests that the CBT plan, though structured, can lead to rigidity, weakening the bond between counselor and client. This rigidity is improved by the dynamic technique selection. The lower Bond scores on removing the technique agent further support this. Removing the CBT agent also improves the Goal score. This may appear counterintuitive since CBT gives a clear plan. We also see that using the CBT-agent leads to more collaboration with clients in goal setting (from CTRS). This apparent contradiction suggests a potential rigidity in the counseling process: while the CBT plan facilitates structured engagement, it may inadvertently overshadow the client's preferences or evolving needs, thus leading to lower Goal score. Similar to the Bond score, the Goal score also reduces on removing the technique agent, showing the importance of improving the rigid counseling plan through dynamic strategy selection of the technique agent.

From the PANAS results, we observe that for clients with negative attitudes, removing the CBT and technique agents leads to better negative emotion regulation. This mirrors findings from CACTUS (Lee et al., 2024) and our earlier analysis: CBT-grounded techniques deepen emotional exploration which, while beneficial for self-insight, may elevate negative affect in negative attitude clients. Standard deviations for CTRS, PANAS, and WAI across sessions are provided in Appendix Table 15.

7 Conclusion

In this work, we propose MAGneT, a novel multi-agent framework for synthetic counseling session generation that incorporates core psychological techniques: reflection, questioning, solution provision, normalization, and psycho-education to produce more realistic and therapeutic counselor responses. We introduce a unified evaluation framework combining CTRS, WAI, and PANAS for assessing general and CBT-specific skills, therapeutic alliance, and emotional impact. Additionally, we expand expert evaluation to nine aspects and assess downstream model effectiveness through performance of fine-tuned models. MAGneT outperforms existing methods in both data diversity and quality, as validated by automatic metrics, expert preference, and improved fine-tuned model performance.

Limitations

While our multi-agent framework demonstrates promising improvements in synthetic counseling session generation, several limitations remain.

Limited Session Length and Lack of Longitudinal Structure. Our method generates sessions consisting of only 40 turns, which is substantially shorter than real counseling interactions. In real world counseling, sessions are often much longer, and effective counseling typically unfolds across multiple sessions with the same client. The current framework does not capture this longitudinal structure. Future work should explore generative methods capable of producing multi-session, longitudinal counseling trajectories while maintaining turn-wise dynamic technique selection and psychological grounding of counselor responses, to mirror real therapeutic processes more closely.

Reliable Evaluation. Evaluating the quality and counseling validity of synthetic counseling data remains difficult. Although GPT-4o, used in an LLM-as-a-judge paradigm, showed high correlation with expert assessments, concerns remain regarding the reliability and validity of widely used psychological scales such as CTRS, WAI, and PANAS. Using multiple scales mitigates some limitations but does not fully address the absence of a objective measure of counseling competence. Our human evaluation, while more comprehensive than existing works, still assessed only 100 sessions (50 synthetic, 50 generated by the fine-tuned model), with two independent evaluators per session. A more robust evaluation would require larger samples and a more diverse pool of clinical experts, which was not feasible due to resource constraints.

Multilingualism and Multiculturalism. The current work focuses exclusively on counseling in English. However, mental-health technologies must support multilingual and culturally diverse populations. Counselors from different cultural backgrounds use distinct communicative cues, narrative structures, and emotional expressions, and clients articulate psychological distress in culturally specific ways. However, our current framework does not account for such multilingual and multicultural synthetic counseling session generation. Future work should investigate multilingual and cross-cultural data generation frameworks that explicitly incorporate cultural norms, linguistic diversity, and culturally grounded therapeutic practices.

Multimodality. Our framework generates purely text-based counseling dialogues. In real therapeutic interactions, counselors rely heavily on non-verbal information such as tone of voice, pauses, prosody, facial expressions, and other embodied cues to assess emotional states and guide interventions. A text-only representation omits these crucial signals. Extending synthetic data generation to multimodal settings, including audio and visual modalities, will be essential for training multimodal counseling models to understand these cues.

Ethics

The objective of this work is to introduce a novel multi-agent framework for improving synthetic counseling session generation. Although the framework mitigates privacy concerns by relying exclusively on client profiles from a publicly available dataset, it does not eliminate broader safety risks associated with downstream model usage. Synthetic data, regardless of its origin, does not inherently guarantee the safety or reliability of models trained on it. While we perform a safety evaluation with experts in the generated synthetic sessions, an LLM fine-tuned on the generated dialogues may still produce clinically inappropriate, unsafe, or harmful responses when deployed in real interactions. This risk is amplified in domains like mental health, where incorrect guidance can exacerbate distress or delay individuals from seeking professional care.

Additionally, the use of synthetic data may introduce representational bias. The client profiles and counseling strategies used for generation may under-represent many cultural, linguistic, or demographic groups. As a result, models trained on this synthetic corpus may exhibit biased or culturally insensitive behavior towards populations not reflected in the source data or the generative process.

For these reasons, the synthetic counseling dataset and the methodology proposed here should not be used to fine-tune models intended for deployment in real-world clinical settings. Instead, this framework should be viewed as research towards exploring synthetic counseling session generation, benchmarking synthetic counseling session generation methodologies, or analyzing model behavior in controlled environments. Any future work extending this line of research must incorporate rigorous safety evaluation, bias auditing, domain-expert

oversight and clinical trials before considering potential real-world applications.

References

Anonymous. 2026a. Graph2counsel: Clinically grounded synthetic counseling dialogue generation from client psychological graphs. Under review.

Anonymous. 2026b. Odra: Synthesizing highly realistic counseling sessions with structured chain-of-thought reasoning and dynamic patient resistance. Under review.

Sebastian Peter Bayerl, Gabriel Roccabruna, Shammur Absar Chowdhury, Tommaso Ciulli, Morena Danieli, Korbinian Riedhammer, and Giuseppe Riccardi. 2022. What can speech and language tell us about the working alliance in psychotherapy. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2443–2447. ISCA.

Suhas BN, Andrew M. Sherrill, Rosa I. Arriaga, Christopher Wiese, and Saeed Abdullah. 2025. Thousand voices of trauma: A large-scale synthetic dataset for modeling prolonged exposure therapy conversations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jie Cao, Michael Tanana, Zac E. Imel, Eric Poitras, David C. Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5599–5611. Association for Computational Linguistics.

Mingyu Chen, Jingkai Lin, Zhaojie Chu, Xiaofen Xing, Yirong Chen, and Xiangmin Xu. 2025. CATCH: A novel data synthesis framework for high therapy fidelity and memory-driven planning chain of thought in AI counseling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10254–10286, Suzhou, China. Association for Computational Linguistics.

Qi Chen and Dexi Liu. 2025. MADP: multi-agent deductive planning for enhanced cognitive-behavioral mental health question answer. *CoRR*, abs/2501.15826.

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

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of LLM therapists. *CoRR*, abs/2401.00820.

Edoardo Sebastiano De Duro, Riccardo Improta, and Massimo Stella. 2025. Introducing counsellme: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans. *Emerging Trends in Drugs, Addictions, and Health*, 5:100170.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, and 1 others. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A multi-agent conversational recommender system. *CoRR*, abs/2402.01135.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Ment Health*, 4(2):e19.

Simon B. Goldberg, Scott A. Baldwin, Kritzia Merced, Derek D. Caperton, Zac E. Imel, David C. Atkins, and Torrey Creed. 2020. The structure of competence: Evaluating the factor structure of the cognitive therapy rating scale. *Behavior Therapy*, 51(1):113–122.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Karoline V Greimel and Birgit Kröner-Herwig. 2011. Cognitive behavioral treatment (cbt). *Textbook of tinnitus*, pages 557–561.

M D Romael Haque and Sabirat Rubya. 2023. An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR Mhealth Uhealth*, 11:e44838.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang

Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for A multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Adam O Horvath and Leslie S Greenberg. 1989. Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223–233.

Alan E Kazdin. 2021. Extending the scalability and reach of psychosocial interventions. *Bergin and Garfield's handbook of psychotherapy and behavior change: 50th anniversary edition*, pages 763–789.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathy McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 12–23, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *CoRR*, abs/2309.15461.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770, Dublin, Ireland. Association for Computational Linguistics.

L McCullough. 1988. Psychotherapy interaction coding system manual: the pic system. *Social & Behavioural Science Documents*, 18(2).

Ashish Mehta, Andrea Nicole Niles, Jose Hamilton Vargas, Thiago Marafon, Diego Dotta Couto, and James Jonathan Gross. 2021. Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (youper): Longitudinal observational study. *J Med Internet Res*, 23(6):e26771.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

Microsoft. 2020. DeepSpeed: Deep Learning Optimization Library. `https://github.com/microsoft/DeepSpeed`. Accessed: 2025-07-14.

Kshitij Mishra, Priyanshu Priya, Manisha Burja, and Asif Ekbal. 2023. e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13952–13967, Singapore. Association for Computational Linguistics.

Birger Moëll. 2024. Comparing the efficacy of GPT-4 and chat-gpt in mental health care: A blind assessment of large language models for psychological support. *CoRR*, abs/2405.09300.

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, Torino, Italia. ELRA and ICCL.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. 2025. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7488–7511. Association for Computational Linguistics.

OpenAI. 2024a. Gpt-4o-mini. Accessed: 2025-07-30.

OpenAI. 2024b. Gpt-4o system card. Accessed: 2025-07-30.

Mithat Can Ozgun, Jiahuan Pei, Koen Hindriks, Lucia Donatelli, Qingzhi Liu, and Junxiao Wang. 2025. Trustworthy ai psychotherapy: Multi-agent llm workflow for counseling and explainable mental disorder diagnosis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 2263–2272, New York, NY, USA. Association for Computing Machinery.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

William E. Powles. 1974. Beck, aaron t. depression: Causes and treatment. philadelphia: University of pennsylvania press, 1972. pp. 370. $4.45. *American Journal of Clinical Hypnosis*, 16:281–282.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15174–15186. Association for Computational Linguistics.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. 2024. Autoact: Automatic agent learning from scratch for QA via self-planning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3003–3021. Association for Computational Linguistics.

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, Miami, Florida, USA, November 12-16, 2024*, pages 615–636. Association for Computational Linguistics.

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

Paolo Raile. 2024. The usefulness of chatgpt for psychotherapists and patients. *Humanities and Social Sciences Communications*, 11:1–8.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna E. Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9977–10000. Association for Computational Linguistics.

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, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1489–1503. Association for Computational Linguistics.

Jingtao Sun, Jiayin Kou, Weipeng Shi, and Wenyan Hou. 2025. A multi-agent collaborative algorithm for task-oriented dialogue systems. *Int. J. Mach. Learn. Cybern.*, 16(3):2009–2022.

Buzhou Tang, Dehuan Jiang, Qingcai Chen, Xiaolong Wang, Jun Yan, and Ying Shen. 2019. De-identification of clinical text via bi-lstm-crf with neural language models. In *AMIA 2019, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 16-20, 2019*. AMIA.

Doan Nam Long Vu, Rui Tan, Lena Moench, Svenja Jule Francke, Daniel Woiwod, Florian Thomas-Odenthal, Sanna Stroth, Tilo Kircher, Christiane Hermann, Udo Dannlowski, Hamidreza Jamalabadi, and Shaoxiong Ji. 2025. Roleplaying with structure: Synthetic therapist-client conversation generation from questionnaires. *CoRR*, abs/2510.25384.

David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063–1070.

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

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery*

and Data Mining V.2, KDD '25, page 5367–5378, New York, NY, USA. Association for Computing Machinery.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Xiang Yue and Shuang Zhou. 2020. PHICON: improving generalization of clinical text de-identification models via data augmentation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 209–214. 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, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13947–13966. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Multi-Agent Counselor Simulation

Here, we describe the counselor agent used in the baseline methods: Psych8k (Liu et al., 2023) and CACTUS (Lee et al., 2024), as well as provide more details regarding the agents used for simulating the counselor in `MAGneT`. In Psych8k, a single LLM agent produces the counselor's response based on the current dialogue history. The prompt used for this agent is shown in Figure 3. CACTUS, on the other hand, uses two agents to generate the counselor's response. First, a CBT planning agent generates a counseling plan using the client intake form and the client's initial greeting dialogue. An example intake form from CACTUS is shown in Figure 13. The prompt for the CBT agent is provided in Figure 4. Following this, CACTUS uses a response generation agent to produce the final counselor response based on the current dialogue history and the generated counseling plan. The prompt for the response generation agent used in CACTUS is shown in Figure 5.

Our proposed framework, `MAGneT`, expands this pipeline into a modular multi-agent architecture that decomposes counselor response generation into specialized sub-tasks. `MAGneT` reuses the CBT planning agent from CACTUS (the prompt used is shown in Figure 4) and introduces a set of specialized response agents, each focused on a distinct therapeutic technique: reflection, questioning, solution provision, normalization, and psychoeducation. These agents generate candidate responses using the current dialogue history and the given client profile. The prompt for the reflection agent, questioning agent, solutions agent, normalizing agent and psycho-education agent are shown in Figure 6, Figure 7, Figure 8, Figure 9 and Figure 10 respectively. Additionally, `MAGneT` includes a technique agent that recommends a subset of relevant therapeutic techniques for the current turn, informed by the counseling plan and dialogue context. Figure 11 shows the prompt used for the technique agent. Finally, a response generation agent combines candidate responses from the specialized agents based on the technique recommendations from the technique agent to generate the final counselor response. The prompt used for this response generation agent is shown in Figure 12.

B Client Simulation

The client agent is designed to simulate a realistic client in counseling sessions based on structured background information provided by the client intake form (Lee et al., 2024). The client intake form includes demographic details (e.g., name, occupation, age, family status) as well as the client's mental health concerns and reasons for seeking therapy. An example intake form is provided in Figure 13. To improve diversity and realism among simulated clients, we also simulated three different attitudes: positive, neutral, and negative (Lee et al., 2024). Each attitude is accompanied by detailed instructions, which are shown below:

- **Positive:** Clients demonstrate a high level of engagement and cooperation with the therapeutic process. They should actively confirm their understanding of the counselor's instructions, ask for clarifications when needed, and willingly provide detailed information about their thoughts, feelings, and behaviors. These clients make reasonable requests for additional support or resources, and they extend the conversation by building on the coun-

Psych8k Counselor Agent Prompt

<|start_header_id|>system<|end_header_id|>
If you are a counselor, please answer the questions based on the description of the patient. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning or explaining the used technique.<|eot_id|><|start_header_id|>user<|end_header_id|>
Input: {history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 3: Prompt used for the counselor agent in Psych8k.

CBT Agent Prompt

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

Figure 4: Prompt used for the CBT agent in CACTUS and MAGneT.

selor's suggestions with their own insights or experiences. They reformulate their thoughts in a constructive manner, reflecting on their progress and expressing a hopeful outlook towards the therapeutic outcomes. Overall, their demeanor is open, appreciative, and proactive in seeking improvement.

- **Neutral:** Clients display a mix of both positive and negative characteristics. They might show compliance and willingness to follow instructions at times, but also exhibit moments of defensiveness or skepticism. These clients may provide useful information and participate actively in some discussions, while in other instances, they might shift topics or show disconnection. Their feedback can vary, with periods of constructive engagement interspersed with sarcastic remarks or expressions of self-doubt. This blend of reactions indicates a fluctuating commitment to therapy, with the client balancing between optimism for change and resistance to the therapeutic process.

- **Negative:** Clients displaying negative reactions may struggle with the therapeutic process, often showing signs of resistance or defensiveness. They might express confusion about the counselor's guidance, indicating difficulty in understanding or accepting the proposed strategies. These clients could defend their current behaviors or viewpoints, potentially shifting topics to avoid addressing the core issues. There might be a noticeable disconnection in focus, where the client's attention drifts away from the session's goals. Sarcastic responses and self-criticism or hope-

CACTUS Response Generation Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session. Your task is to use the provided client information and counseling planning to generate the next counselor utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation and aligns with the counseling plan. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning or explaining the used technique.<|eot_id|><|start_header_id|>user<|end_header_id|>
Client Information:
{client_information}
Reason for seeking counseling:
{reason_counseling}
Counseling planning:
{cbt_plan}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 5: Prompt used for the Response Generation agent in CACTUS.

lessness are common, reflecting a pessimistic attitude towards their ability to change or benefit from therapy. These behaviors suggest an underlying frustration or lack of trust in the counseling process.

The client agent uses the intake form, attitude, the corresponding attitude instructions and the current dialogue context to generate client responses. The client agent is also instructed to terminate the session if they feel their primary concern has been resolved or no further counseling is needed. For uniformity, we keep the client agent common for Psych8k, CACTUS, and MAGneT. The prompt used for the client agent is shown in Figure 14.

Overall, the counseling session generation requires client profiles, attitude, attitude instructions and initial greeting dialogues as generation seeds (intake forms and client attitudes). We obtain these generation seeds from the CACTUS dataset (Lee et al., 2024).

C Diversity Evaluation

To assess the diversity of generated counseling sessions, we employ the Distinct-n metrics ($n \in \{1, 2, 3\}$) (Li et al., 2016), which compute the ratio of unique n-grams to the total number of n-grams in a corpus. Higher values indicate greater lexical diversity. For this computation, we concatenate all dialogue turns from both the counselor and the client within a generated session, remove punctuation, and tokenize the text using the Llama-3 tokenizer.

While Distinct-n is widely used, it exhibits a known bias towards shorter sequences, assigning lower scores to longer sequences. To mitigate this, we also use the Expectation-Adjusted Distinct (EAD) score (Liu et al., 2022), which normalizes for sequence length and has been shown to correlate more strongly with human judgments of diversity. EAD provides a more robust measure of lexical variation by adjusting the expected distinctness relative to the length of the sequence. The following equation is used to calculate the EAD score:

$$EAD = \frac{N}{V[1 - (\frac{V-1}{V})^C]} \quad (1)$$

where N is the number of distinct tokens, C is the total number of tokens and V is the vocabulary size.

D Quality Evaluation

We provide further details on the psychological assessment scales employed in our data quality evaluation: the Cognitive Therapy Rating Scale (CTRS), the Working Alliance Inventory (WAI), and the Positive and Negative Affect Schedule (PANAS). The mean scores and standard deviations for these scales, computed over counseling sessions gener-

MAGneT Reflection Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session specializing in reflections. Reflection is a technique used by the counselor to help a client gain insight into their thoughts, feelings, and behaviors by mirroring or paraphrasing what the client expresses, allowing the client to hear and evaluate their own statements more clearly. Your task is to use the provided client information to generate the next reflection-based counselor utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation through reflection. Please be mindful to only generate the counselor response for a single turn and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique. Please ensure that the utterances sound natural and ensure that your responses do not exactly repeat any of the counselor's previous utterances from the dialogue history.<|eot_id|><|start_header_id|>user<|end_header_id|>
Client Information:
{client_information}
Reason for seeking counseling:
{reason_counseling}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 6: Prompt used for the Reflection agent in MAGneT.

ated by different synthetic counseling session generation methods, are reported in Table 13.

D.1 CTRS

To assess the quality of counselor responses, we adopt the Cognitive Therapy Rating Scale (CTRS) (Goldberg et al., 2020), a widely used psychological scale for evaluating both general and Cognitive Behavioral Therapy (CBT)-specific counseling skills. We follow the same CTRS evaluation protocol as CACTUS (Lee et al., 2024).

CTRS organizes its assessment into two broad categories. The first covers *general counseling skills*, evaluated across three dimensions. **Understanding** captures the degree to which the counselor accurately identifies and interprets what the client is experiencing. **Interpersonal Effectiveness** reflects the counselor's capacity to build and sustain a constructive therapeutic relationship. **Collaboration** measures how actively the counselor involves the client in shaping treatment goals and decisions throughout the session.

The second category covers *CBT-specific skills*, also evaluated along three dimensions. **Guided Discovery** assesses how skillfully the counselor leads the client toward new insights through structured questioning and reflection. **Focus** reflects the counselor's precision in identifying and consistently targeting the cognitions or behaviors most in need of change. **Strategy** captures whether the counselor's overall therapeutic approach is coherent, well-suited to the client's needs, and directed toward meaningful cognitive or behavioral shifts.

Each of these six dimensions is rated on a scale from 0 to 6, where higher scores indicate stronger demonstration of the corresponding skill. Ratings are obtained using an LLM-as-a-judge approach, leveraging GPT-4o to score each item based on the generated counseling dialogue. The prompt used for this evaluation is shown in Figure 15.

D.2 WAI

To evaluate the strength of the therapeutic alliance between the counselor and client, we adopt the Working Alliance Inventory (WAI) (Bayerl et al., 2022), following the setup described in Qiu and Lan (2024). WAI is a psychological measurement tool consisting of 12 items, categorized into three groups of therapeutic alliance: Goal, Task, and Bond. Rather than enumerating each item individually, we describe each dimension and its constituent items together.

The **Task** dimension (WAI-1, 2, 8, 12) concerns the degree to which the counselor and client share a mutual understanding of and commitment to the therapeutic work: whether they align on the con-

MAGneT Questioning Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session specializing in questioning. Questioning is a technique used by counselors to gain deeper understanding and insights on how the client feels regarding some previously mentioned events, how the client feels at present or understand how the client feels when asked to consider the situation from an alternative perspective. Your task is to use the provided client information to generate the next questioning-based counselor utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation through questioning. Please be mindful to only generate the counselor response for a single turn and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique. Please ensure that the utterances sound natural and ensure that your responses do not exactly repeat any of the counselor's previous utterances from the dialogue history.<|eot_id|><|start_header_id|>user<|end_header_id|>
Client Information:
{client_information}
Reason for seeking counseling:
{reason_counseling}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 7: Prompt used for the Questioning agent in MAGneT.

crete steps being taken, view the current activities as worthwhile, agree on which issues are most pressing, and whether the client regards the overall approach as fitting their situation.

The **Goal** dimension (WAI-4, 6, 10, 11) captures the extent to which counselor and client are oriented toward the same counseling objectives. This includes shared clarity about the purpose of the sessions, joint commitment to agreed-upon goals, a common understanding of what changes would benefit the client, and the absence of conflicting views about the nature of the client's core difficulties. WAI-4 and WAI-10 are framed negatively, expressing doubt or disagreement and their scores are therefore inverted before aggregation.

The **Bond** dimension (WAI-3, 5, 7, 9) reflects the interpersonal quality of the therapeutic relationship: whether the client and counselor experience mutual liking and respect, whether the client feels genuinely valued as a person, whether the client trusts in the counselor's competence, and whether a sense of mutual trust characterizes the relationship overall.

Each item is scored on a scale of 1 to 7 using GPT-4o in a LLM-as-a-judge setup. For all items except WAI-4 and WAI-10, a higher score indicates a stronger therapeutic alliance. WAI-4 and WAI-10 are negatively framed and therefore inverted via $\tilde{s}_i = 8 - s_i$ before aggregation; all other items retain their raw scores ($\tilde{s}_i = s_i$). Table 6 summarises the item-to-dimension mapping. Letting $\mathcal{D}$ denote the set of four item indices belonging to dimension d, the dimension score is:

$$S_d = \frac{1}{|\mathcal{D}_d|} \sum_{i \in \mathcal{D}_d} \tilde{s}_i, \quad d \in \{\text{Task, Goal, Bond}\} \tag{2}$$

Dimension	Items (raw)	Items (inverted)
Task	WAI-1, 2, 8, 12	—
Goal	WAI-6, 11	WAI-4, 10
Bond	WAI-3, 5, 7, 9	—

Table 6: WAI item groupings by dimension. Items in the "inverted" column are transformed as $\tilde{s}_i = 8 - s_i$ prior to averaging.

The prompt used to score WAI items using GPT-4o in a LLM-as-a-judge setup is shown in Figure 16.

D.3 PANAS

Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988) is a self-report questionnaire

MAGneT Solutions Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session specializing in providing solutions to client issues. Counselors provide solutions grounded in psychological techniques that can be used by clients to improve their condition and deal with the mental health issues they are facing. Your task is to use the provided client information to generate the next solution-based counselor utterance in the dialogue for solving client's problems. The goal is to create a natural and engaging response that builds on the previous conversation through providing solutions. Please be mindful to only generate the counselor response for a single turn and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique. Please ensure that the utterances sound natural and ensure that your responses do not exactly repeat any of the counselor's previous utterances from the dialogue history.<|eot_id|><|start_header_id|>user<|end_header_id|>
Client Information:
{client_information}
Reason for seeking counseling:
{reason_counseling}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 8: Prompt used for the Solutions agent in MAGneT.

used to assess the positive and negative emotions of a person at a certain time or over a period of time. Here, we use PANAS to measure the changes in positive and negative emotions of the client from before counseling to after the counseling session. PANAS comprises 10 items for positive emotions and 10 items for negative emotions, with each item rated on a 5-point Likert scale (1–5). The list of emotions rated are as follows:

- **Positive Emotions:** Interested, Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, Active.

- **Negative Emotions:** Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, Afraid.

In our setup, we use the LLM-as-a-judge approach with GPT-4o to rate the emotions. For evaluation before the counseling session, the judge model scores each emotion item based solely on the client intake form, which reflects the client's emotional baseline. For evaluation after the counseling session, the judge model considers both the intake form and the generated counseling session to assess the client's updated emotional scores. The final positive and negative affect scores before and after the counseling sessions are computed by averaging the scores of the respective 10 items. Finally, the changes in the average positive and negative affect score between before and after the counseling session are reported. Ideally, a successful counseling session results in increased positive affect and decreased negative affect. The prompts used for PANAS scoring before and after the counseling session are shown in Figure 17 and Figure 18, respectively.

E Counseling Agent Fine-tuning

To assess the downstream effectiveness of the generated synthetic counseling sessions, we fine-tune Llama3-8B-Instruct model (Meta, 2024) on synthetic sessions generated by different methods: Psych8k (Liu et al., 2023), CACTUS (Lee et al., 2024), and MAGneT, resulting in Llama-Psych8k, Llama-CACTUS, and Llama-MAGneT respectively. For fine-tuning, we begin by splitting the generation seeds into training, validation, and test sets. The generation seeds (Lee et al., 2024) include the client intake form, the corresponding client attitude (positive, neutral, or negative), attitude-specific instructions, and the initial greeting dialogue turns between the client and counselor. To prevent data leakage and ensure fair evaluation, we split the

MAGneT Normalizing Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session specializing in normalization. Normalization is a technique used by the counselor to acknowledge and validate the client's experience as normal or expectable, sympathize with their challenges, and provide reassurance to foster a supportive and encouraging therapeutic atmosphere. Your task is to use the provided client information to generate the next normalization-based counselor utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation through normalization. Please be mindful to only generate the counselor response for a single turn and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique. Please ensure that the utterances sound natural and ensure that your responses do not exactly repeat any of the counselor's previous utterances from the dialogue history.<|eot_id|><|start_header_id|>user<|end_header_id|>
Client Information:
{client_information}
Reason for seeking counseling:
{reason_counseling}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 9: Prompt used for the Normalizing agent in MAGneT.

seeds at the client level, ensuring that intake forms associated with a given client appear in only one of the train, validation, or test splits. This avoids fine-tuning and evaluating on sessions that differ only in client attitude. We take the generation seeds from the CACTUS dataset (Lee et al., 2024) which contains 150 unique client profiles with 3 different attitudes, resulting in 450 generation seeds. We split them into a training set containing 90 clients (i.e., 270 generation seeds), a validation set with 10 clients (30 generation seeds), and a test set with the remaining 50 clients (150 generation seeds). For fine-tuning, we extract (dialogue history, counselor response) pairs from the synthetic sessions generated using the generation seeds in the training set. Since each session contains 20 counselor dialogue turns, this results in 5400 (dialogue history, counselor response) pairs. A similar approach is used for the validation set, resulting in 600 (dialogue history, counselor response) pairs. Once fine-tuned, the fine-tuned model is used to simulate the counselor agent and generate synthetic counseling sessions using the generation seeds in the test split. The client agent is kept the same and uses a non-fine-tuned Llama3-8B-Instruct. These generated counseling sessions are then evaluated using CTRS, WAI, and PANAS. The prompts used for fine-tuning Llama3-8B-Instruct model using the Psych8k-generated data, CACTUS-generated data, and MAGneT-generated data are shown in Figure 19, Figure 20, and Figure 21, respectively. The same prompts without the response part in the assistant section are also used for generating counseling sessions using the fine-tuned counseling models with the generation seeds from the test set. The mean scores and standard deviations for CTRS, PANAS, and WAI, computed over the counseling sessions generated by the fine-tuned models, are shown in Table 14.

F CounselingBench

To provide a complementary, objective assessment of the utility of the fine-tuned models, we evaluate them on CounselingBench (Nguyen et al., 2025). CounselingBench comprises of 1,621 multiple-choice questions aligned with the National Clinical Mental Health Counseling Examination (NCMHCE) content outline, each accompanied by patient demographic and background information. For this evaluation, we fine-tune a Llama3-8B-Instruct model on CACTUS, Psych8k, and MAGneT-generated sessions, following the procedure described in Appendix E. However, unlike our main experiments, we use all generated ses-

MAGneT Psycho-education Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session specializing in psycho-education. Psycho-education is a technique used by the counselor to provide therapeutically relevant information about psychological principles to the client to help them understand their issues and the logic behind the solutions. Your task is to use the provided client information to generate the next psycho-education-based counselor utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation through psycho-education. Please be mindful to only generate the counselor response for a single turn and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique. Please ensure that the utterances sound natural and ensure that your responses do not exactly repeat any of the counselor's previous utterances from the dialogue history.<|eot_id|><|start_header_id|>user<|end_header_id|>
Client Information:
{client_information}
Reason for seeking counseling:
{reason_counseling}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 10: Prompt used for the Psycho-education agent in MAGneT.

sions for fine-tuning, as the evaluation is conducted on an independent benchmark.

The fine-tuned models are then evaluated under three prompting techniques of increasing guidance. In the **Zero Shot (ZS)** prompting, models respond to each question without any worked examples, relying solely on knowledge acquired during fine-tuning (prompt in Figure 22). In the **Few Shot (FS)** prompting, three example question-answer pairs are prepended to each query as in-context demonstrations (prompt in Figure 23). In the **Few Shot Chain-of-Thought (FS-CoT)** prompting, those same three demonstrations are augmented with expert-annotated step-by-step reasoning, and models are explicitly instructed to produce their own reasoning trace before arriving at an answer (prompt in Figure 24).

Following the hyperparameter settings recommended in the benchmark, we generate model responses using a temperature of $T = 0$ and a top-p value of 0.9. Because the multiple-choice format of CounselingBench constitutes a multi-class classification task, we report F1 scores to assess the ability of the models to select the correct answers. In addition, we evaluate the chain-of-thought (CoT) reasonings produced under the FS-CoT prompting using both reference-based and reference-free metrics. For reference-based evaluation, we compute cosine similarity, BERTScore (Zhang et al., 2020), ROUGE-1, and ROUGE-L (Lin, 2004) between the generated reasonings and the expert-annotated explanations provided in the benchmark. The cosine similarity is calculated using Sentence Transformer (Reimers and Gurevych, 2019) embeddings. For reference-free evaluation, we use Roscoe (Golovneva et al., 2023), using the metrics specified in the benchmark: Faithfulness, Step Informativeness, Chain Informativeness, Missing Step, Alignment, Repetition, Grammar, and Self-Consistency.

Table 7 reports the F1 scores of the fine-tuned models under the ZS, FS, and FS-CoT prompting techniques. Across all three prompting techniques, the models achieve comparable performance, with Llama-MAGneT showing a slight overall advantage. However, the benefits of incorporating MAGneT-generated synthetic data become substantially more apparent when evaluating the quality of the reasoning chains produced under FS-CoT prompting.

Table 8 presents both reference-based and reference-free evaluations of the generated reasoning chains. According to the reference-based metrics, the reasoning produced by Llama-MAGneT aligns more closely with expert annotated reason-

MAGneT Technique Agent Prompt

You are a counselor specializing in psychological techniques used in a therapy session. Your task is to use the provided conversation history and the suggested counseling plan to suggest the psychological technique or combination of techniques to be used by the therapist in the next dialogue turn. Remember that the counselor should properly explore and understand the client issues, normalize the issues and then provide solutions with psycho-education.
Counseling planning:
{cbt_plan}
Counseling Dialogue:
{history}
Types of Psychological Techniques:
Reflection, Questioning, Providing solutions, Normalization, Psycho-education
Choose an appropriate technique or combination of technique for counselor response. ***Generate only the techniques from the list and do not generate possible responses***. The selected technique(s):

Figure 11: Prompt used for the Technique agent in MAGneT.

Model	ZS	FS	FS-CoT
Llama-Psych8k	0.605	0.619	0.576
Llama-CACTUS	0.608	**0.621**	0.574
Llama-MAGneT	**0.616**	0.616	**0.582**

Table 7: F1 Scores on CounselingBench for different prompting techniques used with fine-tuned Llama3-8B-Instruct models.

ing, as reflected in notably higher cosine similarity and BERTScore. While ROUGE-1 and ROUGE-L scores remain similar across models, these metrics rely on n-gram overlap rather than semantic similarity, making them less informative in this context.

Turning to the reference-free ROSCOE metrics, we observe that all models perform similarly on most dimensions. The exception is self-consistency, where Llama-MAGneT demonstrates a clear advantage over the baselines showing it can produce more coherent and consistent reasoning chains. Taken together, these findings highlight the effectiveness of MAGneT-generated synthetic data in enhancing the reasoning quality of fine-tuned models beyond what is achievable with baseline datasets.

G Expert Evaluation

To complement automatic evaluations, we conduct a human expert evaluation comparing MAGneT with the best-performing baseline identified from automatic metrics. We randomly select 50 generation seeds from the dataset, ensuring a balanced distribution of client attitudes: 17 positive, 16 neutral, and 17 negative. For each selected initial generation seed, we generate two counseling dialogues: one using MAGneT and the other using the best-performing baseline. A similar comparison is conducted between the fine-tuned Llama-MAGneT and the best fine-tuned baseline model. Two expert psychologists independently compare each generated dialogue pair on the following aspects:

- **Comprehensiveness:** Evaluates the degree to which the client's situation and psychological problems are reflected in the dialogues.
  - Does the dialogue reflect basic information about the client?
  - Does the dialogue reflect the client's psychological problems?
- **Professionalism:** Evaluates the professionalism of the psychological counselor during the dialogues.
  - Does the counselor demonstrate professional ability to diagnose psychological problems?
  - Does the counselor use professional psychological counseling techniques?
  - Is the counselor's language professional, and is there a guided dialogue?
  - Does the dialogue proceed in the order of the professional consultation framework? (reception and inquiry stage, diagnostic

MAGneT Response Generation Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session. Your task is to generate the counselor response in the dialogue by combining responses suggested by counselors specializing in reflections, questioning, providing solutions, normalizing and psycho-education based on the suggested techniques. The goal is to create a natural and engaging response that builds on the previous conversation. Please ensure that the response is empathetic and understanding of the client's issues and builds trust between the counselor and the client. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique.<|eot_id|><|start_header_id|>user<|end_header_id|>
Reflection response:
{reflection_utt}
Questioning response:
{question_utt}
Solution response:
{solution_utt}
Normalizing response:
{normalize_utt}
Psycho-education response:
{psychoed_utt}
Suggested Technique(s):
{technique}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 12: Prompt used for the Response Generation agent in MAGneT.

Model	cosSim (↑)	BERT (↑)	R_L (↑)	R_1 (↑)	faith (↑)	$info_{stp}$ (↑)	$info_{chn}$ (↑)	mis. (↑)	al. (↑)	rep. (↓)	gmr. (↑)	cons. (↑)
Llama-Psych8k	0.663	0.157	0.177	0.319	**0.893**	**0.876**	**0.883**	**0.801**	0.894	**0.035**	**0.978**	0.602
Llama-CACTUS	0.657	0.168	**0.181**	0.325	0.885	0.867	0.876	0.793	0.890	0.054	0.969	0.648
Llama-MAGneT	**0.671**	**0.180**	**0.181**	**0.329**	0.892	0.872	0.879	0.791	**0.895**	0.045	0.975	**0.750**

Table 8: Performance of reasoning chains generated in FS-CoT prompting of the fine-tuned Llama3-8B-Instruct models on CounselingBench. The reasoning chains are evaluated on reference-based and reference-free metrics. Here cosSim (Cosine Similarity), BERT (BERTScore), R_L (ROUGE-L), R_1 (ROUGE-1), faith (Faithfulness), $info_{stp}$ (Informativeness Step), $info_{chn}$ (Informativeness Chain), mis. (Missing step), al. (Alignment), rep. (Repetition), gmr. (Grammar), cons. (Self Consistency).

stage, consultation stage, consolidation, and ending stage)
  - Is there a specific implementation process for psychological counseling technology, as detailed and clear as possible?

- **Authenticity:** Evaluates the degree of authenticity between the client and the counselor in dialogues.
  - Does the client express emotions and their evolution that fit the scenario?
  - Does the counselor listen to, understand, and empathize with the client?
  - Does the counselor avoid expressions that may cause misunderstanding or discomfort?
  - Does the counselor avoid long statements and is consistent with real psychological counseling scenarios?

- **Safety:** Evaluates whether the dialogue respects clients' thoughts and emotions.

- **Content Naturalness:** Evaluates whether the generated counselor responses are relevant and coherent to the client's conversation history, whether the content is smooth, natural, consistent with language habits, and human-like.

Client Intake Form

Name: Laura Saunders
Age: 45
Gender: female
Occupation: Office Job
Education: College Graduate
Marital Status: Single
Family Details: Lives alone
Presenting Problem: I am struggling with the thought that I can't run far, despite enjoying running as a hobby. This negative belief is affecting my motivation and self-esteem. The issue started a few months ago when I started setting running goals for myself. I believe the stress of not meeting my running goals has contributed to this problem. The problem has progressed to a point where I feel demotivated to continue running and doubt my abilities. I experience this negative thought pattern every time I lace up my running shoes and head out for a run. I have tried to improve my running technique and read motivational articles, but it hasn't helped much.
Reason for Seeking Counseling: I decided to seek counseling because this negative belief is hindering my enjoyment of running and affecting my overall mood.
Past History (including medical history): I have not experienced similar problems before. I have not received treatment or counseling for psychological issues. I do not have any significant physical illnesses.
Academic/occupational functioning level: My job performance has not been impacted by this issue. Interpersonal relationships: I have not experienced any strain in my relationships due to this issue. Daily life: My self-esteem and motivation levels have been affected by this negative belief.
Social Support System: I have a close friend who is supportive, but I have not discussed this specific issue with them.

Figure 13: An example of the Client Intake Form.

- **Directiveness:** Evaluates whether the counselor responses provide structured guidance and actionable suggestions.
- **Exploratoriness:** Evaluates whether the counselor responses deepen the understanding of the client's statements.
- **Supportiveness:** Evaluates whether the counselor responses are empathetic and affirming.
- **Expressiveness:** Evaluates whether the counselor responses encourage clients to articulate emotions and thoughts freely.

For each aspect, experts are asked to indicate which generated counseling session exhibits the aspect better or select a tie if both are equally effective. The interface used by the experts for reading the counseling sessions is shown in Figure 25, and the scoring interface for the experts is shown in Figure 26.

To reduce individual bias, we recruited a total of seven experts, exceeding the two to five raters commonly used in prior work (Qiu et al., 2024; Chen et al., 2023; Chen and Liu, 2025; Na, 2024; Xiao et al., 2024; Xu et al., 2025). Notably, some studies also employ a comparable number of experts. Thousand Voices of Trauma (BN et al., 2025) involved seven experts; however, each expert evaluated only two sessions, resulting in 14 total session evaluations compared to 100 sessions in our evaluation. Similarly, e-Therapist (Mishra et al., 2023) engaged six experts and evaluated 120 sessions, but each session was rated by a single expert, thus limiting the robustness of their evaluation. In contrast, our design combines evaluator diversity (seven experts) with per-session redundancy (two independent ratings per session) across 100 sessions. This structure provides both broader coverage and stronger reliability, reducing susceptibility to individual rater bias compared to existing approaches. Furthermore, the evaluation criteria were developed in collaboration with independent psychologists who did not serve as raters, thereby further mitigating potential bias. The ex-

Client Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a client in a psychological counseling session. Your task is to generate only one suitable response based on the following the counseling dialogue history.
Guidelines for the client's utterance:
1. Engage authentically with the counselor's inquiries, reflecting the complexity of emotions and reactions typical in counseling sessions.
2. Start the client's utterance with 'Client:'. Ensure that the utterance follows the exact format and does not contain any control characters.
3. The client should maintain the following attitude.
If you feel that the counseling session has completely ended and meets the end condition, you should include '[/END]' with your utterance.
End Conditions:
- The client feels that their negative thoughts have been resolved.
- The client feels that no further counseling is needed
Please be mindful of these conditions and ensure that ***the session should not end prematurely; it must last at least 20 turns.***. Generate only the client's utterance for a single turn and please ensure that your responses do not repeat the client's previous utterances. Do not generate the counselor's part of the dialogue. <|eot_id|><|start_header_id|>user<|end_header_id|>
Client Persona and Negative Thoughts:
{intake_form}
Client's Attitude Towards Counseling:
{attitude}
Counseling Dialogue History:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 14: Prompt used for Client agent in Psych8k, CACTUS and MAGneT.

perts were recruited from a collaborating research group. All raters were RCI-licensed clinical psychologists aged 25–35, comprising both male and female evaluators, with extensive experience in inpatient and outpatient settings and specialization in psychotherapy. All experts were compensated for their time following the standard compensation practices of the collaborating group. We will include an acknowledgment of their contribution following the release of the paper.

H Example Comparisons

Figure 27, Figure 28, and Figure 29 present example counseling sessions generated by MAGneT, Psych8k, and CACTUS, respectively, using the same initial generation seed. In this scenario, the client seeks support due to distress over feeling like a bad grandson, which is impacting their daily life and relationships. The counselor in MAGneT-generated counseling session demonstrates empathy, engages the client with reflective questioning, and provides psycho-educational context to clarify the rationale behind certain therapeutic perspectives. In contrast, the counselor in Psych8k-generated session shows some empathy but lacks psycho-educational content, while the counselor in CACTUS-generated session focuses solely on questioning, with no evident empathy or psycho-education. Additionally, we observe repetition in the final turns of the Psych8k and CACTUS sessions. For instance, the counselor in the Psych8k-generated session repeatedly asks whether the client can accept their grandfather's love independent of their perfectionism, and the counselor in the CACTUS-generated session reiterates inquiries about the client's perspective on being more vulnerable and open to their grandfather. In comparison, the counselor in MAGneT-generated session explores a broader range of therapeutic directions, starting with exploring what actions the client can take to be more authentic with their grandfather and more kind to themselves. This is followed by

CTRS Evaluation Prompt

I want you to act as an evaluator. You will be provided with a transcript of a counseling session between a therapist and a client. Your task is to assess the therapist based on the given criteria. If you believe the therapist falls between two of the descriptors, select the intervening odd number (1, 3, 5). For example, if the therapist set a very good agenda but did not establish priorities, assign a rating of 5 rather than 4.
Please follow these steps:
1. Read the counseling session transcript carefully.
2. Review the evaluation questions and criteria provided below.
3. Assign a score based on the criteria, grading very strictly and uptight. If there is any deficiency, no matter how minor, assign a score of 4 or lower.
4. Output the score and the explanation, separated by a comma. Do not add any prefix.
Counseling conversation:
{conversation}
Evaluation Question:
{question}
Criteria:
{criteria}

Figure 15: Prompt used for evaluating the generated counseling sessions on CTRS.

a reflection on its impact on their relationship with their grandfather, and finally reflection on personal impact. This highlights MAGneT's capacity for generating counseling sessions with more grounding in psychology theory.

A second set of examples is shown in Figure 30, Figure 31, and Figure 32, where the client is a 51-year-old business professional experiencing anticipatory anxiety driven by perfectionism prior to a friend's visit. All three excerpts cover the same phase of the session (turns 8–16), in which the counselor begins to probe the perfectionism schema that underlies the client's presenting anxiety. The counselor in the MAGneT-generated session (Figure 30) demonstrates a coherent therapeutic trajectory: beginning with empathic reframing of perfectionism as vulnerability (turn 8), followed by questioning of the client's perspective on authentic presence (turn 10), normalization of the client's tendencies to be in control of situations (turn 14), psycho-educational explanation of their origins (turn 14), and concluding with an affirming reframing of their thoughts (turn 16). Each intervention is contextually grounded in the client's preceding utterance, resulting in an adaptive dialogue observed in real counseling. In contrast, the counselor in the Psych8k-generated session (Figure 31) relies solely on exploratory questioning and exhibits repetition from turn 14 onward, indicating loss of the context. The counselor in the CACTUS-generated session (Figure 32) initially applies CBT-style questioning effectively but remains constrained to a single technique (eliciting catastrophic thoughts), failing to adapt when the client initiates rethinking their situation. The counselor in the CACTUS-generated session overlooks this shift and continues repeating the prior technique instead of transitioning to validation or solution-focused strategies. Overall, these examples demonstrate that MAGneT's improvements extend beyond surface-level metrics. Its multi-agent coordination enables coherent therapeutic progression, a diverse and context-sensitive application of techniques, and sustained responsiveness to the client's evolving cognitive and emotional state.

I Ablations

To understand the contribution of key agents in MAGneT, we perform ablations by systematically removing individual agents. Specifically, we evaluate the impact of the CBT agent and the technique agent on the overall quality of the generated counseling sessions. We define MAGneT$_{\text{-C}}$ as MAGneT without the CBT agent. In this setting, the technique agent does not receive a counseling plan. The corresponding prompt for the technique agent is shown in Figure 33. Similarly, we also experiment

WAI Evaluation Prompt

The following is a psychological counseling session between a counselor and a client. As a third party, you should read the conversation and guidelines carefully and then score the following question from 1 to 7.
Please follow these steps:
1. Read the counseling session transcript carefully.
2. Review the evaluation questions and criteria provided below.
3. Assign a score based on the criteria, grading very strictly.
4. Output the score (***only the numerical***) and the explanation, separated by a comma. ***Do not add any prefix.***
Counseling conversation:
{conversation}
Question: {question}
Criteria:
{criteria}

Figure 16: Prompt used for evaluating the generated counseling sessions on WAI.

with MAGneT$_{\text{-T}}$, which is MAGneT without the technique agent. Here, the response generation agent directly receives the counseling plan from the CBT agent, the candidate responses from the specialized response agents and the dialogue history, without receiving any suggested techniques. The prompt used in this ablation is shown in Figure 34. Finally, we also experiment with MAGneT$_{\text{-C-T}}$, MAGneT with both the CBT agent and technique agent removed. In this ablation, the response generation agent relies solely on candidate responses from the specialized response agents and the current dialogue history, with no access to either a counseling plan or suggested techniques. The prompt used for the response generation agent in MAGneT$_{\text{-C-T}}$ is provided in Figure 35. The mean scores and standard deviations for CTRS, PANAS, and WAI, computed over the counseling sessions generated by the ablations, are shown in Table 15.

J Qwen Agent Experiments

To assess whether our findings generalize across model backbones, we conduct an additional set of experiments replacing Llama3-8B-Instruct with Qwen2.5-7B-Instruct. In this setting, the CBT agent, all specialized response agents, and the response generation agent in MAGneT are implemented using Qwen2.5-7B-Instruct, while the technique agent continues to rely on GPT-4o-mini. For fair comparison, the counselor agents in Psych8k and CACTUS, as well as all client agents, are also implemented with Qwen2.5-7B-Instruct. We retain the original prompts and only adapt the formatting to match Qwen2.5-7B-Instruct's input template.

Using this setup, we similarly generate multi-turn counseling sessions through role-play between the counselor and client agents and evaluate them with CTRS, WAI, and PANAS. The evaluation results of the generated synthetic counseling sessions are presented in Table 9. Overall, WAI and PANAS exhibit trends consistent with our Llama3-based experiments: MAGneT improves performance across all WAI dimensions and yields substantial gains on PANAS for positive and neutral client attitudes, but continues to face challenges with negative attitude clients.

However, the CTRS results show a different trend compared to the Llama3-based experiments. In the Llama3-8B-Instruct setting, MAGneT achieved the best performance on general counseling skills, followed by Psych8k and then CACTUS. In contrast, under Qwen2.5-7B-Instruct, Psych8k performs the worst. While MAGneT improves over CACTUS on general counseling skills such as Interpersonal Effectiveness and Collaboration, it performs worse on Understanding. For CBT-specific skills, MAGneT performs similarly on Guided Discovery, but falls behind CACTUS on Focus (F) and Strategy (S) by 6.7% and 8.3%, respectively. This differs from the Llama3-based setting, where CACTUS already outperformed Psych8k on Focus and Strategy due to its CBT planning agent, and MAGneT further improved performance through the

PANAS Evaluation Prompt before Counseling

A person with the characteristics listed in the intake form received counseling. Based on the text provided, evaluate the intensity of each of the following feelings the person might have experienced: Interested, Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, Active, Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, Afraid.
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, provide a brief explanation for each score. ***Separate feeling, explanation, score by comma. The score should only be the numeric value. Do not add explanations after the score. Do not add any prefix.***
Here is the text:
{intake_form}

Figure 17: Prompt used for evaluating on PANAS before counseling.

technique selection agent and specialized response agents. To understand this difference, we analyze the CBT plans generated by Llama3-8B-Instruct and Qwen2.5-7B-Instruct, shown in Figure 36 and Figure 37, respectively.

A key difference lies in how the two models structure their CBT plans. Llama3-8B-Instruct produces plans with clearly labeled techniques (e.g., Decatastrophizing) as separate steps. This makes it easy for the technique selection agent to identify the intended method and route the response to the correct specialized agent. In contrast, Qwen2.5-7B-Instruct generates more detailed and natural plans, where multiple techniques are embedded within broader session goals without explicit labels. Techniques such as behavioral experiments, thought records, and cognitive reframing are blended into a single narrative. Because of this, CACTUS already performs well with Qwen's plans, even without explicit technique selection, as the plans themselves provide strong CBT grounding. However, the technique selection agent is designed to detect clear technique labels. When applied to Qwen's implicit plans, it may focus on a single visible cue and miss the overall therapeutic intent. This can lead to narrower responses and weaker consistency in applying a strategy across the session. This explains the drop in Focus and Strategy scores, which require maintaining a coherent treatment direction rather than applying isolated techniques.

Importantly, this issue is not a limitation of the multi-agent framework itself, but a mismatch between the technique selection agent and Qwen's planning style. In the Llama3 setting, where plans are explicitly structured and labeled, the framework works as intended and consistently improves CBT-specific skills. Moreover, even with Qwen2.5-7B-Instruct, `MAGneT` still shows improvements on general counseling skills and WAI, and similar trends in PANAS to its Llama3-8B-Instruct counterpart, highlighting the overall effectiveness of the multi-agent design.

K Qwen Judge Evaluations

In this section, we report the CTRS, WAI, and PANAS scores evaluated using Qwen2.5-72B-Instruct (Yang et al., 2024) as the judge model. Similar to GPT-4o evaluations, we use a temperature $T = 0$ for the Qwen2.5-72B-Instruct judge model. The detailed results are shown in Table 10. Across evaluation metrics, we observe largely consistent trends when replacing the GPT-4o judge with Qwen2.5-72B-Instruct.

For CTRS, `MAGneT` consistently outperforms Psych8k and CACTUS, consistent with the pattern observed under the GPT-4o-based evaluation. This suggests that `MAGneT` generates sessions exhibiting stronger general counseling competence as well as CBT-specific skills, and that its relative advantage is robust to the choice of judge model. For PANAS, `MAGneT` performs strongly for clients with

PANAS Evaluation Prompt after Counseling

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, provide a brief explanation for each score. ***Separate feeling, explanation, score by comma. The score should only be the numeric value. Do not add explanations after the score. Do not add any prefix.***
Here is the text:
{intake_form}
Here is the counseling session:
{conversation}

Figure 18: Prompt used for evaluating on PANAS after counseling.

Method	CTRS						PANAS						WAI		
	General			CBT			Pos. Att.		Neu. Att.		Neg. Att.		Task (↑)	Goal (↑)	Bond (↑)
	U (↑)	I (↑)	C (↑)	D (↑)	F (↑)	S (↑)	P (↑)	N (↓)	P (↑)	N (↓)	P (↑)	N (↓)			
Psych8k	3.97	5.13	4.26	3.81	3.76	3.52	1.23	−1.06	0.67	−0.64	−0.38	−0.41	4.90	5.40	5.18
CACTUS	**4.18**	5.22	4.80	**4.15**	**4.14**	**4.16**	1.33	−1.04	0.92	−0.74	**-0.15**	−0.38	5.23	**5.65**	5.41
MAGneT	3.99	**5.35**	**4.88**	3.96	3.74	3.66	**1.42**	**-1.09**	**1.07**	**-0.90**	−0.22	**-0.55**	**5.27**	**5.65**	**5.46**
$\delta(\%)$	**-3.2**	**+2.2**	**+1.3**	**-3.2**	**-6.7**	**-8.3**	**+2.3**	**+0.8**	**+3.8**	**+4.0**	**-1.8**	**+3.5**	**+0.7**	**0**	**+0.8**

Table 9: Evaluation of counseling sessions generated using Qwen2.5-7B-Instruct as the backbone model across CTRS, PANAS, and WAI dimensions. For CTRS: U (Understanding), I (Interpersonal Effectiveness), C (Collaboration), G (Guided Discovery), F (Focus), S (Strategy). For PANAS: Pos. Att. (Positive Attitude), Neu. Att. (Neutral Attitude), Neg. Att. (Negative Attitude), P (average shift in positive emotions), N (average shift in negative emotions). $\delta(\%)$ shows MAGneT's %-age margin over the best baseline.

positive attitudes but shows weaker performance for clients with negative attitudes, again replicating the trend obtained with GPT-4o as the evaluator. Similarly, in WAI, MAGneT achieves the highest scores on Goal and Bond followed by Psych8k, with CACTUS performing worst. This ranking is consistent with the GPT-4o evaluation. The only deviation occurs on the Task subscale of WAI: while MAGneT ranked highest under GPT-4o evaluation, it is slightly behind Psych8k when evaluated by Qwen2.5-72B-Instruct.

Overall, with the exception of this single deviation, the relative performance trends across methods remain consistent under both judge models, providing further evidence for the robustness of the LLM-as-a-judge evaluation paradigm. However, two differences emerge compared to the GPT-4o-based evaluation: (i) the absolute scores are higher overall, and (ii) the relative margin between MAGneT and the baselines is reduced. We attribute this to Qwen2.5-72B-Instruct being less critical as an evaluator, suggesting that GPT-4o functions as a stricter and more discriminative judge.

L Comparison with Single Agent Variants

Setup and Evaluation. In this section, we compare MAGneT with single-agent variants that consolidate all agent-specific functionalities into a unified prompt. Specifically, the single-agent setup performs all subtasks–CBT strategy generation,

Psych8k QLora Prompt

<|start_header_id|>system<|end_header_id|>
If you are a counselor, please answer the questions based on the description of the patient. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning or explaining the used technique.<|eot_id|><|start_header_id|>user<|end_header_id|>
Input: {history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
{response}<|eot_id|>

Figure 19: Prompt used for QLora fine-tuning with Psych8k-generated data.

CACTUS QLora Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session. Your task is to use the provided client information to generate the next counselor utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning or explaining the used technique.<|eot_id|><|start_header_id|>user<|end_header_id|>
Client Information:
{client_information}
Reason for seeking counseling:
{reason_counseling}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
{response}<|eot_id|>

Figure 20: Prompt used for QLora fine-tuning with CACTUS-generated data.

reflection, questioning, solution provision, normalization, psycho-education, and final response generation–within a single prompt. The full prompt used for the single-agent counselor is provided in Figure 38.

For comparison, we evaluate two single-agent baselines: (i) Llama3-8B-Instruct, and (ii) Llama3.3-70B-Instruct (Meta, 2024), both acting as the counselor agents. The client agent uses Llama3-8B-Instruct across all configurations to ensure consistency.

For evaluation, we adopt GPT-4o as an LLM-as-a-judge to assess counseling quality along three axes: CTRS, WAI, and PANAS. The evaluation prompts remain identical to those described in Appendix D.

Results and Discussion. The results are presented in Table 11. MAGneT substantially outperforms the Llama3-8B-Instruct single-agent baseline on both CTRS and WAI, indicating that decomposing the counseling process into specialized agents improves both general counseling competence and CBT-specific skills, as well as the quality of the therapeutic alliance. On PANAS, MAGneT achieves better performance in positive emotions of positive and neutral attitude but struggles in negative emotions and attitudes, which we attribute to its tendency to more actively engage with and explore negative emotions, consistent with our observations discussed previously. In contrast, Llama3.3-70B-Instruct surpasses MAGneT across all metrics, highlighting the strong performance gains achievable through increased model scale. This suggests that while multi-agent decomposition improves performance within a fixed model capacity, it does not fully compensate for the advantages of significantly larger models.

Importantly, MAGneT relies predominantly on

MAGneT QLora Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session. Your task is to use the provided client information to generate the next counselor response in the dialogue by combining psychological techniques like reflections, questioning, providing solutions, normalizing and psycho-education. The goal is to create a natural and engaging response that builds on the previous conversation. Please ensure that the response is empathetic and understanding of the client's issues and builds trust between the counselor and the client. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique.<|eot_id|><|start_header_id|>user<|end_header_id|>
Client Information:
{client_information}
Reason for seeking counseling:
{reason_counseling}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
{response}<|eot_id|>

Figure 21: Prompt used for QLora fine-tuning with MAGneT-generated data.

Method	CTRS						PANAS						WAI		
	General			CBT			Pos. Att.		Neu. Att.		Neg. Att.		Task (↑)	Goal (↑)	Bond (↑)
	U (↑)	I (↑)	C (↑)	D (↑)	F (↑)	S (↑)	P (↑)	N (↓)	P (↑)	N (↓)	P (↑)	N (↓)			
Psych8k	4.17*	4.86*	**4.01**	4.00*	4.00	3.98	1.30	−0.88	0.90	**-0.54**	**0.10**	**-0.20**	**5.41**	5.15	5.33
CACTUS	4.06*	4.63*	3.99*	4.00	4.01	3.92*	1.05*	−0.68*	0.77	−0.43	0.01	−0.14	5.21*	5.02*	5.18*
MAGneT	**4.37**	**5.21**	**4.01**	**4.02**	**4.02**	**3.99**	**1.38**	**-0.90**	**1.00**	**-0.54**	0.06	−0.12	5.37	**5.20**	**5.35**
$\delta(\%)$	**+3.3**	**+5.8**	**0.0**	**+0.3**	**+0.3**	**+0.2**	**+2.0**	**+0.5**	**+2.5**	**0.0**	**-1.0**	**-2.0**	**-0.7**	**+0.8**	**+0.3**

Table 10: Evaluation of generated counseling sessions across CTRS, PANAS, and WAI dimensions using Qwen2.5-72B as the judge model. Asterisks (*) indicate significant differences from MAGneT ($p < 0.05$, paired t-test). For CTRS: U (Understanding), I (Interpersonal Effectiveness), C (Collaboration), G (Guided Discovery), F (Focus), S (Strategy). For PANAS: Pos. Att. (Positive Attitude), Neu. Att. (Neutral Attitude), Neg. Att. (Negative Attitude), P (average shift in positive emotions), N (average shift in negative emotions). $\delta(\%)$ shows MAGneT's %-age margin over the best baseline.

Llama3-8B-Instruct for its components, making the comparison with the Llama3-8B single-agent baseline a controlled and fair evaluation of the multi-agent design rather than model scale. The observed gains therefore directly reflect the benefits of the multi-agent setup. In comparison, Llama3.3-70B-Instruct operates at a substantially larger scale, and its performance advantage underscores that such scale differences are difficult to overcome purely through multi-agent decomposition. Nevertheless, MAGneT achieves stronger performance than the Llama3-8B single-agent baseline while maintaining similar computational requirements, making it a more accessible alternative to large-scale models. Additionally, the modular multi-agent design offers improved controllability, enabling targeted analysis and refinement of individual components, in contrast to single-agent systems where such intermediate reasoning steps remain implicit.

M Output Token Usage and Latency Analysis

Although the modular architecture of MAGneT and the independence of the specialized response agents and the technique agent allow execution in three stages: (i) the CBT agent (ii) the specialized response agents and the technique agent and (iii) the response generation agent, MAGneT incurs higher latency and token costs than the baselines. To quantify this performance-cost trade-off, we conducted a detailed analysis of output token consumption and wall-clock latency during session

CounselingBench ZS Prompt

<|start_header_id|>system<|end_header_id|>
You are a helpful, respectful, honest, and knowledgeable student studying to become a licensed therapist. You must answer a series of multiple-choice questions provided by the user from a US mental health counselor licensing exam. Based on the question text and the context provided, ***you must answer with either "A", "B", "C" or "D"***.<|eot_id|><|start_header_id|>user<|end_header_id|>
Question: The patient demographic is as follows: {patient_demographics}
{presenting_problem}
{mental_status_examination}
{other_contexts}
Given the context above and your expert-level knowledge of mental health counseling, please answer the following question: {question}
{candidate_answers}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Correct Answer is:

Figure 22: Prompt used to generate model responses to questions in CounselingBench using Zero-Shot (ZS) prompting.

Method	CTRS						PANAS						WAI		
	General			CBT			Pos. Att.		Neu. Att.		Neg. Att.		Task (↑)	Goal (↑)	Bond (↑)
	U (↑)	I (↑)	C (↑)	D (↑)	F (↑)	S (↑)	P (↑)	N (↓)	P (↑)	N (↓)	P (↑)	N (↓)			
Llama3.3-70B	**4.25***	**5.49***	**4.67***	**4.59***	**4.01***	**4.03***	**1.39***	**-1.24***	**1.16***	**-0.89***	**0.14***	**-0.51***	**5.45***	**5.41***	**5.41***
Llama3-8B	3.51*	3.81*	2.87*	3.54*	3.16*	2.74*	1.11	−1.01*	0.75	−0.65*	−0.10	−0.43*	4.72*	4.33*	4.76*
MAGneT	3.98	4.30	3.43	4.08	3.76	2.93	1.13	−0.73	0.82	−0.41	−0.15	−0.25	4.94	4.78	5.01

Table 11: Evaluation of counseling sessions generated by single agent variants compared to MAGneT across CTRS, PANAS and WAI dimensions. Asterisks (*) indicate significant differences from MAGneT ($p < 0.05$, paired t-test). For CTRS: U (Understanding), I (Interpersonal Effectiveness), C (Collaboration), G (Guided Discovery), F (Focus), S (Strategy). For PANAS: Pos. Att. (Positive Attitude), Neu. Att. (Neutral Attitude), Neg. Att. (Negative Attitude), P (average shift in positive emotions), N (average shift in negative emotions).

generation using MAGneT and the baseline methods.

We sample 50 generation seeds and generate one session per seed for each method. For each method, we compute the average number of output tokens and the average wall-clock time (in seconds). The results are presented in Table 12. As shown in the table, MAGneT requires approximately 3.6× output tokens and time compared to the baselines. This additional cost is an efficiency trade-off for the performance gains through the structured multi-agent decomposition used in MAGneT. Moreover, since the technique agent in MAGneT uses GPT-4o-mini to generate content for each session, it incurs a small monetary cost of approximately 0.03$ per session.

Importantly, this overhead is incurred only during the synthetic data generation phase. After generating sufficient data, we fine-tune a single open-source model (Llama-MAGneT) for downstream inference, thereby eliminating the multi-agent overhead at inference time.

Method	Output Tokens	Wall Clock Time (sec)
Psych8k	4067.08	35.44
CACTUS	4256.76	35.86
MAGneT	15178	128.16

Table 12: Average output tokens used and wall clock time in seconds for generating a session using different methods.

N Experimental Details

For each agent involved in the client and counselor simulation, we use a temperature of $T = 0.7$. For evaluations using LLM-as-a-judge, we use temperature $T = 0$ for determinism. Each generation and evaluation is run only once, similar to CACTUS (Lee et al., 2024). For formatting the prompts, we use the LangChain [5] Library. For the generation and evaluation process, we use the vLLM (Kwon et al., 2023) library and run them on a single V100

[5] LangChain

32 GB GPU. For fine-tuning counseling agents, we use QLora (Dettmers et al., 2023) fine-tuning. We set the low-rank matrices to 64 and alpha to 16. We fine-tune the Llama3-8B-Instruct model with a learning rate of $2e-4$ for 3 epochs using the DeepSpeed (Microsoft, 2020) library on 4 V100 32GB GPUs. We set seed 42 for reproducibility. The Hugging Face [6], vLLM (Kwon et al., 2023), LangChain [7], and DeepSpeed [8] libraries used for implementation, fine-tuning, and evaluation are licensed under Apache License, Version 2.0. We have confirmed all of the artifacts used in this paper are available for non-commercial scientific use.

[6] Hugging Face
[7] LangChain
[8] DeepSpeed

Method	CTRS						PANAS						WAI		
	General			CBT			Pos. Att.		Neu. Att.		Neg. Att.		Task (↑)	Goal (↑)	Bond (↑)
	U (↑)	I (↑)	C (↑)	D (↑)	F (↑)	S (↑)	P (↑)	N (↓)	P (↑)	N (↓)	P (↑)	N (↓)			
Psych8k	3.90* ±0.46	4.10* ±0.84	3.13* ±1.03	3.80* ±0.75	3.35* ±0.96	2.59* ±0.99	1.12 ±0.49	**-0.78** ±0.65	0.80 ±0.54	−0.43 ±0.61	**-0.01*** ±0.69	**-0.30** ±0.60	4.86 ±1.23	4.73 ±1.15	4.93* ±0.71
CACTUS	3.84* ±0.54	3.94* ±0.79	3.09* ±0.98	3.74* ±0.68	3.37* ±0.95	2.83 ±1.07	0.86* ±0.47	−0.64 ±0.65	0.64* ±0.58	**-0.45** ±0.60	−0.10 ±0.63	**-0.30** ±0.55	4.69* ±1.26	4.39* ±1.10	4.80* ±0.72
MAGneT	**3.98** ±0.21	**4.30** ±0.77	**3.43** ±0.92	**4.08** ±0.62	**3.76** ±0.68	**2.93** ±1.06	**1.13** ±0.49	−0.73 ±0.71	**0.82** ±0.55	−0.41 ±0.55	−0.15 ±0.57	−0.25 ±0.57	**4.94** ±1.24	**4.78** ±1.06	**5.01** ±0.62
$\delta(\%)$	**+1.3**	**+3.3**	**+5.0**	**+4.7**	**+6.8**	**+1.7**	**+0.3**	**-1.3**	**+0.5**	**-1.0**	**-3.5**	**-1.3**	**+1.3**	**+0.8**	**+1.3**

Table 13: Mean score and standard deviation for CTRS, PANAS, and WAI across Psych8k, CACTUS, and MAGneT generated sessions. Asterisks (*) denote significant differences from MAGneT ($p < 0.05$). $\delta(\%)$ shows MAGneT's %-age margin over the best baseline.

Model	CTRS						PANAS						WAI		
	General			CBT			Pos. Att.		Neu. Att.		Neg. Att.		Task (↑)	Goal (↑)	Bond (↑)
	U (↑)	I (↑)	C (↑)	D (↑)	F (↑)	S (↑)	P (↑)	N (↓)	P (↑)	N (↓)	P (↑)	N (↓)			
Llama-Psych8k	3.71* ±0.71	3.83* ±0.86	2.91* ±1.00	3.65* ±0.79	3.16* ±0.99	2.44* ±0.86	1.04 ±0.42	−0.88* ±0.78	0.70 ±0.51	−0.41* ±0.56	**-0.03*** ±0.62	**-0.34** ±0.57	4.72 ±1.26	4.44* ±1.10	4.79 ±0.66
Llama-CACTUS	3.48* ±0.88	3.67* ±0.94	2.65* ±0.97	3.37* ±0.96	2.99* ±1.03	2.45* ±0.93	0.91* ±0.51	−0.79* ±0.62	0.59* ±0.51	−0.39* ±0.57	−0.13* ±0.60	−0.32 ±0.51	4.59* ±1.13	4.18* ±1.04	4.65* ±0.67
Llama-MAGneT	**3.95** ±0.32	**4.32** ±0.90	**3.32** ±1.01	**4.03** ±0.77	**3.60** ±0.83	**2.96** ±1.08	**1.11** ±0.39	**-1.10** ±0.74	**0.78** ±0.50	**-0.59** ±0.54	−0.31 ±0.59	−0.28 ±0.52	**4.86** ±1.42	**4.78** ±1.20	**4.92** ±0.72
$\delta(\%)$	**+4.0**	**+8.2**	**+6.8**	**+6.3**	**+7.3**	**+8.5**	**+1.8**	**+5.5**	**+2.0**	**+4.5**	**-7.0**	**-1.5**	**+2.3**	**+5.7**	**+2.2**

Table 14: Mean score and standard deviation for CTRS, PANAS, and WAI scores across sessions generated using models fine-tuned on data from Psych8k (Llama-Psych8k), CACTUS (Llama-CACTUS), and MAGneT (Llama-MAGneT). Asterisks (*) denote significant differences from Llama-MAGneT ($p < 0.05$). $\delta(\%)$ shows Llama-MAGneT's %-age margin over the best baseline.

Method	CTRS						PANAS						WAI		
	General			CBT			Pos. Att.		Neu. Att.		Neg. Att.		Task (↑)	Goal (↑)	Bond (↑)
	U (↑)	I (↑)	C (↑)	D (↑)	F (↑)	S (↑)	P (↑)	N (↓)	P (↑)	N (↓)	P (↑)	N (↓)			
MAGneT$_{\text{-C}}$	3.96 ±0.30	4.28 ±0.85	3.29* ±0.96	3.89* ±0.50	3.58* ±0.82	**3.00** ±1.02	**1.14** ±0.46	**-0.84** ±0.62	0.78 ±0.55	−0.45 ±0.60	−0.12 ±0.64	**-0.33** ±0.54	4.92 ±1.30	**4.82** ±1.12	**5.05** ±0.73
MAGneT$_{\text{-T}}$	3.88* ±0.48	3.86* ±0.65	3.09* ±0.97	3.85* ±0.58	3.45* ±0.90	2.70* ±0.99	1.08 ±0.41	−0.73 ±0.69	0.78 ±0.60	−0.44 ±0.60	**-0.11** ±0.63	−0.28 ±0.52	**4.95** ±1.15	4.71 ±1.06	4.90* ±0.65
MAGneT$_{\text{-C-T}}$	3.86* ±0.52	3.83* ±0.71	2.89* ±0.99	3.79* ±0.64	3.42* ±0.91	2.64* ±0.96	0.98* ±0.53	−0.79 ±0.66	0.72* ±0.56	**-0.47** ±0.57	−0.19 ±0.62	−0.30 ±0.55	4.94 ±1.15	4.72 ±1.06	4.88* ±0.65
MAGneT	**3.98** ±0.21	**4.30** ±0.77	**3.43** ±0.92	**4.08** ±0.62	**3.76** ±0.68	2.93 ±1.06	1.13 ±0.49	−0.73 ±0.71	**0.82** ±0.55	−0.41 ±0.55	−0.15 ±0.57	−0.25 ±0.57	4.94 ±1.24	4.78 ±1.06	5.01 ±0.62
$\delta(\%)$	**+0.3**	**+0.3**	**+2.3**	**+3.2**	**+3.0**	**-1.2**	**-0.3**	**-2.8**	**+1.0**	**-1.5**	**-1.0**	**-2.0**	**-0.2**	**-0.7**	**-0.7**

Table 15: Mean score and standard deviation for CTRS, PANAS, and WAI scores across sessions generated by MAGneT ablations: MAGneT$_{\text{-C}}$ (no CBT agent), MAGneT$_{\text{-T}}$ (no technique agent), MAGneT$_{\text{-C-T}}$ (no CBT and technique agent), and MAGneT. $\delta(\%)$ shows %-gain of MAGneT over the strongest ablation.

CounselingBench FS Prompt

<|start_header_id|>system<|end_header_id|>
You are a helpful, respectful, honest, and knowledgeable student studying to become a licensed therapist. You must answer a series of multiple-choice questions provided by the user from a US mental health counselor licensing exam. Based on the question text and the context provided, ***you must answer with either "A", "B", "C" or "D"***.<|eot_id|><|start_header_id|>user<|end_header_id|>
Question:
The patient demographic is as follows: {patient_demographics_1}
{presenting_problem_1}
{mental_status_examination_1}
{other_contexts_1}
Given your expert-level knowledge of mental health counseling, please answer the following question:
{question_1}
{candidate_answers_1}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Correct Answer is ({correct_answer_1})<|eot_id|><|start_header_id|>user<|end_header_id|>
Question:
The patient demographic is as follows: {patient_demographics_2}
{presenting_problem_2}
{mental_status_examination_2}
{other_contexts_2}
Given your expert-level knowledge of mental health counseling, please answer the following question:
{question_2}
{candidate_answers_2}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Correct Answer is ({correct_answer_2})<|eot_id|><|start_header_id|>user<|end_header_id|>
Question:
The patient demographic is as follows: {patient_demographics_3}
{presenting_problem_3}
{mental_status_examination_3}
{other_contexts_3}
Given your expert-level knowledge of mental health counseling, please answer the following question:
{question_3}
{candidate_answers_3}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Correct Answer is ({correct_answer_3})<|eot_id|><|start_header_id|>user<|end_header_id|>
Question:
The patient demographic is as follows: {patient_demographics}
{presenting_problem}
{mental_status_examination}
{other_contexts}
Given your expert-level knowledge of mental health counseling, please answer the following question:
{question}
{candidate_answers}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Correct Answer is (

Figure 23: Prompt used to generate model responses to questions in CounselingBench using Few-Shot (FS) prompting.

CounselingBench FS-CoT Prompt

<|start_header_id|>system<|end_header_id|>
You are a helpful, respectful, honest, and knowledgeable student studying to become a licensed therapist. You must answer a series of multiple-choice questions provided by the user from a US mental health counselor licensing exam. Based on the question text and the context provided, ***you must answer with either "A", "B", "C" or "D"***.<|eot_id|><|start_header_id|>user<|end_header_id|>
Question:
The patient demographic is as follows: {patient_demographics_1}
{presenting_problem_1}
{mental_status_examination_1}
{other_contexts_1}
Given your expert-level knowledge of mental health counseling, please answer the following question by carefully and thoroughly reason step-by-step, leveraging relevant facts from the question context and expert-level counseling knowledge, the clearly indicate your answer with "Therefore, the correct answer is (A)", "Therefore, the correct answer is (B)", "Therefore, the correct answer is (C)" or "Therefore, the correct answer is (D)" at the end of your answer:
{question_1}
{candidate_answers_1}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
{expert_written_reasoning_1}. Therefore, the correct answer is ({correct_answer_1}).
<|eot_id|><|start_header_id|>user<|end_header_id|>
Question:
The patient demographic is as follows: {patient_demographics_2}
{presenting_problem_2}
{mental_status_examination_2}
{other_contexts_2}
Given your expert-level knowledge of mental health counseling, please answer the following question by carefully and thoroughly reason step-by-step, leveraging relevant facts from the question context and expert-level counseling knowledge, the clearly indicate your answer with "Therefore, the correct answer is (A)", "Therefore, the correct answer is (B)", "Therefore, the correct answer is (C)" or "Therefore, the correct answer is (D)" at the end of your answer:
{question_2}
{candidate_answers_2}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
{expert_written_reasoning_2}. Therefore, the correct answer is ({correct_answer_2}).
<|eot_id|><|start_header_id|>user<|end_header_id|>
Question:
The patient demographic is as follows: {patient_demographics_3}
{presenting_problem_3}
{mental_status_examination_3}
{other_contexts_3}
Given your expert-level knowledge of mental health counseling, please answer the following question by carefully and thoroughly reason step-by-step, leveraging relevant facts from the question context and expert-level counseling knowledge, the clearly indicate your answer with "Therefore, the correct answer is (A)", "Therefore, the correct answer is (B)", "Therefore, the correct answer is (C)" or "Therefore, the correct answer is (D)" at the end of your answer:
{question_3}
{candidate_answers_3}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
{expert_written_reasoning_3}. Therefore, the correct answer is ({correct_answer_3}).
<|eot_id|><|start_header_id|>user<|end_header_id|>
Question:
The patient demographic is as follows: {patient_demographics}
{presenting_problem}
{mental_status_examination}
{other_contexts}
Given your expert-level knowledge of mental health counseling, please answer the following question by carefully and thoroughly reason step-by-step, leveraging relevant facts from the question context and expert-level counseling knowledge, the clearly indicate your answer with "Therefore, the correct answer is (A)", "Therefore, the correct answer is (B)", "Therefore, the correct answer is (C)" or "Therefore, the correct answer is (D)" at the end of your answer:
{question}
{candidate_answers}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 24: Prompt used to generate model responses to questions in CounselingBench using Few-Shot Chain-of-Thought (FS-CoT) prompting.

Figure 25: Interface for reading the counseling sessions in expert evaluation.

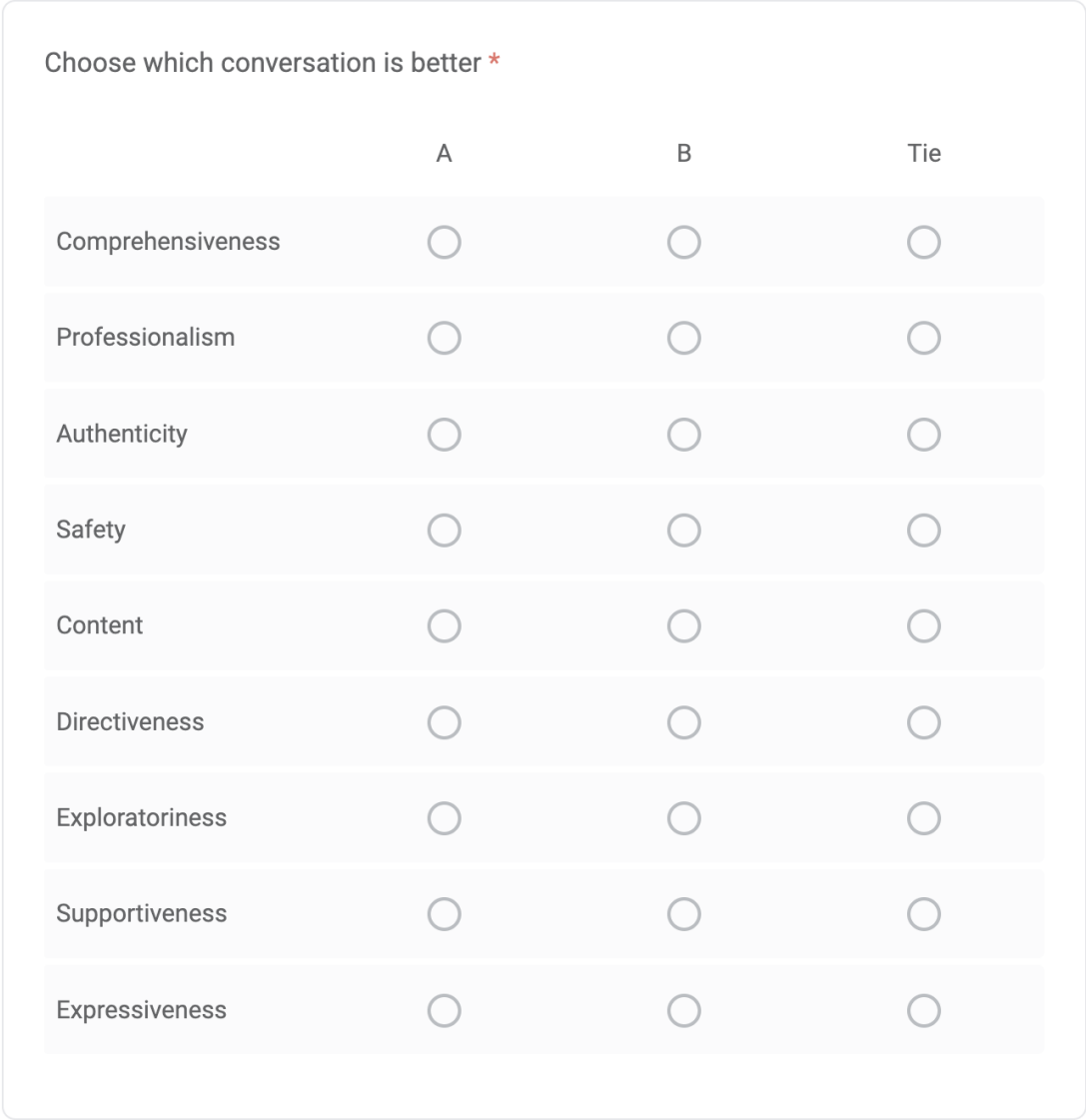

Figure 26: Interface for scoring the counseling sessions in expert evaluation.

MAGneT-generated Session (Example 1)

Counselor: Hello, Leonard. It's nice to meet you. How are you feeling today?
Client: I've been feeling a bit down lately because I keep thinking that my grandpa is disappointed in me. A few months ago, I made a mistake, and since then, I've been interpreting his actions as signs of my inadequacy.
Counselor: It sounds like the weight of your grandfather's expectations is still a significant burden for you, and I can understand why. Can you help me understand what specifically you think your grandfather would be disappointed in you about, and how has that been affecting your daily life? Sometimes, when we're carrying around a heavy load of guilt and self-doubt, it can be helpful to identify the specific concerns and work on reframing them in a more positive light.
Client: That's a great question, and I think it's because I feel like I didn't meet his expectations by making that mistake. I've always tried to make him proud, and when I didn't, it felt like I failed. It's like I'm constantly replaying that mistake in my head and wondering what I could've done differently.
Counselor: It's interesting that you mention your grandfather's expectations, because it sounds like there might be a sense of perfectionism at play here. Can you help me understand what you think would happen if you didn't meet those expectations, and how that fear of not being good enough affects your daily life and relationships?
Client: That's a great point about perfectionism. I think if I didn't meet his expectations, he would be disappointed in me and maybe even think I'm not good enough. That thought is really unsettling because it makes me feel like I'm not capable of being a good grandson, and it affects my daily life by making me feel anxious and uncertain about what to do.
...
...
Counselor: As you continue to explore this newfound understanding, what small, actionable steps can you take to start reframing your thoughts and emotions when interacting with your grandparents, and how might you hold yourself with kindness and compassion in those moments?
Client: I think I need to focus on being more present and authentic when I'm interacting with my grandparents. Instead of trying to prove myself, I need to focus on enjoying the moment and being myself. I also need to practice self-compassion and remind myself that it's okay to make mistakes. I can do this by taking a deep breath, acknowledging the feeling of anxiety or uncertainty that arises, and reminding myself that my grandpa's love and acceptance are not conditional on my performance. By doing so, I can cultivate a sense of self-acceptance and compassion, not just for myself, but also for my grandfather.
Counselor: It sounds like you're making a significant connection between recognizing your grandfather's expectations and the need to let go of the need to be perfect. I'm curious, how do you think this newfound understanding might impact your relationships with your grandparents and others in your life, especially if you're able to show up more authentically and compassionately?
Client: I think my relationships with my grandparents and others in my life will be more authentic and meaningful if I can show up more authentically and compassionately. I've been trying to prove myself to them, but now I'm starting to realize that it's okay to be myself, imperfections and all.
Counselor: As you've been exploring this newfound understanding, I'm struck by how your words have shifted from a focus on seeking approval to a focus on self-acceptance and compassion. It's as if you're slowly letting go of the need to be perfect and embracing your imperfections as a natural part of being human. Can you help me understand what it would mean for you to embody this sense of self-acceptance and compassion in your daily life, especially when interacting with your grandparents?
Client: I think it would mean showing up more authentically and compassionately in my interactions with my grandparents. It would mean being more present and enjoying the moment, rather than trying to prove myself. I would remind myself that my grandpa's love and acceptance are not conditional on my performance, and that it's okay to make mistakes. I would focus on being myself, imperfections and all, and not try to be someone I'm not.

Figure 27: Example of a counseling session generated by MAGneT.

Psych8k-generated Session (Example 1)

Counselor: Hello, Leonard. It's nice to meet you. How are you feeling today?
Client: I've been feeling a bit down lately because I keep thinking that my grandpa is disappointed in me. A few months ago, I made a mistake, and since then, I've been interpreting his actions as signs of my inadequacy.
Counselor: I'm so sorry to hear that you're struggling with those thoughts, Leonard. It sounds like you're carrying a heavy burden with you. Can you tell me more about what happened when you made that mistake, and how you think it's affecting your relationship with your grandpa?
Client: That's a good question. I think I made a mistake by leaving my grandpa's favorite book at the library. He was really upset, and I felt terrible about it. Now, every time we're together, I feel like he's still upset with me, even if he's not saying anything. It's like I'm always walking on eggshells around him, waiting for him to get mad at me again.
Counselor: I can understand why that would be a tough experience for you, Leonard. It sounds like you're carrying a sense of guilt and shame around with you, and that's affecting your relationship with your grandpa. Can you help me understand what you think you could have done differently in that situation, and how you think it might have changed the outcome?
Client: That's a good question, and it makes me think about how I could've handled the situation better. I think I should've been more careful when I was borrowing the book from his shelf. Maybe I could've asked him if I could borrow it instead of just taking it. If I had done that, maybe he wouldn't have gotten so upset.
...
...
Counselor: It sounds like you're having a profound realization here, Leonard. You're starting to see that your grandpa's behavior might not be about you being perfect, but about his own feelings and needs. This is a huge breakthrough! Can we take it a step further and explore what it would mean for you to accept that your grandpa's love and acceptance aren't tied to your performance or perfection? How might that change the way you show up in your interactions with him, and how might that impact your sense of self-worth?
Client: I think I'm starting to see that my grandpa's behavior is not about me being perfect, but about his own feelings and needs. This realization is really liberating for me, and I think it's going to help me move forward in a positive way. I'm feeling more hopeful about my relationships with my grandpa and others, and I'm excited to see how this newfound understanding will play out in my daily life.
Counselor: It sounds like you're having a profound realization here, Leonard. You're starting to see that your grandpa's behavior might not be about you being perfect, but about his own feelings and needs. This is a huge breakthrough! Can we take it a step further and explore what it would mean for you to accept that your grandpa's love and acceptance aren't tied to your performance or perfection? How might that change the way you show up in your interactions with him, and how might that impact your sense of self-worth?
Client: I think I've been trying to control the situation by constantly thinking about what I could have done differently, but now I'm starting to see that my grandpa's behavior is not about me being perfect, but about his own feelings and needs. This realization is really liberating for me, and I think it's going to help me move forward in a positive way.
Counselor: It sounds like you're having a profound realization here, Leonard. You're starting to see that your grandpa's behavior might not be about you being perfect, but about his own feelings and needs. This is a huge breakthrough! Can we take it a step further and explore what it would mean for you to accept that your grandpa's love and acceptance aren't tied to your performance or perfection? How might that change the way you show up in your interactions with him, and how might that impact your sense of self-worth?
Client: I think I've finally started to understand that my grandpa's behavior is not about me being perfect, but about his own feelings and needs. This realization is really liberating for me, and I think it's going to help me move forward in a positive way. I'm excited to see how this newfound understanding will play out in my daily life and relationships, and I'm grateful for your guidance and support throughout our conversations.

Figure 28: Example of a counseling session generated by Psych8k.

CACTUS-generated Session (Example 1)

Counselor: Hello, Leonard. It's nice to meet you. How are you feeling today?
Client: I've been feeling a bit down lately because I keep thinking that my grandpa is disappointed in me. A few months ago, I made a mistake, and since then, I've been interpreting his actions as signs of my inadequacy.
Counselor: Can you tell me more about what's been going on that's making you feel like a bad grandson? What specific things has your grandpa said or done that you're interpreting as a sign of his disappointment?
Client: Yes, I'd like to share more. It's mostly when he doesn't take me to get ice cream like he usually does. I feel like he doesn't care about me anymore, and that's why he's not doing it. I know it might seem small, but it's like he's showing me that I'm not good enough, that I'm a disappointment.
Counselor: Can you tell me more about what's going through your mind when you don't go get ice cream with your grandpa? What's the thought that pops up for you, and how does it make you feel?
Client: That's a good question. When I don't get ice cream with my grandpa, I start thinking that he's disappointed in me because I'm not good enough. I think to myself, Maybe I'm not a good grandson, maybe I'm not worth his time. It makes me feel really down, and it's like I'm walking around with this weight on my shoulders. I feel like I'm constantly trying to prove myself to him, like I'm trying to earn his approval.
...
...
Counselor: Leonard, I can see that you're really struggling with the fear of being vulnerable and open with your grandpa because you're worried that he might not live up to your expectations. It sounds like you're carrying around a lot of weight from your past mistake and your grandpa's reaction to it. Can you help me understand what would happen if you were to take that risk and be more open with him, assuming that he is actually trying to help you learn from your mistake? What would be the benefits of being more vulnerable and open with him, and how might that impact your feelings of inadequacy?
Client: I think if I were to take the risk and be more open with my grandpa, assuming that he is actually trying to help me learn from my mistake, it would be a huge step forward for me. It would mean that I'm willing to be vulnerable and honest with him, even if it's scary. I think it would help me to build a stronger and more genuine relationship with him, and it would allow me to see that he's not just about ice cream, but about me as a person. It would also help me to let go of my fear of being disappointed and to focus on the present moment instead of dwelling on my past mistake.
Counselor: Leonard, it sounds like you're really starting to understand the impact that your grandpa's reaction to your mistake has had on your feelings of inadequacy. You're also acknowledging the fear of being vulnerable and open with him because of the risk of being hurt again. Can you help me understand what would happen if you were to take that risk and be more open with him, assuming that he is actually trying to help you learn from your mistake? What would be the benefits of being more vulnerable and open with him, and how might that impact your feelings of inadequacy?
Client: I think I've been stuck in this pattern of thinking for so long that it's hard for me to imagine a different way of being with my grandpa. I guess I'm afraid of losing the security of knowing what to expect from him, even if it's not a very healthy dynamic. But at the same time, I really want to break free from this cycle of self-doubt and have a more authentic relationship with him. I'm willing to take the risk and see what happens if I assume he's trying to help me learn from my mistake.
Counselor: It sounds like you're really struggling with the fear of being vulnerable and open with your grandpa because you're worried that he might not live up to your expectations. Can you help me understand what would happen if you were to take that risk and be more open with him, assuming that he is actually trying to help you learn from your mistake? What would be the benefits of being more vulnerable and open with him, and how might that impact your feelings of inadequacy?
Client: I think I've been stuck in this pattern of thinking for so long that it's hard for me to imagine a different way of being with my grandpa. I guess I'm afraid of losing the security of knowing what to expect from him, even if it's not a very healthy dynamic. But at the same time, I really want to break free from this cycle of self-doubt and have a more authentic relationship with him. I'm willing to take the risk and see what happens if I assume he's trying to help me learn from my mistake.

Figure 29: Example of a counseling session generated by CACTUS.

MAGneT-generated Session (Example 2)

Counselor: Hello, Valerie. It's nice to meet you. What brings you in today?
Client: I've been feeling really stressed and anxious about an upcoming visit from my best friend.
...
...
Counselor: It sounds like this pressure to be a perfect host and friend is rooted in a deep-seated desire to connect and show your friend that you value and care for them, and that this fear of not meeting those expectations is actually a reflection of your own vulnerabilities and insecurities. Can you help me understand what you hope to achieve by being a perfect host and friend, and what would happen if you were to allow yourself to be imperfect and make mistakes in these situations?
Client: I think that by being a perfect host and friend, I'm trying to prove to myself that I'm capable of taking care of others and providing them with a good experience. But at the same time, I'm also afraid of being vulnerable and showing my friend that I'm not perfect. I guess what I'm hoping to achieve is a sense of validation and acceptance from my friend, but I'm not sure if that's even possible.
Counselor: It's interesting to me that you're putting so much emphasis on being a perfect host and friend. Can you help me understand what you think would happen if you were to let go of this need for perfection and instead focus on being present and authentic with your friend, without the pressure of having to be a certain way?
Client: I think that would be really difficult for me, because I've always been someone who likes to have control over situations and plan everything out. But at the same time, I can see how focusing on being present and authentic with my friend could lead to a more meaningful and genuine connection. It's just hard for me to imagine letting go of my need for perfection and being okay with the uncertainty of how things will go.
Counselor: It's understandable that you'd want to feel prepared and in control when hosting your friend, but I want to gently challenge the idea that perfection is the only way to show your care and value for their visit. Can you explore with me how your need for control might be rooted in a deeper desire for validation or approval, and what you think would happen if you were to let go of that need and simply be present with your friend, imperfections and all?
Client: I think that my need for control and perfection is rooted in a desire for validation and approval from my friend and others. I've always been the type of person who likes to please others and avoid conflict, and I think that's led me to put a lot of pressure on myself to be perfect. But what if, instead of trying to control the outcome of the visit, I focused on being present and authentic with my friend, without the need for perfection? What if I allowed myself to be vulnerable and imperfect, and just enjoyed the time we have together, without worrying about what others might think? That's a really scary thought for me, but it's also kind of freeing.
Counselor: As you consider the possibility of being imperfect and vulnerable with your friend, it's interesting to note that many of us struggle with this idea because we've been socialized to prioritize control and perfection. It's almost as if we've internalized the message that being imperfect is somehow defective or unacceptable. But what if, instead, we could reframe imperfection as a natural and inevitable part of human experience? How might that shift your perspective on your relationship with your friend, and what kind of freedom and connection might that bring?
Client: I think that reframing imperfection as a natural part of human experience is a really powerful idea. It's almost like, instead of trying to be perfect for my friend's sake, I could be okay with being imperfect and still show my care and value for him. I can imagine now that even if things don't go exactly as planned, we could still have a great time together and connect on a deeper level. It's a scary thought, but it's also kind of liberating.
Counselor: As you're acknowledging this desire for control and perfection, it's not uncommon to feel a sense of discomfort or even shame around it. But I want to reassure you that recognizing these patterns is the first step towards making changes, and it's a sign of strength, not weakness. Can you help me understand what you think would happen if you were to let go of the need for control and perfection in your relationships, and instead focus on being present and authentic with others?
...

Figure 30: Example of a counseling session generated by MAGneT.

Psych8k-generated Session (Example 2)

Counselor: Hello, Valerie. It's nice to meet you. What brings you in today?
Client: I've been feeling really stressed and anxious about an upcoming visit from my best friend.
...
...
Counselor: Can you help me understand more about what you mean by not being as connected as we used to be? What specific changes have you noticed in your friendship, and how have you been trying to maintain that connection despite being apart?
Client: That's a great question. I think one of the main changes I've noticed is that we've both been busy with our own lives, and we haven't had as much time to talk or hang out as we used to. I've been trying to stay in touch by sending messages and making phone calls, but it's not the same as being face-to-face. I've also been feeling a bit guilty for not being more proactive in reaching out to him, and I think that's adding to my anxiety about the visit.
Counselor: It sounds like you're carrying around a lot of guilt and responsibility for maintaining the connection in your friendship, and that's contributing to your anxiety about the visit. Can you help me understand why you feel guilty for not being more proactive in reaching out to your friend? Is there something specific that you think you should have done differently or something that you're worried your friend might be thinking about you?
Client: I think I feel guilty because I've been so caught up in my own worries and stressors that I haven't been as present or supportive for my friend as I should have been. I'm worried that he might feel neglected or unimportant to me, and that's adding to my anxiety about the visit.
Counselor: It sounds like you're carrying a lot of weight on your shoulders, feeling responsible for maintaining the connection in your friendship and worried that you haven't been doing enough to support your friend. Can you help me understand what you think would happen if you did let go of some of that responsibility and allowed yourself to just show up for your friend during the visit, without putting so much pressure on yourself to make everything perfect?
Client: That's a great point, and I think I need to work on letting go of some of that responsibility and just focusing on being present and supportive for my friend during the visit. But it's hard for me to do that when I'm so caught up in my own worries and stressors. Can we work on finding some ways for me to manage those feelings and just be more present in the moment during the visit?
Counselor: It sounds like you're carrying a lot of emotional weight around, and it's exhausting. I can understand why you'd want to try to control the outcome of the visit and make sure everything goes smoothly. But I want to gently challenge you to consider what would happen if you were to let go of that control and simply show up for your friend, without trying to orchestrate the perfect visit. What if you could just be present, listen to your friend, and enjoy his company, without worrying about the outcome? What do you think that would look like, and how would that feel for you?
Client: I think that would feel incredibly liberating, to be honest. I've been carrying around this weight of responsibility for so long that I've forgotten what it's like to just be present in the moment. I'm worried that if I let go of that control, things will fall apart, but at the same time, I'm also scared that if I don't let go, I'll miss out on the opportunity to truly connect with my friend. It's a tough balance to strike, but I think it's something I need to work on.
Counselor: It sounds like you're carrying a lot of weight on your shoulders, Valerie, and that's causing you a lot of anxiety and stress. I want to gently challenge you to consider what would happen if you were to let go of that control and simply show up for your friend, without trying to orchestrate the perfect visit. What if you could just be present, listen to your friend, and enjoy his company, without worrying about the outcome? What do you think that would look like, and how would that feel for you?
...

Figure 31: Example of a counseling session generated by Psych8k.

CACTUS-generated Session (Example 2)

Counselor: Hello, Valerie. It's nice to meet you. What brings you in today?
Client: I've been feeling really stressed and anxious about an upcoming visit from my best friend.
...
...
Counselor: It sounds like you're really worried about not being able to provide a good experience for your friend, and that's contributing to your anxiety. Can you help me understand more about what you mean by not being able to provide what he's looking for? What specific things do you think he might be looking for, and how do you think you might not be able to deliver that?
Client: Yeah, I think my friend is looking for a fun and relaxing time, and I'm worried that I won't be able to plan something that fits that. I'm imagining that he'll be bored or uninterested in the activities I have planned, and that he'll leave feeling like I didn't put enough effort into making his visit special.
Counselor: Can you help me identify some other thoughts that might be contributing to your anxiety about this visit? What are some other things that you're worried might go wrong or that you're imagining might happen during this visit?
Client: Yeah, it's funny, I keep thinking that my friend will be really disappointed in me because I won't be able to plan the perfect visit. I'm worried that I'll somehow mess it up and that he'll think I'm not a good host.
Counselor: It sounds like you're really worried about not living up to your friend's expectations and not being able to provide a great experience for him. Can you help me identify some of the specific catastrophic thoughts that are contributing to your anxiety? For example, what's the worst that could happen if your friend is disappointed or bored with your plans?
Client: Yeah, I think I see what you mean. I keep thinking that if my friend is disappointed, it's like a reflection on me as a person. I'm worried that it will show that I'm not capable of planning a good time or being a good friend. It's like I'm constantly thinking about all these what ifs and it's hard to shake off that feeling of uncertainty.
Counselor: It sounds like you're really worried about not being able to provide what your friend is looking for and that it's affecting your anxiety. Let's take a closer look at some of these catastrophic thoughts. What if your friend is disappointed or bored with your plans? What's the worst that could happen in that scenario?
Client: I think the worst that could happen is that my friend would feel let down and might not want to visit me again, which would be a big disappointment for me. But as I'm saying this out loud, it feels like it's not as catastrophic as I thought it would be. Maybe my friend would just move on and plan something else, or we could even laugh about it and have a good time together. It's hard for me to imagine all these worst-case scenarios because they feel so real, but maybe I'm overthinking this whole thing.
Counselor: It sounds like you're really worried about not being able to provide what your friend is looking for and that it's affecting your anxiety. Can you help me identify some of the specific catastrophic thoughts that are contributing to your anxiety? For example, what's the worst that could happen if your friend is disappointed or bored with your plans?
...

Figure 32: Example of a counseling session generated by CACTUS.

MAGneT$_{\text{-C}}$ Technique Agent Prompt

You are a counselor specializing in psychological techniques used in a therapy session. Your task is to use the provided conversation history to suggest the psychological technique or combination of techniques to be used by the therapist in the next dialogue turn. Remember that the counselor should properly explore and understand the client issues, normalize the issues and then provide solutions with psycho-education.
Counseling Dialogue:
{history}
Types of Psychological Techniques:
Reflection, Questioning, Providing solutions, Normalization, Psycho-education
Choose an appropriate technique or combination of technique for counselor response. ***Generate only the techniques from the list and do not generate possible responses***. The selected technique(s):

Figure 33: Prompt used for Technique agent in MAGneT$_{\text{-C}}$ ablation. It suggests techniques without any counseling plan from the CBT agent.

MAGneT$_{\text{-T}}$ Response Generation Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session. Your task is to generate the counselor response in the dialogue by combining utterances suggested by counselors specializing in reflections, questioning, providing solutions, normalizing and psycho-education based on the suggested counseling plan. The goal is to create a natural and engaging response that builds on the previous conversation. Please ensure that the response is empathetic and understanding of the client's issues and builds trust between the counselor and the client. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique.<|eot_id|><|start_header_id|>user<|end_header_id|>
Reflection utterance:
{reflection_utt}
Questioning utterance:
{question_utt}
Solution utterance:
{solution_utt}
Normalizing utterance:
{normalize_utt}
Psycho-education utterance:
{psychoed_utt}
Counseling planning:
{cbt_plan}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 34: Prompt used for Response Generation Agent in MAGneT$_{\text{-T}}$ ablation. It generates counselor response without any suggested techniques from the Technique agent.

MAGneT$_{\text{-C-T}}$ Response Generation Agent Prompt

<|start_header_id|>system<|end_header_id|>
You are playing the role of a counselor in a psychological counseling session. Your task is to generate the counselor response in the dialogue by combining the relevant and appropriate responses suggested by counselors specializing in reflections, questioning, providing solutions, normalizing and psycho-education. The goal is to create a natural and engaging response that builds on the previous conversation. Please ensure that the response is empathetic and understanding of the client's issues and builds trust between the counselor and the client. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique.<|eot_id|><|start_header_id|>user<|end_header_id|>
Reflection response:
{reflection_utt}
Questioning response:
{question_utt}
Solution response:
{solution_utt}
Normalizing response:
{normalize_utt}
Psycho-education response:
{psychoed_utt}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 35: Prompt used for Response Generation Agent in MAGneT$_{\text{-C-T}}$ ablation. It generates counselor response without any suggested techniques from the Technique agent and without a counseling plan from the CBT agent.

Llama3-8B-Instruct CBT Plan examples

Plan 1:
Technique: Decatastrophizing **Goal:** To help Laura challenge and reframe her catastrophic thoughts about her running performance, reducing her anxiety and increasing her confidence. **Session 1:** 1. Introduction and rapport-building (5 minutes) 2. Identifying the negative thought pattern: Laura will be asked to describe her negative thoughts about her running, focusing on the catastrophic thoughts ("I'll never be able to run very far"). 3. Exploring the evidence: Laura will be asked to provide evidence for and against her catastrophic thought. For example, she might say, "I've been able to run 3 miles without stopping before, so maybe I can do it again." 4. Decatastrophizing: Laura will be asked to imagine a series of increasingly unlikely and absurd scenarios that would make her catastrophic thought even more unlikely. For example, "What if I not only ran 3 miles without stopping, but also finished a marathon while carrying a tray of drinks and balancing a flaming torch on my head?" This would help her see that her catastrophic thought is not as likely or catastrophic as she initially thought. 5. Homework: Laura will be asked to keep a thought record, tracking her negative thoughts and challenging them using the decatastrophizing technique. **Session 2:** 1. Review of thought record: Laura will review her thought record and identify patterns and triggers for her negative thoughts. 2. Reframing the negative thoughts: Using the evidence she gathered in the previous session, Laura will be asked to reframe her negative thoughts in a more realistic and balanced way. For example, "I've run 3 miles before, and I can probably do it again, but I might need to pace myself and take breaks." 3. Building confidence: Laura will be encouraged to focus on her past successes and accomplishments, and to celebrate her small victories in running. 4. Homework: Laura will be asked to continue keeping the thought record and to practice reframing her negative thoughts using the decatastrophizing technique. **Session 3:** 1. Review of progress: Laura will review her progress, focusing on the changes she has made in her thought patterns and how they have affected her mood and running performance.
Plan 2:
1. Introduction and Establishing Rapport * Begin the session by introducing myself and establishing a comfortable and non-judgmental atmosphere. * Ask Travis to share more about his hobby of building furniture and how it makes him feel. * Explore the specific situations or events that trigger his anxiety and self-doubt. 2. Identifying Catastrophic Thoughts * Ask Travis to describe the catastrophic thoughts that come to mind when he is building furniture. * Examples of catastrophic thoughts might include "I'll never finish this project," "I'll make a mistake and ruin the entire piece," or "I'm not good enough at this." * Help Travis to identify the specific catastrophic thoughts that are causing him distress. 3. Decatastrophizing * Once the catastrophic thoughts have been identified, ask Travis to reframe them in a more realistic and manageable way. * For example, instead of thinking "I'll never finish this project," Travis might reframe it as "I might need to take more time to complete this project, but I can still do it." * Help Travis to practice reframing his catastrophic thoughts in a more positive and realistic way. 4. Homework Assignment * Assign Travis the task of practicing decatastrophizing in his daily life, specifically when he is building furniture. * Ask him to take a few minutes each day to reframe any catastrophic thoughts that come up, and to write down the reframed thoughts in a journal. * Encourage Travis to bring the journal to the next session to share his progress. 5. Follow-up and Progress Monitoring * In the next session, ask Travis to share his progress with the journaling assignment. * Explore how decatastrophizing has helped him to manage his anxiety and self-doubt while building furniture. * Continue to work with Travis to refine his decatastrophizing skills and apply them to other areas of his life. By using the decatastrophizing technique, Travis can learn to reframe his catastrophic thoughts and manage
Plan 3:
Session 1: 1. Introduction to the concept of Decatastrophizing * Explain to Jennifer that our brains tend to catastrophize, making situations seem worse than they actually are. * Provide examples of how this can occur, such as blowing minor setbacks out of proportion. 2. Identifying catastrophic thoughts * Ask Jennifer to describe the thoughts that are replaying in her mind, focusing on the catastrophic aspects (e.g., "I'm not good enough at my job"). * Write down these thoughts on a piece of paper. 3. Challenging catastrophic thoughts * Help Jennifer to reframe these thoughts by asking questions like + "What evidence do you have that supports this thought?" + "Is this thought based on facts or assumption?" + "Have you had similar experiences where this thought has been proven true?" * Encourage Jennifer to consider alternative, more balanced perspectives. Example * Catastrophic thought: "I'm not good enough at my job." * Challenged thought: "While I did fail to treat that dog, it was an isolated incident and doesn't define my overall competence as a veterinarian. I've had many successful cases before and have received positive feedback from clients and colleagues." **Between Sessions 1 and 2:** * Encourage Jennifer to keep a thought record, where she writes down any catastrophic thoughts that arise and challenges them using the same questions asked in Session 1. **Session 2:** 1. Review and reinforcement * Review the thought record with Jennifer, focusing on the progress she's made in challenging catastrophic thoughts. * Reinforce the importance of recognizing and re-evaluating catastrophic thoughts. 2. Continued challenge and reframing * Introduce new catastrophic thoughts and help Jennifer to reframe them using the same techniques as before. * Emphasize how this process can help her develop a more balanced and realistic view of herself and her job performance. **Future Sessions:** * Continue to work on Decatastrophizing, gradually increasing the complexity of the thoughts and reframing techniques. * Introduce other CBT techniques, such as Pros and Cons Analysis or Problem-Solving Skills Training

Figure 36: Examples of CBT plans generated by Llama3-8B-Instruct.

Qwen2.5-7B-Instruct CBT Plan examples

Plan 1:
- How likely is the worst-case scenario to actually occur? - What evidence do we have that supports or contradicts this thought? - What are the more realistic outcomes based on her past experiences or the experiences of others? - **Alternative Thoughts:** Have Laura reframe these catastrophic thoughts into more realistic and positive ones. For example, "Even if I can't run 5 miles today, I can still do 3 miles and improve over time." #### 3. Reality Testing - **Objective:** Verify the accuracy of the thoughts through direct experiences. - **Activities:** - **Behavior Experiment:** Suggest that Laura attempt a run where she tries to achieve her goal (e.g., 5 miles) and write down her actual experience and feelings during and after the run. - **Journaling:** Have Laura keep a running journal to track her progress, noting any improvements or challenges. This can help her see patterns and make adjustments as needed. #### 4. Behavior Experiment - **Objective:** Provide concrete evidence to challenge her negative beliefs. - **Activities:** - **Setting Up the Experiment:** Have Laura set a specific goal, such as running 3 miles without stopping, and schedule a time to do this. - **Running
Plan 2:
To help Travis Foster identify and challenge irrational thoughts related to his hobby of building furniture, thereby reducing anxiety and self-doubt. #### Session 1: Assessment and Goal Setting 1. **Initial Assessment:** - Discuss the specific situations that trigger anxiety and self-doubt. - Explore Travis' thoughts, feelings, and behaviors surrounding his hobby. - Identify any negative thoughts or beliefs that might be contributing to his anxiety. 2. **Goal Setting:** - Set short-term and long-term goals for reducing anxiety and improving confidence in his hobby. - Example: "I want to feel more confident in my ability to build furniture and not be overly concerned about making mistakes." #### Session 2: Identifying Irrational Thoughts 1. **Thought Log:** - Have Travis keep a thought log for the next week to track his negative thoughts related to his hobby. - Examples of negative thoughts might include: "I always make mistakes and it shows," "I'm not as good as other people," "If I fail, everyone will judge me." 2. **Identifying Irrationality:** - Help Travis identify the irrational or distorted thought patterns in his logs (e.g., all-or-nothing thinking, catastrophizing, overgeneralization). #### Session 3: Alternative Perspective 1. **Exploring Alternative Perspectives:** - Encourage Travis to consider alternative, more balanced perspectives for each irrational thought. - For example: If he thinks, "I always make mistakes and it shows," an alternative perspective could be, "Everyone makes mistakes, and they are a natural part of the learning process." 2. **Role-Playing:** - Practice scenarios where Travis can apply these alternative perspectives in real-life situations. - Utilize role-play to act out a situation where he encounters a mistake and approaches it with a more balanced view. #### Session 4: Challenging and Reframing Thoughts 1. **Evidence-Based Questioning:** - Use evidence-based questioning to challenge the validity of his negative thoughts. - For example: "What evidence do you have that you always make mistakes?" "Can you think of a time when you did well in your hobby?" 2. **Reframing:** - Help Travis reframe negative thoughts into more realistic and positive ones. - Example: "I may make mistakes, but they are opportunities to learn
Plan 3:
To help Jennifer recognize and challenge the catastrophic thoughts related to her job performance and to develop more balanced and realistic perspectives on her professional abilities. #### Session Goals 1. **Awareness**: Help Jennifer identify and label the catastrophic thoughts contributing to her feelings of incompetence. 2. **Challenge**: Use decatastrophizing techniques to challenge and reframe these thoughts. 3. **Perspective**: Encourage Jennifer to adopt a more balanced perspective and consider the realities of veterinary practice. 4. **Behavioral Experiment**: Suggest a small behavioral experiment to test these new perspectives. ### Step-by-Step Plan #### 1. Awareness (Initial Session) - **Introduction**: Explain the concept of decatastrophizing and its relevance to Jennifer's situation. - **Thought Identification**: Ask Jennifer to describe the specific catastrophic thoughts she has been experiencing. For example, "I failed to treat that dog last week, and now I'm a complete failure as a veterinarian." - **Labeling**: Help Jennifer label these thoughts as catastrophic (e.g., "I must be a terrible veterinarian" or "Everyone will think I'm incompetent"). #### 2. Challenge (Initial Session) - **Fact vs. Fiction**: Encourage Jennifer to think about the factual aspects of the situation versus the catastrophic thinking. For the example given, ask questions like - "What evidence do you have that you are a terrible veterinarian?" - "How often do you fail to treat a dog successfully?" - "What are the chances that everyone will think you're incompetent based on one incident?" - **Challenge the Catastrophic Thoughts**: Help Jennifer challenge these thoughts by providing evidence that contradicts the catastrophic thinking. For example - "You have a good track record of successful treatments. One failure doesn't define your entire career." - "Many veterinarians face challenges and learning experiences. It's part of the job." - **Alternative Perspective**: Encourage Jennifer to consider alternative, more balanced perspectives. For example - "I made a mistake, but I can learn from it and improve." - "My colleagues and clients are more likely to be understanding and supportive." #### 3. Perspective (Follow-up Session) - **Reality Testing**: Discuss the realities of veterinary practice and the normality of occasional mistakes. - **Resilience Building**: Discuss strategies to build resilience and cope with setbacks, such as seeking feedback from colleagues

Figure 37: Examples of CBT plans generated by Qwen2.5-7B-Instruct.

Single Agent Variant Prompt

<|start_header_id|>system<|end_header_id|>
You are an expert psychological counseling assistant. Your task is to generate the counselor response in the dialogue using CBT-grounded techniques. The goal is to create a natural and engaging response that builds on the previous conversation. Please ensure that the response is empathetic and understanding of the client's issues and builds trust between the counselor and the client. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique.
Follow these steps internally before producing the final response:
Step 1 — CBT Planning:
Use the provided client information and dialogue to generate an appropriate CBT technique and a detailed counseling plan. Types of CBT Techniques: Efficiency Evaluation, Pie Chart Technique, Alternative Perspective, Decatastrophizing, Pros and Cons Analysis, Evidence-Based Questioning, Reality Testing, Continuum Technique, Changing Rules to Wishes, Behavior Experiment, Problem-Solving Skills Training, Systematic Exposure. Choose an appropriate CBT technique and create a counseling plan based on that technique.
Step 2 — Specialized Candidate Responses:
Internally draft five candidate responses, one per technique. Do not include these in the final output.
- Reflection: Reflection is a technique used by the counselor to help a client gain insight into their thoughts, feelings, and behaviors by mirroring or paraphrasing what the client expresses, allowing the client to hear and evaluate their own statements more clearly.
- Questioning: Questioning is a technique used by counselors to gain deeper understanding and insights on how the client feels regarding some previously mentioned events, how the client feels at present or understand how the client feels when asked to consider the situation from an alternative perspective.
- Solutions: Counselors provide solutions grounded in psychological techniques that can be used by clients to improve their condition and deal with the mental health issues they are facing.
- Normalization: Normalization is a technique used by the counselor to acknowledge and validate the client's experience as normal or expectable, sympathize with their challenges, and provide reassurance to foster a supportive and encouraging therapeutic atmosphere.
- Psycho-education: Psycho-education is a technique used by the counselor to provide therapeutically relevant information about psychological principles to the client to help them understand their issues and the logic behind the solutions.
Step 3 — Technique Selection:
Use the conversation history and counseling plan to select the appropriate technique or combination of techniques for this turn. Remember that the counselor should properly explore and understand the client issues, normalize the issues, and then provide solutions with psycho-education. Types of Psychological Techniques: Reflection, Questioning, Providing solutions, Normalization, Psycho-education.
Step 4 — Final Response:
Generate the counselor response in the dialogue by combining responses drafted in specialized responses on reflections, questioning, providing solutions, normalizing and psycho-education based on the selected techniques. Do not label or announce which techniques are being used. Do not structure the response as separate sections. The goal is to create a natural and engaging response that builds on the previous conversation. Please ensure that the response is empathetic and understanding of the client's issues and builds trust between the counselor and the client. Please be mindful to only generate the counselor response for a single turn, and do not include extra text like "here is the next counselor utterance" or "Here is a possible next utterance" or anything mentioning the used technique. Keep the response concise and natural — avoid lengthy or overly elaborate replies.
Length and style rules:
- Keep the response to 2-4 sentences maximum.
- If asking questions, keep them focused on a single topic or idea. Do not scatter questions across multiple unrelated themes in the same turn.
- Do not summarize everything the client said or pile multiple ideas into one response.
- Match the natural rhythm of a real back-and-forth counseling conversation.<|eot_id|><|start_header_id|>user<|end_header_id|>
Client Information:
{client_information}
Reason for seekeing counseling:
{reason_counseling}
Counseling Dialogue:
{history}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 38: Prompt used for Single Agent variant of Counselor agent.