Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy

XIN SUN, University of Amsterdam, The Netherlands and National Institute of Informatics (NII), Japan
JAN DE WIT, Tilburg University, The Netherlands
ZHUYING LI, Southeast University, China
JIAHUAN PEI, Vrije Universiteit Amsterdam, The Netherlands
ABDALLAH EL ALI, Centrum Wiskunde & Informatica (CWI) and Utrecht University, The Netherlands
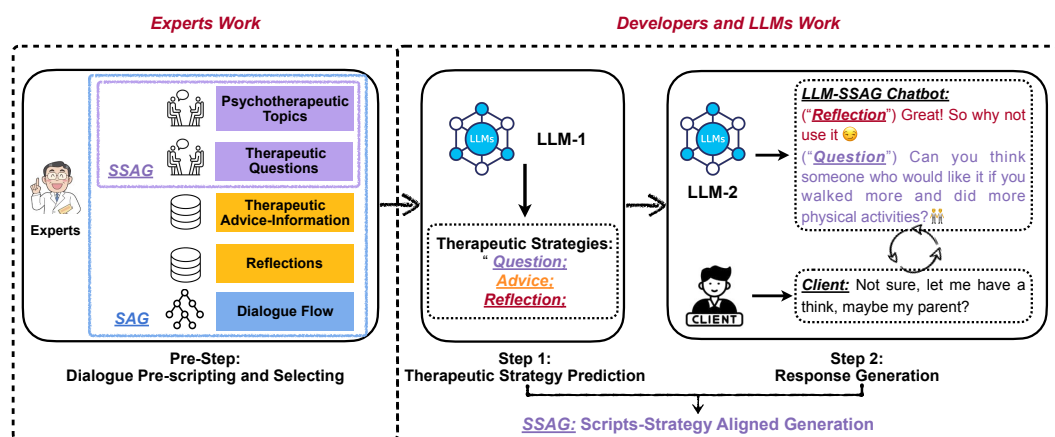JOS A. BOSCH, University of Amsterdam, The Netherlands

Fig. 1. Visualization of SSAG and its comparison with SAG. In SSAG, experts provide core therapeutic content, while LLMs first predict therapeutic strategy and then generate responses accordingly. In comparison, SAG relies on fully expert-crafted dialogue scripts.

Chatbots or conversational agents (CAs) are increasingly used to improve access to digital psychotherapy. Many current systems rely on rigid, rule-based designs, heavily dependent on expert-crafted dialogue scripts for guiding therapeutic conversations. Although advances in large language models (LLMs) offer potential for more flexible interactions, their lack of controllability and explanability poses challenges in high-stakes contexts like psychotherapy. To address this, we conducted two studies in this work to explore how aligning LLMs with expert-crafted scripts can enhance psychotherapeutic chatbot performance. In Study 1 (N=43), an online experiment with a within-subjects design, we compared rule-based, pure LLM, and LLMs aligned with expert-crafted scripts via fine-tuning and prompting. Results showed that aligned LLMs significantly

Authors' Contact Information: Xin Sun, x.sun2@uva.nl, University of Amsterdam, Amsterdam, , The Netherlands and and National Institute of Informatics (NII), Tokyo, , Japan; Jan de Wit, Tilburg University, Tilburg, , The Netherlands; Zhuying Li, Southeast University, China; Jiahuan Pei, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; Abdallah El Ali, Centrum Wiskunde & Informatica (CWI) and Utrecht University, Amsterdam, The Netherlands; Jos A. Bosch, University of Amsterdam, Amsterdam, , The Netherlands.

outperformed the other types of chatbots in empathy, dialogue relevance, and adherence to therapeutic principles. Building on findings, we proposed "Script-Strategy Aligned Generation (SSAG)", a more flexible alignment approach that reduces reliance on fully scripted content while maintaining LLMs' therapeutic adherence and controllability. In a 10-day field Study 2 (N=21), SSAG achieved comparable therapeutic effectiveness to full-scripted LLMs while requiring less than 40% of expert-crafted dialogue content. Beyond these results, this work advances LLM applications in psychotherapy by providing a controllable and scalable solution, reducing reliance on expert effort. By enabling domain experts to align LLMs through high-level strategies rather than full scripts, SSAG supports more efficient co-development and expands access to a broader context of psychotherapy.

CCS Concepts: • **Human-centered computing** → **Computer supported cooperative work**.

Additional Key Words and Phrases: Large language model, Alignment, Chatbot-delivered psychotherapy

1 Introduction

Chatbots, or conversational agents (CAs), are increasingly used in psychotherapy for behavioral interventions. These agents utilize evidence-based conversational techniques such as Motivational Interviewing (MI) [59] and Cognitive Behavioral Therapy (CBT) [9] to provide round-the-clock mental health support [3, 41, 42, 45, 70, 85, 99]. Traditionally, these agents have relied on rule-based approaches [58] grounded in expert-crafted scripts [82, 93] to ensure therapeutic adherence and safety. While effective, these approaches often produce rigid, non-adaptive conversations and require substantial expert effort to not only author dialogue content but also design the structured dialogue flows, as illustrated in Fig 2. Recent advances in Natural Language Generation (NLG) [25] have enabled more dynamic chatbot interactions. For example, MI chatbots have incorporated rephrasing and template-based generation[23] to improve engagement [3, 41, 62], while models trained on CBT data [45] demonstrated the potential to deliver structured and principle-based therapeutic dialogues. Hybrid approaches [6, 86] have emerged to integrate LLMs with rule-based systems, with the aim of combining generative flexibility with expert-guided structure and safety. However, these systems still require large domain-specific datasets, which are scarce in psychotherapy due to privacy concerns and data sensitivity.

The rise of large language models (LLMs) [14] opens new opportunities for personalized, empathetic, and engaging digital psychotherapy [64] and mental heath treatments [43]. Psychotherapy requires a real-time adaptation to subtle motivational, emotional, and behavioral cues. Unlike simpler NLP models, advanced LLMs like ChatGPT [66] are capable of generating responses that are naturally fluent, contextually appropriate, and emotionally attuned. Psychotherapeutic strategies such as reflective listening, a core element in MI and CBT, demand nuanced interpretation and reflective response, which requires far more than surface-level text manipulation such as paraphrasing or summarization. Furthermore, managing therapeutic dialogues involves pacing, topic transitions, goal alignment, and not merely following scripted dialogue flows. Yet, applying LLMs in the psychotherapeutic context to develop LLM-powered or hybrid [6] psychotherapy chatbots still poses three technical challenges: **(1)** LLMs lack domain-specific knowledge to initiate and sustain conversations on specific *psychotherapeutic topics* (see Fig 4), such as increasing intrinsic motivation, or cognitive behavioral practices [9] like "mindfulness" [31], and cognitive distortions such as "should statements" [1]. **(2)** LLMs also struggle to generate structured and goal-directed *therapeutic questions* (see Fig 2) that guide behavioral interventions. Moreover, the reliance on

fully expert-crafted dialogue scripts presents another challenge: **(3)** Creating these expert-crafted dialogue scripts is time-consuming and labor-intensive, posing barriers to scalability and broader adoption in real-world therapeutic contexts.
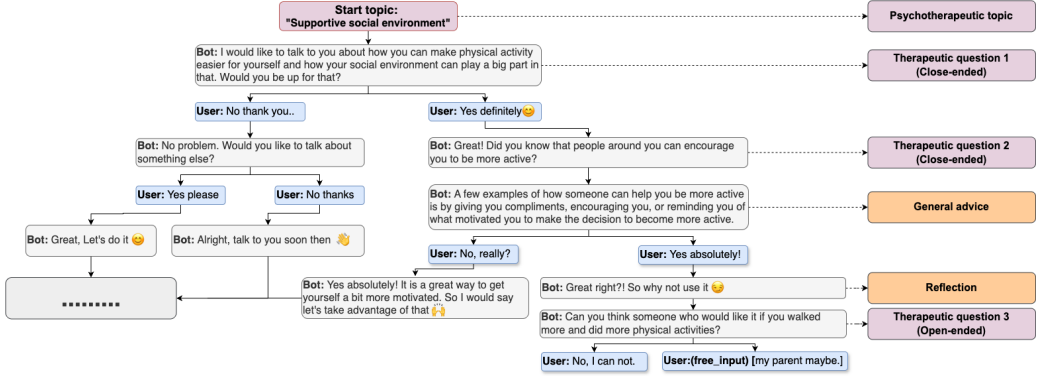

Fig. 2. A dialogue example: parts of tree-structured dialogue scripts pre-crafted by experts under the psychotherapeutic topic "Supportive social environment" for behavioral intervention.

Despite the advancements of LLMs, it remains unclear whether expert-crafted dialogue scripts are still necessary to guide LLMs for psychotherapy, specifically in balancing generative flexibility with the structured and goal-oriented nature of evidence-based therapeutic conversations. This leads to our first research question. *RQ1: Do expert-crafted dialogue scripts remain essential for chatbot-delivered psychotherapy in the era of LLMs?* To explore this question, we propose Script-Aligned Generation (SAG), a concept that aligns LLMs with expert-crafted dialogue scripts to retain therapeutic structure while enabling conversational flexibility. To rigorously examine RQ1 and validate SAG, we define two sub-questions: *RQ1.1: How do LLM-powered chatbots aligned with expert-crafted dialogue scripts compare to rule-based chatbots and pure LLMs for psychotherapy?* and *RQ1.2: How do LLMs aligned via fine-tuning and prompting differ in performing psychotherapy?* To answer these questions and address the first two challenges, we conducted Study 1 (Fig 3.a), comparing four chatbot types: a rule-based chatbot using static expert-crafted scripts; a pure LLM without alignment; and two LLM-powered chatbots aligned with expert-crafted scripts (i.e., SAG) via either fine-tuning or prompting. Results showed that LLM-powered chatbots by SAG significantly outperformed both rule-based and pure LLMs across key assessing metrics, including linguistic quality, therapeutic relevance, empathy, engagement, MI adherence, and motivation enhancement. These findings highlighted the continuing importance of expert-crafted scripts in enabling LLMs to deliver safe and therapeutically effective psychotherapy.

Although SAG aims to enable LLMs to balance conversational flexibility with therapeutic effectiveness by aligning them with expert-crafted dialogue scripts, it still relies on fully expert-crafted scripts, posing scalability and cost challenges for real-world psychotherapy applications. To address challenge (3), we proposed Script-Strategy Aligned Generation (SSAG), a more flexible and collaborative alignment approach. Unlike SAG, SSAG requires only partial expert input: core psychotherapeutic topics, key therapeutic questions, and optional advice (see Fig 1). Drawing on prior work in CSCW [78, 100] and NLP [89] that employ therapeutic strategies to guide human therapists and LLMs for facilitating effective therapy, SSAG enabled LLMs to dynamically generate dialogues aligned with these therapeutic strategies, such as asking questions, reflective listening, and giving advice, enabling alignments with both partial expert-crafted scripts and evidence-based therapeutic strategies [61, 78]. We therefore ask our second research question: *RQ2: Can psychotherapy*
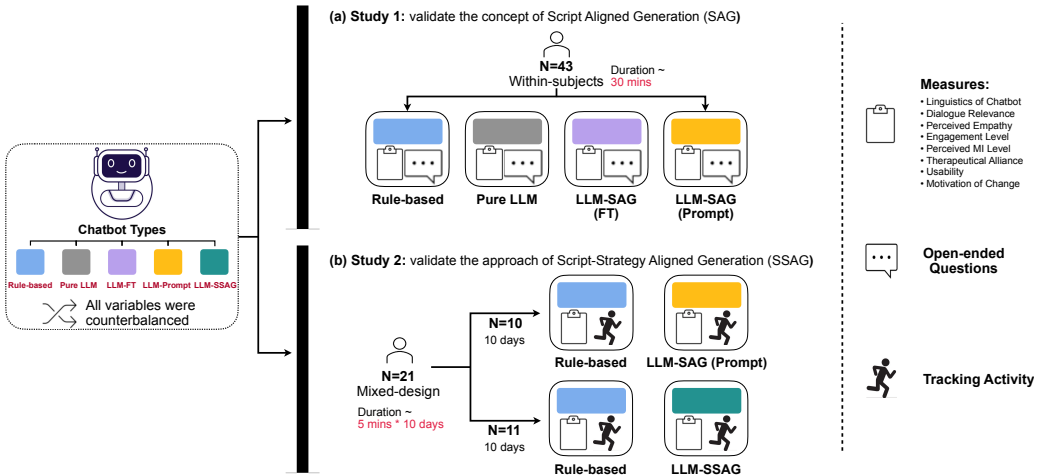
Fig. 3. The procedure and design of the two studies conducted to evaluate the chatbots in delivering the psychotherapy for health behavioral intervention

chatbots using SSAG achieve comparable conversational quality and therapeutic effectiveness to those using SAG? To evaluate this, we conducted Study 2 (Fig 3.b), a 10-day field study comparing three chatbots: (1) a rule-based chatbot as baseline, (2) an SAG-aligned chatbot (via prompting) as in Study 1, and (3) an SSAG-aligned chatbot. Results showed that SSAG matched SAG in therapeutic effectiveness while sustaining empathy and engagement, suggesting that flexible alignment can deliver comparable outcomes with less expert scripts.

To distinguish between SAG and SSAG, we defined SAG as strict alignment that relies on fully scripted dialogue content and flows, whereas SSAG is a flexible alignment approach that requires only partial expert input. SSAG enabled domain experts to contribute at a higher level of abstraction by specifying key therapeutic elements (e.g., psychotherapeutic topics and questions as illustrated in Fig 1) rather than scripting the entire dialogues. This leads to our third research question: *RQ3: To what extent can SSAG reduce reliance on expert-scripted content for developing psychotherapy chatbots?* Findings from Study 2 showed that SSAG chatbots achieve comparable performance to those using SAG while reducing the need for fully expert-crafted dialogue scripts, positioning LLMs as co-authors of therapeutic dialogues. SSAG is situated within CSCW by facilitating collaboration among psychotherapy experts, developers, and LLMs. This helps enable the scalable, co-designed development of chatbots that have the potential to be both engaging and therapeutically effective for health behavioral interventions.

We chose to focus on MI and CBT due to their strong evidence base and complementary structures. MI, a flexible, client-centered approach, emphasizes reflective listening and open-ended questioning to enhance intrinsic motivation [4, 56, 59, 60], making it well-suited for evaluating the ability of LLMs to support engaging and context-aware conversations. CBT, by comparison, offers structured, goal-oriented dialogues like cognitive reframing [9], aligning well with LLMs guided by specific therapeutic objectives. By combining MI's flexibility with CBT's structure, we can better explore whether aligned LLMs can uphold therapeutic adherence without sacrificing perceived conversational quality such as empathy, a tradeoff investigated in prior work about CBT [45, 90]. Importantly, SSAG can be generalized beyond MI and CBT. By grounding generation in high-level therapeutic strategies, such as asking, reflecting, and advising, identified in prior CSCW work [78, 100], SSAG can be extended to other frameworks like Acceptance and Commitment Therapy (ACT) [39] or digital mental health coaching [24, 77] by substituting context-appropriate,

expert-defined strategies. This positions SSAG as a scalable, adaptable approach for a wide range of evidence-based digital health interventions.

This work contributes to CSCW by advancing the effective development of LLM-powered psychotherapy chatbots for health behavioral interventions. First, we proposed and evaluated Script-Aligned Generation (SAG), a concept that aligns LLMs with fully expert-crafted scripts to balance therapeutic effectiveness and conversational flexibility, highlighting the ongoing importance of human expertise in guiding LLMs. Second, we introduced Script-Strategy Aligned Generation (SSAG), a more flexible approach that reduces reliance on expert-scripted content by requiring only partial human input (e.g., therapeutic topics and questions), enabling efficient development of psychotherapy chatbots supported by LLMs. Third, we contributed the first dataset of expert-authored psychotherapy dialogues grounded in evidence-based techniques (MI and CBT), supporting future research on LLM alignment for psychotherapy. Together, these contributions demonstrate that expert guidance remains critical for safe and effective digital psychotherapy, and that SSAG offers a practical path toward scalable, controllable, and efficient development of therapeutically aligned chatbots in the era of LLMs.

2 Related Work

2.1 Chatbots and Conversational Design for Digital Psychotherapy and Behavioral Intervention

The application of chatbots or conversational agents (CAs) in delivering psychotherapy and health-interventions has gained significant attention as a means of broadening access to psychotherapeutic and mental health support services like health behavioral intervention [3, 41, 85] and mental healthcare [70, 99]. Rule-based chatbots [58], which operate on predefined dialogue scripts, are widely used in digital psychotherapy for their high controllability and explainability, essential qualities in sensitive fields such as mental healthcare and psychotherapy. These systems ensure precision in addressing therapeutic needs and prevent deviations from clinically validated pathways. For example, studies have used rule-based chatbots for Motivational Interviewing (MI) [56, 59] and Cognitive Behavioral Therapy (CBT) [9] to support interventions like smoking cessation [3, 41] and physical activity promotion [55, 85], where controlled, structured responses are crucial.

Despite the benefits of controllability and transparency, rule-based chatbots face limitations in dialogue empathy and flexibility. Their interactions tend to be rigid, lacking the dynamic adaptability of human therapists, which is particularly essential in psychotherapy, where a nuanced understanding of user queries and empathetic responses is critical [100]. For instance, a core principle of MI or CBT is reflective listening [61], where the therapist mirrors the client's expressions to promote self-reflection and insight. However, rule-based chatbots struggle to emulate such responsive empathy [42], as their responses are confined to predefined scripts that may not fully capture the depth of user needs or emotional nuances. More importantly, rule-based chatbots typically rely on intent-based dialogue systems [58], where domain experts design expert-crafted scripts to define chatbot's response to various user intents, ensuring consistent therapeutic guidance and simulating therapeutic conversations with a high degree of accuracy and reliability. However, this dependence on expert-written scripts makes rule-based chatbots highly resource-intensive, as they require extensive input from experts to develop domain-specific dialogue scripts [20, 82, 93]. Implementing these expert-crafted dialogues often involves encoding intricate conversation designs, including conversational content and flows, that incorporate strategies from psychotherapy such as MI or CBT, making the development process time-consuming and costly. Therefore, our work created a dataset with extensive expert-crafted dialogue scripts tailored for physical activity interventions using MI and CBT.

2.2 Generative Language Model for Digital Psychotherapy and Behavioral Intervention

The integration of generative language models [64, 81] into digital psychotherapy shows promise in enhancing conversational flexibility and depth in chatbot-driven therapeutic interventions. Initial efforts to improve rigid, rule-based chatbots led to hybrid systems from prior work by [6] that combined Natural Language Generation (NLG) [25] with predefined dialogue scripts. These systems aimed to improve conversational fluidity by incorporating generative elements to create more empathetic, reflective responses, a core principle of many psychotherapeutic techniques such as MI [56, 59] and CBT [9]. For example, hybrid models generate reflective statements or paraphrases to encourage clients to explore their thoughts while staying aligned with therapeutic goals [6, 23, 86]. With advancements in large language models (LLMs) [14, 34], the potential for more adaptive and contextually aware dialogues in digital psychotherapy has grown considerably. LLMs offer a more powerful ability to generate varied and nuanced responses, which can help maintain engagement and improve the user experience during psychotherapy [7, 89]. In MI or CBT, where client engagement and reflective listening are critical [42], LLMs have shown promise by producing responses that align with therapeutic techniques [45, 89], with a more flexible and engaging conversational style [43, 71, 100]. A recent study [43] demonstrated the effectiveness of generative AI chatbots in providing clinical-level mental health treatments, supported by evidence from a long-term randomized controlled trial.

Despite these advantages, the use of LLMs in psychotherapy presents considerable challenges [21], especially concerning technical controllability, model transparency [26], cultural understanding [45] and ethical concerns [95] in sensitive contexts. Unlike rule-based systems that follow experts' predefined guidance or expertise, LLMs operate with a black-box nature that generates responses based on probabilistic models [10] that may lack explicit adherence to therapeutic principles or ethical standards [95]. This lack of predictability and controllability poses risks, as LLMs may inadvertently generate responses that are insensitive, ethically inappropriate, or even harmful, especially in sensitive interactions where mental health and emotional well-being are involved. Furthermore, LLMs struggle to consistently apply psychotherapeutic counseling techniques, such as MI or CBT, as they lack the ability to systematically follow therapeutic strategies without expertise guidance. These challenges have prompted further investigations into how LLMs and generative AI systems can be instructed or aligned with therapeutic principles [89, 98, 100], human values [46, 65, 68], external knowledge [22], and ethical standards [72, 91] to ensure safe and effective deployment in sensitive contexts such as psychotherapy for health intervention.

2.3 Aligning LLMs with Domain Expertise and Human Instructions for Psychotherapy

Recent advances in aligning LLMs has evolved beyond linguistic fluency to emphasize adherence to goal-directed and domain-informed dialogue. This shift from purely probabilistic generative models [10] to more adaptive systems has introduced approaches such as instructed dialogue generation [37], reinforcement learning from human feedback (RLHF) [65, 68], and context-aware generation [102]. These approaches enable LLMs to tailor response generation based on specific conversational strategies, user intent, and domain objectives [50, 89, 98]. This also enables more effective mixed-initiative dialogue [22, 92], where both the user and model collaboratively guide the conversational interaction.

In psychotherapy, LLMs offer transformative potential to deliver flexible and engaging interventions; however, prior work by Iftikhar et al. [45] indicates that instructed LLMs may overly rely on therapeutic techniques with compromising conversational qualities such as empathy. To address this, alignment approaches should guide LLMs to balance conversational quality with adherence to structured therapeutic principles, which makes the alignment of LLMs with domain

expertise critical to ensure safety, engaging, and goal-directed conversations. Aligned dialogue generation is expected to support models to embed psychological and empathetic principles within responses [43, 80, 83, 88], allowing LLMs to effectively emulate the precision of rule-based systems while retaining the flexibility needed for nuanced, context-sensitive support in psychotherapeutic settings.

Two dominant alignment approaches have emerged: fine-tuning and prompting. Fine-tuning [105] integrates domain expertise directly into model weights, offering strong domain fidelity [43, 45, 98]. However, it is resource-intensive and dependent on large, sensitive datasets that are difficult to obtain in psychotherapy. Prompting [52], especially with in-context learning methods like Chain-of-Thought [96, 97, 104] and Tree-of-Thoughts [101], provides a more scalable and flexible alternative, guiding LLM outputs with structured cues instead of continuous model retraining. Recent work [8, 89] has shown that incorporating therapeutic strategies into prompts allows LLMs to produce more therapeutic-aligned responses [78, 100]. Comparative studies [48] also validate both prompting and fine-tuning as effective approaches for mental health tasks, supporting their potential in psychotherapeutic interventions.

Despite this progress, these approaches typically assume access to comprehensive domain-specific data for aligning the LLMs, which can be costly and time-consuming for domain experts to produce. To address this, we proposed Script-Strategy Aligned Generation (SSAG) in this work, an alignment approach that requires only partial expert-crafted dialogue scripts, thereby reducing reliance on fully scripted dialogue content or flows, while using LLMs to dynamically manage dialogue flow and generate responses. SSAG supports a collaborative expert-LLM co-authoring workflows and enables a controllable alignment through the stepwise therapeutic strategy prediction mechanism. Compared to prompting and fine-tuning, SSAG offers a more flexible and efficient alternative for the real-world development of psychotherapy chatbots, especially in the contexts where data is limited and expert labor is costly.

3 Creating Dataset with Expert-Crafted Dialogue Scripts

A team of fifteen experts (i.e., research associates, each holding a BSc or higher degree in clinical or health psychology), collaboratively developed a comprehensive dataset of pre-scripted, tree-structured dialogues designed to support physical activity interventions. Pre-scripted dialogues include both dialogue content and flows in tree-structure as shown in Fig 2. Grounded in the psychotherapeutic techniques of MI [56, 59] and CBT [9], these scripts target common therapeutic scenarios in behavioral interventions. Examples of crafted dialogues are shown in Fig 4, covering topics such as "rating confidence of change" in MI and "supportive social environment" in CBT.

The creation process began with brainstorming sessions, where experts identified key psychotherapeutic topics relevant to behavioral interventions, focusing on physical activity. Topics like overcoming exercise barriers and improving sleep hygiene were chosen to cover diverse intervention scenarios. Each expert then crafted dialogues for specific topics, ensuring the dataset captured the depth and variability of real therapy while adhering to MI and CBT principles. To ensure clarity and authenticity, a think-aloud protocol [33] was used.

The final dataset includes over 1,800 expert-written utterances across 26 **psychotherapeutic topics** as detailed in Appendix A.1. The dataset includes a mix of closed- and open-ended questions, reflections, and general advice, key elements for eliciting client self-reflection and promoting engagement, which are core objectives of both MI and CBT. Each topic is structured by a series of **therapeutic questions**. The tree-structured format as shown in Fig 2 ensures a coherent flow and branching, designed to mirror real-world psychotherapeutic dynamics. In order to support reproducibility of our work, we aim to make the dataset available upon request, with the creation of a data sharing agreement following EU intellectual property management [18] guidelines and

Psychotherapy Technique	Psychotherapeutic Topic	Example of Expert-crafted Dialogues	Therapeutic Strategy
Motivational Interviewing (MI)	"Rating Confidence of Change"	**Therapist**: So my first question is: on a scale from 0 to 10, how confident are you that you will reach your step goals tomorrow 👟?	Therapeutic Question 1
		Client: I would rate as 7 maybe.	
		Therapist: Thanks for sharing with me😊 Seems like you're already quite confident about reaching your goal. Can you tell me why you gave it a 7 but not a 5? Any aspects make you confident right now to achieve your step count goal?	Reflection / Therapeutic Question 2
		Client: Because I have to walk to school every day, so I should be able to have enough steps.	
		Therapist: That's indeed something that could make you more confident! Do you know what would need to change so you will rate your confidence a 9?	Reflection / Therapeutic Question 3
		Client: Hmm, I don't know …	
		Therapist: I understand.. trying to raise your confidence is easier said than done. But I am here to help! I have a few small tips you can try to boost your confidence to reach your step count goal. Would you like to know?	Reflection / Therapeutic Question 4
		Client: Sure, why not.	
		Therapist: The best way to gain more confidence is to see yourself succeed, so the first tip is: split your goal into small pieces and try to achieve those! My second tip is: ……	General Advice
Cognitive Behavioral Therapy (CBT)	"Supportive Social Environment"	**Therapist**: I would like to talk to you about how your social environment can play a big part in making your physical active easier. Did you know that people around you can encourage you to be more active?	Therapeutic Question 1
		Client: Maybe there are someone …	
		Therapist: Great! So why not use it😊 Can you think someone who would like it if you walked more and did more physical activities?	Reflection / Therapeutic Question 2
		Client: I think, my mum would like that if I can walk more or do more physical activities.	
		Therapist: Nice! This person can be a great supporter in your journey to become more active 🧑‍🤝‍🧑 Can you think of two ways in which he or she can be helpful?	Reflection / Therapeutic Question 3
		Client: Hmm, can you maybe have an example?	
		Therapist: Of course! An example of how someone can help you be more active is by giving you compliments, encouragement, or reminding you of what motivated you to make decision to be more active.	General Advice

Fig. 4. Examples of expert-crafted dialogues in motivational interviewing (MI) and cognitive behavioral therapy (CBT) with different psychotherapeutic topics. The dialogues contain 1) psychotherapeutic topics, 2) therapeutic questions, 3) reflections, 4) general advice, and 5) dialogue flows pre-designed as in Fig 2.

the respective institute wherein this data belongs. More detailed dialogue examples are provided in Appendix A.2 and Appendix A.3

4 Study 1: Concept of Aligning LLM with Full Expert-Crafted Dialogue Scripts

4.1 Concept Validation: Script-Aligned Generation (SAG) by Fine-tuning and Prompting

Study 1 aimed to validate the concept of Script-Aligned Generation (SAG), which aligns LLMs with full expert-crafted dialogue scripts for psychotherapy. SAG seeks to balance therapeutic fidelity with the conversational flexibility and engagement inherent in LLMs. The goal of Study 1 was to assess SAG's potential to enhance LLMs' potential to deliver flexible, engaging, and expert-guided psychotherapy. To implement the concept of SAG, we proposed two alignment approaches: fine-tuning (LLM-SAG (FT)) [105] and prompting (LLM-SAG (Prompt)) [52], to align LLMs with expert-crafted, tree-structured dialogue scripts designed for health intervention as shown in Fig 2, aiming to determine which provides the best combination of therapeutic effectiveness and conversational qualities.

4.2 Study Methods

4.2.1 Study design and procedure. A comprehensive overview of the study procedure is illustrated in Fig 3 (a). We conducted an evaluation study employing a within-subjects design to compare four chatbot types: 1) a rule-based chatbot ("rule-based") strictly followed the expert-crafted dialogues; 2) a pure LLM ("pure LLM") without any alignment with expert-crafted dialogues; and LLM-powered chatbots strictly aligned with the full expert-crafted scripts through either 3) prompting ("LLM-SAG (Prompt)") or 4) fine-tuning ("LLM-SAG (FT)").

Before the study, each participant received an information letter and provided informed consent. After, participants interacted with each chatbot in counterbalanced orders and completed a survey

```
Searching Method: Top-down Breadth-First Search (BFS)
procedure BFS(G: connected simple graph with vertices v₁, v₂, . . . , vₙ)
T := tree consisting only of vertex v₁
Q := [v₁]  {Q is the queue of unprocessed vertices}
while Q is not empty do
begin
      v := first vertex in Q
      remove v from Q
      {now process vertex v}
      while v has a neighbour not in T do
      begin
            w := first neighbour of v not in T
            append w to the end of Q
            T := T + vw
      end
end
{T is the breadth-first spanning tree of G}
```
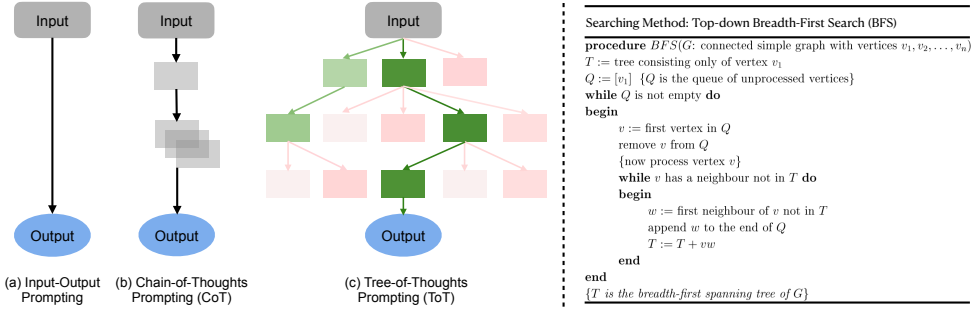
Fig. 5. Visualization of aligning LLMs with expert-crafted tree-structured dialogue scripts via Tree-of-Thoughts prompting [101].

immediately after interaction to assess their experience with that specific chatbot. This within-subjects design allowed for a direct comparison of user assessments across chatbot types under consistent conditions, aiming to identify which type of chatbot best performs in facilitating the therapeutic interactions.

4.2.2 Types of chatbots. Study 1 examined the following four types of chatbot:

1) **rule-based chatbot.** This chatbot was implemented using the RASA framework [11], a popular open-source platform for building rule-based chatbots. It strictly followed the expert-crafted dialogue scripts detailed in Section 3, using predefined intents, responses, and dialogue flow rules encoded in RASA's dialogue management model [11]. This chatbot did not generate responses beyond what is explicitly pre-scripted, ensuring high controllability. This condition served as a baseline, representing a traditional chatbot approach where all conversational paths are predetermined by experts.

2) **pure LLM.** This chatbot used a GPT-4o model [29] without alignment to expert-crafted dialogue scripts. At the start of each session, the LLM was prompted with: "You are a psychotherapist conducting a session to promote healthier behavior using Motivational Interviewing (or Cognitive Behavioral Therapy). The current therapeutic topic is [given topic]. The definition of this topic is [description of the given topic]." The topic was selected from a set of expert-defined **psychotherapeutic topics** (as demonstrated in Section 3 and detailed in Appendix A.1). Participants can interact with this chatbot freely within the specified topic. This setup evaluated the pure LLM's performance in delivering psychotherapy with high-level instructions, but without script guidance or alignment.

3) **LLM-SAG (FT).** We fine-tuned a GPT-4o model using our expert-crafted dialogue dataset, which consisted of tree-structured dialogues authored and reviewed by health psychology experts for clarity and coherence (see Section 3). We employed supervised fine-tuning involving input-output pairs, ensuring the LLM's responses aligned with expert-crafted dialogue responses and flows. Additionally, to evaluate the chatbot's adherence to the predefined dialogue flow, participants can ask "out-of-script" questions to test the chatbot's ability to respond appropriately while maintaining alignment with the specific dialogue flow. This method was also applied to the LLM-SAG (Prompt) chatbot.

4) **LLM-SAG (Prompt).** For a more scalable and computationally efficient alternative to fine-tuning [67], we implemented a prompting-based alignment using the Tree-of-Thoughts (ToT) technique [101], as visualized in Fig 5. ToT guided the GPT-4o model through our pre-authored, tree-structured dialogue scripts using a Breadth-First Search (BFS) algorithm [49]. We chose ToT for its effectiveness to align LLM with tree-structured expert-crafted dialogue scripts. Unlike other

prompting techniques (e.g., few-shot or chain-of-thought [97, 104]), ToT ensures consistency in multi-turn dialogue scenarios by enabling controlled, step-by-step navigation through predefined tree-based dialogue branches. At each turn, the LLM was prompted with the current dialogue context and a constrained set of valid next-turn response options through predefined dialogue branches in the tree structure, preserving the dialogue logic and therapeutic progression embedded in expert scripts. Compared to fine-tuning, ToT is a more lightweight and flexible approach that avoids retraining while maintaining integrity to expert scripts. While ToT was well-suited for aligning with tree-structured scripts in this work, we did not compare it with alternative prompting strategies (e.g., chain-of-thought). Future work could evaluate how different prompting strategies perform in this context.

The prompting templates used for the LLM-powered chatbots are provided in Appendix B. The implementation details of the models and settings are presented in Appendix C.

4.2.3 Web-based interfaces for chatbot interaction and evaluation. The study was conducted using a self-developed web-based application that integrated both chatbot interaction and an evaluation survey into one seamless interface, as shown in Fig 6. This design allowed participants to interact with the chatbot and provided immediate feedback in a continuous session, streamlining user experience. The interface was optimized for simplicity, minimizing distractions and ensuring smooth navigation, enabling participants to focus on their interactions without difficulties.
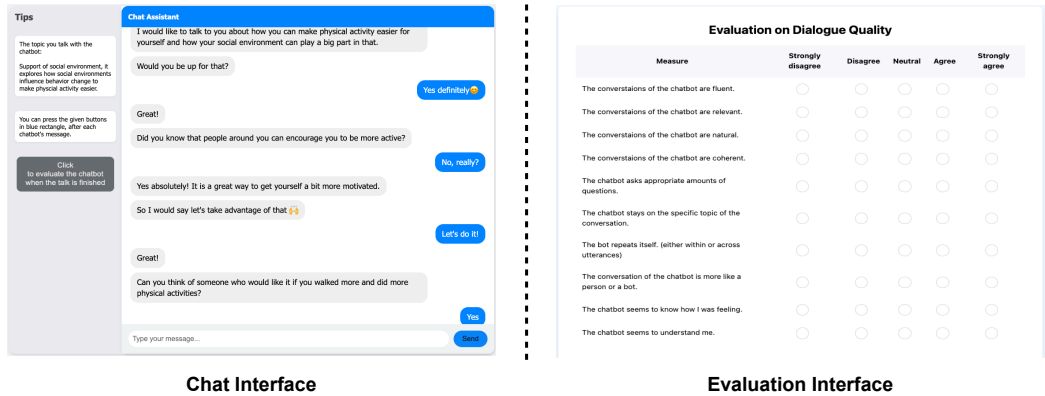

Chat Interface | **Evaluation Interface**

Fig. 6. Web-based interfaces were developed for Study 1 to enable chatbot interactions and facilitate immediate evaluation.

4.2.4 Measures. We assessed above four chatbot types using the measures as follows. All quantitative self-report measures used a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

Linguistic Quality: Evaluated the fluency, relevance, naturalness, coherence, and the human-likeness of the chatbot's responses, using five assessing items adopted from [16, 36]. Example items are "The conversations of the chatbot are fluent." and "The conversation of the chatbot is more like a person rather than a bot."

Dialogue Relevance: Inspired by work [13], two self-constructed items were used to specifically assess the alignment of expert-crafted therapeutic dialogues. Example items are "The chatbot asks an appropriate amount of questions." and "The chatbot stays on the specific topic of the conversation."

Empathy and Engagement: Gauged perceived empathy of the chatbot's conversations, which is crucial for therapeutic effectiveness. We adopted the questionnaire from [42, 76] with two items:

"The chatbot seems to know how I felt." and "The chatbot seems to understand me." To assess the chatbot's ability to maintain user interest and interaction as the engagement level, a questionnaire from [69] was adopted. Example items are "I lost myself in the interaction with this chatbot." and "The chatbot is enjoyable to talk to."

Perceived Motivational Interviewing (MI) Adherence: Evaluated how well the chatbot simulates an MI session and adheres to the MI principles. We adopted a questionnaire from [42, 54] with four items. Example items are "The chatbot helped me talk about changing my behavior." and "The chatbot helped me feel hopeful about changing my behavior."

Motivation for Change: Measured the chatbot's impact on participants' motivation to change, a key therapeutic outcome. We self-constructed an item "I am motivated to make changes in the behavior after I talk with the chatbot."

Therapeutic Alliance: Explored the rapport and connection between the chatbot and the user. We adopted the questionnaire from [44] with six items. Example items are "I believe this chatbot can help me to address my problem." and "This chatbot encourages me to accomplish tasks and make progress for the change we discussed."

Usability: Measured the ease of interaction with the chatbots. We adopted the questionnaire called Bot Usability Scale from [12] with seven items. Example items are "Communicating with the chatbot was clear." and "The chatbot's responses were easy to understand."

Two Open-Ended Questions were included to capture participants' feelings and opinions about chatbots: "What have you enjoyed most or least about interacting with this chatbot?" and "What could be improved about this chatbot?"

Automatic Evaluation Metrics: We also developed two self-defined, tailored metrics to evaluate the chatbots' performance in delivering expert-crafted *psychotherapeutic topics* and *therapeutic questions* (shown in Fig 2 and Fig 4). *"Auto-Metric 1"* assesses the number of psychotherapeutic topics that reached a natural conclusion, reflecting the level of user engagement with chatbots. *"Auto-Metric 2"* measures the total number of expert-crafted therapeutic questions asked by the chatbot, indicating its effectiveness in guiding the conversation toward specified therapeutic goals.

4.2.5 Participants. A power analysis using G*Power [30] indicated that at least 30 participants were required to detect a medium effect size ($d = 0.25$) and $\alpha = 0.05$ with 90% power. To ensure robustness, we recruited 43 participants (N=43) through institutional channels and social media. Participants had a diverse demographic profile, including variations in age, gender, and educational background. Eligibility criteria required participants to be at least 18 years old and fluent in English. Participation was voluntary, and participants were compensated with money or study credits for completing a 30-minute online session. This study was approved by the institutional ethics committee. Participant demographics are shown in Table 1.

4.2.6 Data analysis. To analyze the impact of different chatbots on participants' perceptions of psychotherapeutic interventions, we first tested the data's suitability for statistical analysis. Normality was assessed using the Shapiro-Wilk test [79], and homogeneity of variance via Bartlett's test [5]. As the data violated normality assumptions, we applied a Generalized Estimating Equation (GEE) model [38] to compare mean scores across chatbot conditions. Post-hoc pairwise comparisons between the LLM-SAG (Prompt) and LLM-SAG (FT) chatbots were conducted using Wilcoxon signed-rank tests [74] with Bonferroni correction [40].

For the qualitative data, we conducted an inductive content analysis [27] of responses to two open-ended questions. The first two authors developed an initial codebook using ATLAS.ti [2] based on participants' perceptions and suggested improvements for each chatbot type. Both coders independently coded the responses, revising the codebook as new themes emerged. Codes were refined, merged where appropriate, and re-coded to ensure consistency.

Demographic	Categories	Numbers of Participants (%)
Gender	Female	22 (51.2%)
	Male	21 (48.8%)
Age	18-24	23 (53.5%)
	25-34	14 (32.6%)
	35-44	2 (4.7%)
	45-54	3 (7.0%)
	65+	1 (2.3%)
Education	High school degree or equivalent	8 (18.6%)
	Bachelor's degree	15 (34.9%)
	Master's degree	16 (37.2%)
	Doctorate or higher	4 (9.3%)
Professional Domain	Health and Medical Science	7 (16.3%)
	Science, Technology, Engineering, Mathematics (STEM)	14 (32.6%)
	Business, Economics, and Law	6 (14.0%)
	Communication, Arts, Culture and Entertainment	5 (11.6%)
	Education and Social Science	8 (18.6%)
	Government and Public Sector	2 (4.7%)
	Other	1 (2.3%)

Table 1. Characteristics of participants in Study 1.

4.3 Quantitative Findings

4.3.1 Automatic evaluation metrics. Table 2 shows the results of the automatic evaluation metrics, as demonstrated in Section 4.2.4. For "Auto-Metric 1" (i.e., measuring completion of each psychotherapeutic topic), the rule-based chatbot scores the highest, followed by LLM-SAG (Prompt) and pure LLM, which have similar scores. In "Auto-Metric 2" (i.e., measuring therapeutic questions asked), the rule-based chatbot again leads, with LLM-SAG (Prompt) close behind. The pure LLM scores the lowest, indicating a significant deviation from the expert-crafted dialogue scripts.

Chatbot Types	Auto-Metric 1	Auto-Metric 2
Rule-based (as oracle)	100.00	98.81
Pure LLM	85.71	12.62
LLM-SAG (FT)	71.43	70.24
LLM-SAG (Prompt)	97.62	96.42

Table 2. Results (ratio) of automatic metrics for evaluating rule-based, pure LLM, and LLM-SAG chatbots.

4.3.2 Descriptive statistics. Participants reported moderate prior exposure to chatbots. The average frequency of chatbot use was 3.14 (SD=0.83), prior experience was 3.23 (SD=1.08), familiarity with chatbots was 3.44 (SD=0.73), and overall attitude toward chatbots was 3.40 (SD=0.79).

Across the four chatbot types: rule-based, pure LLM, LLM-SAG (FT), and LLM-SAG (Prompt), the LLM-SAG (Prompt) consistently outperformed others across multiple measures. As shown in Table 3, LLM-SAG (Prompt) achieved the highest ratings in linguistic quality (M=3.95, SD=0.23), dialogue relevance (M=3.39, SD=0.68), and empathy (M=3.07, SD=0.16), and showed strong performance in engagement (M=2.90, SD=0.33), perceived MI adherence (M=3.74, SD=0.13), motivation (M=3.42, SD=0.00), usability (M=3.82, SD=0.26), and therapeutic alliance (M=3.58, SD=0.46), surpassing other chatbot types. This indicated that it not only followed expert-crafted dialogue scripts closely but also delivered a more flexible and effective therapeutic experience overall.

4.3.3 Rule-based vs. LLM-powered chatbots. To address RQ1.1, we compared the performance of rule-based and LLM-powered chatbots, including pure LLM, LLM-SAG (FT) and LLM-SAG

Comparison	Measure	Mean (SD)	Coefficient	Effect ($Std.B$)	p-value
Rule-based vs. Pure LLM					
	Linguistic Quality	3.66 (0.22) vs. 3.76 (0.14)	0.10	0.14 (small)	0.45
	Dialogue Relevance	3.26 (0.88) vs. 3.09 (0.33)	-0.13	0.22 (medium)	0.34
	Empathy	2.53 (0.10) vs. 2.99 (0.18)	0.45	0.44 (medium)	0.04 *
	Engagement	2.58 (0.53) vs. 2.72 (0.42)	0.14	0.18 (small)	0.33
	Perceived MI	3.13 (0.15) vs. 3.13 (0.08)	0.00	0.00 (small)	1.00
	Motivation Change	2.86 (0.00) vs. 3.00 (0.00)	0.14	0.13 (small)	0.47
	Therapeutic Alliance	3.31 (0.64) vs. 3.33 (0.36)	0.02	0.02 (small)	0.92
	Usability	3.60 (0.40) vs. 3.56 (0.24)	-0.04	0.05 (small)	0.82
Rule-based vs. LLM-SAG (FT)					
	Linguistic Quality	3.66 (0.22) vs. 3.60 (0.18)	-0.06	0.08 (small)	0.65
	Dialogue Relevance	3.26 (0.88) vs. 3.38 (0.51)	0.15	0.25 (medium)	0.21
	Empathy	2.53 (0.10) vs. 2.94 (0.15)	0.41	0.40 (medium)	0.04 *
	Engagement	2.58 (0.53) vs. 2.58 (0.54)	0.00	0.00 (small)	1.00
	Perceived MI	3.13 (0.15) vs. 3.31 (0.12)	0.19	0.17 (small)	0.41
	Motivation Change	2.86 (0.00) vs. 3.00 (0.00)	0.14	0.13 (small)	0.46
	Therapeutic Alliance	3.31 (0.64) vs. 3.33 (0.48)	0.01	0.01 (small)	0.94
	Usability	3.60 (0.40) vs. 3.61 (0.26)	0.01	0.01 (small)	0.97
Rule-based vs. LLM-SAG (Prompt)					
	Linguistic Quality	3.66 (0.22) vs. 3.95 (0.23)	0.30	0.40 (medium)	0.02 *
	Dialogue Relevance	3.26 (0.88) vs. 3.39 (0.68)	0.16	0.27 (medium)	0.20
	Empathy	2.53 (0.10) vs. 3.07 (0.16)	0.53	0.52 (large)	0.01 **
	Engagement	2.58 (0.53) vs. 2.90 (0.33)	0.32	0.40 (medium)	0.03 *
	Perceived MI	3.13 (0.15) vs. 3.74 (0.13)	0.62	0.57 (large)	0.01 **
	Motivation Change	2.86 (0.00) vs. 3.42 (0.00)	0.56	0.51 (large)	0.01 **
	Therapeutic Alliance	3.31 (0.64) vs. 3.58 (0.46)	0.26	0.32 (medium)	0.10
	Usability	3.60 (0.40) vs. 3.82 (0.26)	0.22	0.30 (medium)	0.19

Table 3. Results of generalized estimating equations (GEE) [38] comparing rule-based with three types of LLM-powered chatbots (pure LLM, LLM-SAG via prompting and fine-tuning). (**p<.01, *p<.05, "Coefficient" represents the unstandardized regression coefficient.)
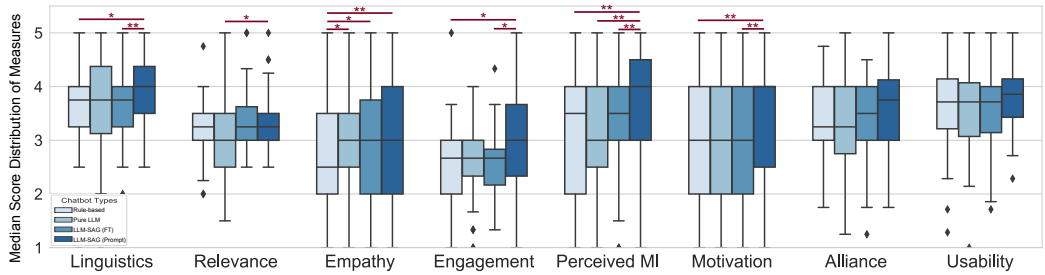

Fig. 7. Distribution of median value of measures in Study 1 across four chatbot types. Y-axis ranges from 1 to 5 (Likert scale). Each box shows the interquartile range, with the midline indicating the median. Dots outside represent outliers. Horizontal red lines with asterisks indicate statistically significant differences between conditions (*p < .05, **p < .01).

(Prompt). As depicted in Table 3 and Fig 7, LLM-SAG (Prompt) significantly outperformed the rule-based chatbot in linguistic quality (p=0.02), empathy (p=0.01), engagement (p=0.03), perceived MI adherence (p=0.01), and motivation change (p=0.01). Pure LLM and LLM-SAG (FT) chatbots only showed significantly higher empathy ratings than the rule-based chatbot (p=0.04), but performed similarly or even worse in dialogue relevance and MI adherence. LLM-SAG (Prompt) chatbot scored significantly higher than pure LLM in dialogue relevance and perceived MI, shown in Fig 7.

Moreover, all chatbots performed similarly in usability and therapeutic alliance. These findings affirmed the value of aligning LLMs with expert-crafted scripts to preserve therapeutic quality and conversational engagement.

4.3.4   Prompting (LLM-SAG (Prompt)) vs. Fine-tuning (LLM-SAG (FT)). To address RQ1.2, we compared two alignment approaches used to implement the concept of SAG: fine-tuning (LLM-SAG (FT)) and prompting (LLM-SAG (Prompt)). Results in Table 4 and Fig 8 showed that LLM-SAG (Prompt) significantly outperformed LLM-SAG (FT) in assessing diemensions, including linguistic quality (p=0.01), engagement (p=0.02), perceived MI (p=0.01), and motivation change (p=0.02). These findings suggested that prompting provides a more effective and flexible alignment approach for psychotherapy, likely due to its ability to flexibly navigate tree-structured dialogues while preserving expert-scripted dialogues for effective therapeutic experience, whereas fine-tuning may struggle to balance the adherence to the pre-crafted complex dialogue scripts and the adaptability to the "out-of-scripts" inputs from the participants.

Comparison	Measure	Mean (SD)	Effect ($Std.B$)	p-value
LLM-SAG (Prompt) vs. LLM-SAG (FT)				
	Linguistic Quality	3.95 (0.23) vs. 3.60 (0.18)	0.43 (large)	0.01 ∗∗
	Dialogue Relevance	3.39 (0.68) vs. 3.38 (0.51)	0.09 (small)	0.54
	Empathy	3.07 (0.16) vs. 2.94 (0.15)	0.07 (small)	0.67
	Engagement	2.90 (0.33) vs. 2.58 (0.54)	0.36 (medium)	0.02 ∗
	Perceived MI	3.74 (0.13) vs. 3.31 (0.12)	0.40 (medium)	0.01 ∗∗
	Motivation Change	3.42 (0.00) vs. 3.00 (0.00)	0.35 (medium)	0.02 ∗
	Therapeutic Alliance	3.58 (0.46) vs. 3.33 (0.48)	0.27 (medium)	0.08
	Usability	3.82 (0.26) vs. 3.61 (0.26)	0.24 (medium)	0.11

Table 4. Results from Wilcoxon signed-rank test [74] for pairwisely comparing the LLM-powered chatbots aligned with the full expert-crafted dialogue scripts (SAG) through prompting or fine-tuning. (∗∗*p*<.01, ∗*p*<.05)
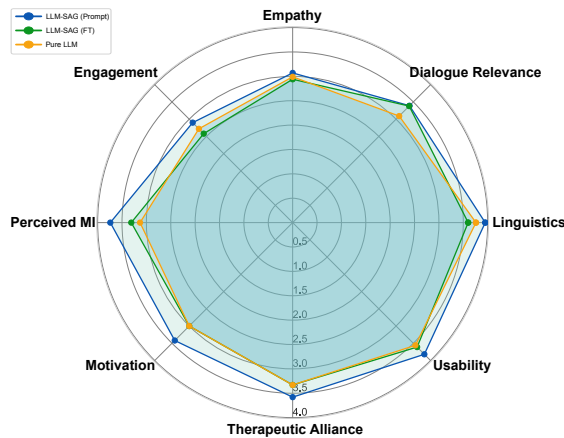

Fig. 8. Visualization of the comparison (mean value) between the pure LLM chatbot and LLM-SAG chatbots through prompting (LLM-SAG (Prompt) and fine-tuning (LLM-SAG (FT)). Each axis represents one of eight evaluation dimensions, with values plotted on a radial scale from 0 to 5 reflecting mean scores for each chatbot type.

4.4 Qualitative Findings: Open-ended Questions

A total of 344 free-text responses were collected in Study 1 from 43 participants across four chatbot types (i.e, rule-based vs. pure LLM vs. LLM-SAG (FT) vs LLM-SAG (Prompt)). 298 non-empty responses were included for analysis. Findings revealed four key themes shaping participant perceptions of chatbots in psychotherapy for health behavioral intervention that specifically addressing RQ1: (1) conversational quality and effectiveness, (2) balancing engagement with therapeutic adherence, (3) LLM-powered intelligences for digital psychotherapy, and (4) suggestions for improving chatbot-delivered psychotherapy.

*4.4.1 **Effective chatbot-delivered psychotherapy through clarity, relevance and coherence**. In therapeutic contexts, the effectiveness of chatbot interactions depends heavily on two intertwined factors: clarity and coherence in conversations, and the therapeutic relevance of the content. This theme highlights the need for chatbots to respond in a way that is easy to understand, logically structured, and directly aligned with users' therapeutic goals.

The qualitative analysis showed that participants responded favorably to the rule-based chatbots for their straightforward responses that were built on expert-crafted dialogue scripts. These chatbots were seen as effective in guiding users through therapeutic content with clarity. One participant remarked, "I liked the fact that the chatbot was really straightforward and clear in providing information, leading the users into questions. (rule-based chatbot)" However, participants also highlighted instances where clarity was lacking, especially in pure LLM-based chatbots. One participant shared, "It is not clear, and sometimes I don't know what I should say. Some examples are not clear at all. I wanted to see real instances, but it said ambiguously. (pure LLM)" This feedback highlights the need for clearer instructions, concrete examples, and more direct conversational guidance.

Therapeutic and contextual relevance were also viewed as essential. Some participants felt pure LLMs lacked focus, as one noted: "The topic is not quite relevant to the topic given; it would be good if it gave more specific advice on the topic we talked about for helping me in behavioral change." This highlights the importance of responses grounded in therapeutic objectives and user needs. While rule-based chatbots were appreciated for clarity, their rigidity was a frequent criticism. One participant noted, "I do not like chatbots with pre-set responses. There was no flexibility in what I could talk about." This inflexibility was seen as a barrier to deeper engagement. In contrast, LLM-based chatbots, particularly those aligned with expert-guided prompts (LLM-SAG), were praised for their adaptability and contextual responsiveness. As one participant described, "The chatbot can understand what I talk about and always stay within the topic given. It also asks relevant questions to move the talk forward. (LLM-SAG (Prompt))"

Maintaining logical flow and staying on topic was viewed as essential of effective interventions. Participants valued when chatbots responded coherently. One participant noted the improved coherence in LLM-SAG chatbots compared to rule-based one: "I can feel the dialogue from the (rule-based) chatbot is similar to the previous one, and this time it (LLM-SAG (Prompt)) can understand what I said and give correct reply. I can feel it quite encouraging for change." Another participant emphasized the capability of aligned LLMs to stay on topic: "It guides my behavior change step by step in human-like conversations. And it is not overwhelming. It can understand my specific questions and give correct answers, and then it can go back to previous talk after my questions, interesting. (LLM-SAG (Prompt))"

*4.4.2 **Balancing conversational engagement and therapeutic adherence: Flexible vs. Structured**. Conversational style plays a pivotal role in shaping users' perceptions and relationships

with chatbots. Participants consistently highlighted how conversational style impacts their sense of engagement and therapeutic effectiveness.

The rule-based chatbot provided well-structured dialogues for psychotherapeutic intervention following expert-crafted content and was noted as, "I love it guiding my behavior change step by step. Instead of directly asking me to do certain things, it guide me to think about how to deliver such changes." However, participants described LLM-powered chatbots as more relatable and engaging than rule-based ones. The LLM-powered chatbots' ability to mimic natural conversation made interactions feel natural and fluid. As one participant noted, "The tone of this chatbot is interesting and appealing. It talks like a human being. (LLM-SAG (FT))" In contrast, rule-based chatbots with their rigid, pre-set responses were disliked in this context as, "I do not like (rule-based) chatbots with pre-set responses. There was no flexibility in what I could talk about with the chatbot." This lack of conversational flexibility left some participants feeling disconnected and constrained.

Interestingly, while pure LLM chatbots have been able to generate contextually relevant responses, participants noted that those aligned with expert-crafted dialogues provided more structured, professional, supportive and motivational conversations. These chatbots asked thoughtful questions, provided relevant examples, and offered advice in a structured, encouraging tone. One participant shared, "I liked the way it asked questions, gave examples and advices. I found that they were very relatable. (LLM-SAG (Prompt))". This well-balanced flexibility and therapeutic structures for the engaging and therapeutically effective psychotherapy for behavioral interventions. However, some participants felt that overly professional or friendly tones could undermine the seriousness of certain topics. "I did not really like the super friend-like way of talking; I prefer a more professional type of way. (LLM-SAG (FT))" one participant noted. This points to the need for chatbots to dynamically adjust their conversational styles based on context, avoiding extremes and tailoring communication to user expectations and emotional needs.

In sum, the conversational styles of chatbots, particularly how they balance flexibility with structure, and friendliness with professionalism, directly influence user-perceived comfort, relatability, and engagement with the chatbots.

4.4.3 *LLM-Powered Intelligence for Engaging Psychotherapy: Personalization, Memory, and Empowerment*. This theme highlights the vital role of LLM-powered intelligence for enhancing digital psychotherapy for behavioral intervention, including features such as personalization, chatbot's memory, and user empowerment in creating engaging and effective chatbot-delivered interventions. The rule-based chatbot was disliked because of its pre-set responses and fixed dialogue structures without personalization, as noted, "I enjoy receiving the intended lesson and information in a very straightforward manner. [...] However, it did feel less personal and tailored. It was more like a good lesson than a conversation and coaching." Most participants appreciated chatbots that responded in a personalized but still structured way, offering tailored advice rather than generic responses. One participant remarked, "It guides my behavior change step by step in human-like conversations. And it is not overwhelming. It can understand my specific questions and give correct answers, and then it can go back to the previous talk after my questions, interesting. (LLM-SAG (Prompt))" This kind of personalized guidance, delivered at a comfortable pace, helped participants feel heard, supported, and empowered in their behavioral change journey.

In addition, participants also emphasized that chatbots capable of remembering prior interactions and adapting their responses accordingly made the conversations more coherent, supportive, and relatable. One participant shared, "The initial advice was also very generic and not useful for me specifically but in the rest of the conversation, the bot switched to suggestions, which were much more tailored to me and actually did inspire me. I even made a note and am going to implement one of the suggestions. (pure-LLM)" This illustrates how contextual awareness can strengthen

engagement and empower the effectiveness of psychotherapeutic interventions. This ability to build on past interactions can also enhance the continuity and personalization of long-term conversations.

Therefore, these LLM-powered intelligent features could help build more engaging, empowering, and effective chatbot-delivered interventions. By integrating these features and maintaining the right balance of conversational styles, chatbots can better support users' therapeutic goals and foster more engaging experiences.

*4.4.4   **User-Informed Suggestions for Enhancing Chatbot-Delivered Psychotherapy**.* Participants shared valuable suggestions to improve chatbot-delivered psychotherapy for health behavior change, emphasizing the need for more engaging, adaptable, and therapeutically effective interactions.

A key suggestion was to integrate LLM capabilities into rule-based systems. Participants believed this would combine the structured guidance of scripted dialogues with the conversational flexibility, pointing to the potential for more dynamic and personalized interactions. Alongside this, participants highlighted the value of open-ended questions, which LLMs are well-suited to generate. Closed-question formats were seen as limiting, while open dialogue can encourages reflection and deeper engagement in the therapeutic process.

Participants also stressed the importance of modulating the chatbot's conversational style. A balanced tone, friendly yet professional, was viewed as critical. One participant shared, "It could have used a slightly different tone to balance the formality that enhances its reliability and friendliness."

Personalization emerged as another vital area for improvement. Participants wanted chatbots to tailor responses to their specific circumstances and offer actionable advice. One participant noted, "It should have given clearer instructions for users for their change" highlighting the need for more individualized support. In line with this, participants emphasized the importance of memory and contextual awareness. The ability of a chatbot to recall past interactions was seen as essential for maintaining continuity, deepening personalization, and improving therapeutic relevance.

Collectively, these user insights point toward a future of chatbot-delivered therapy that is more personalized, flexible, engaging, and therapeutically effective, ultimately leading to more impactful psychotherapeutic interventions.

5   Study 2: A Collaborative and Scalable Approach to Develop LLM-Powered Psychotherapy Chatbots Aligned with Partial Expert-Crafted Dialogues

Study 1 underscored the value of expert-crafted dialogue scripts for aligning LLMs in psychotherapy. However, as shown in Fig 1 and Fig 2, the most labor-intensive aspects of the expert-scripting involve both creating dialogue content (e.g., questions, reflections, advice) and designing structured dialogue flows. SAG relies heavily on fully expert-authored scripts, which are costly and difficult to scale in expert-driven domains like psychotherapy.

To address this, we introduced Script-Strategy Aligned Generation (SSAG), a more flexible and efficient alignment approach that requires only partial expert input. SSAG drew on evidence-based therapeutic strategies [78, 100] adapted from Motivational Interviewing (MI) [61, 89] to guide LLMs in generating responses that balance therapeutic adherence with conversational flexibility. SSAG enabled LLMs to co-author therapeutic dialogue content and dynamically manage dialogue flow based on predicted therapeutic strategies, reducing reliance on experts' fully scripted dialogue. In Study 2, we compared SSAG (partial alignment) with LLM-SAG (full alignment) to assess whether SSAG could maintain therapeutic effectiveness while reducing dependence on expert-authored dialogue scripts (RQ2 and RQ3). By streamlining expert input demands and enabling flexible generation, SSAG facilitated more efficient and scalable development of psychotherapy chatbots supported by LLMs.

5.1 Approach: Script-Strategy Aligned Generation (SSAG)

Script-Strategy Aligned Generation (SSAG) combined partial expert-scripted dialogue content with therapeutic strategies to reduce reliance on fully expert-authored dialogue scripts. As shown in Fig 1 and Fig 2, expert-crafted dialogue scripts typically include five components: psychotherapeutic topics, therapeutic questions, reflections, and general advice, along with tree-structured dialogue flows. SSAG simplified this by requiring only the three most essential components: psychotherapeutic topics, therapeutic questions, and general advice (optional), as expert pre-authored input. Unlike prior alignment approaches such as unconstrained prompting/fine-tuning that rely heavily on full scripting and limit the therapeutic explainability, SSAG introduced a middle-ground alignment grounded in strategy prediction, offering greater modularity, efficiency, and explainability than direct prompting or fine-tuning alone. Specifically, SSAG operated in the following two steps as illustrated in Fig 1:

Step 1: Predicting the therapeutic strategy: SSAG first predicts the next therapeutic strategy based on dialogue context, selecting from core therapeutic behaviors: "asking questions," "reflective listening," and "giving advice."

Step 2: Generating the response: Once the therapeutic strategies are selected, LLM generates a response according to the predicted strategies. For example, if "asking questions" or "giving advice" is selected, the LLM retrieves and paraphrases the pre-written expert content (therapeutic questions or advice) based on dialogue context. If "reflective listening" is selected, the LLM generates reflections freely based on the context, allowing for dynamic, empathetic responses without scripting. By anchoring responses to specific therapeutic strategies, SSAG aligns the conversation with therapeutic principles [78, 100], maintaining control over the dialogue flow while enabling flexibility in generation.

5.2 Study Methods

5.2.1 Study design and procedure. The procedure of Study 2 is illustrated in Fig 3 (b). We conducted a 10-day field study employing a mixed design to compare three chatbot types: (1) a rule-based chatbot (baseline), (2) an LLM-powered chatbot aligned via SAG using prompting (LLM-SAG (Prompt)), and (3) an LLM-powered chatbot aligned via the flexible SSAG approach (LLM-SSAG). Participants were divided into two groups: one interacted with the rule-based chatbot and LLM-SAG (Prompt); the other with the rule-based chatbot and LLM-SSAG. Each group followed a counterbalanced design, where participants used one chatbot for five consecutive days, then switched to the other, totaling ten days. After each daily interaction, participants completed evaluation to assess their experience with the chatbot. This design enabled direct comparisons under consistent conditions to evaluate whether SSAG (partial alignment) could deliver psychotherapeutic interventions as effectively as SAG (full alignment).

5.2.2 Types of chatbots. Study 2 examined the following three types of chatbot:
1) **rule-based chatbot.** This chatbot was implemented using the RASA framework [11], as same as the rule-based chatbot used in Study 1 (see Section 4.2.2 (1)).
2) **LLM-SAG (Prompt).** This chatbot was implemented by aligning with the full expert-crafted dialogue scripts using SAG via prompting as in Study 1 (see Section 4.2.2 (4))
3) **LLM-SSAG.** As detailed in Section 5.1, SSAG offered a flexible alternative to SAG by reducing the need for fully scripted dialogues. It relied on three core expert-crafted elements: psychotherapeutic topics, a sequence of therapeutic questions, and general advice (optional). LLM-SSAG chatbot operated in two steps (see Fig 1). In Step 1, LLM-1 predicted the therapeutic strategies (i.e., asking questions, reflective listening, or giving advice) based on dialogue context. In Step 2, LLM-2 generated the dialogue response, ensuring alignment with the predicted strategies from Step 1.

In this study, SSAG restricted completely "free" generation to reflective responses, while therapeutic questions and advice were retrieved and paraphrased from expert-crafted dialogues. This balanced flexibility and safety: reflections are lower-risk and promote empathy and engagement, while questions and advice, being more directive and sensitive, are anchored in validated expert input to maintain therapeutic integrity and ethical standards. However, SSAG was designed to be extensible, similar to Retrieval-Augmented Generation (RAG) [35], but with greater controllability over which response types can be freely generated and which must remain grounded in expert data. Although this work restricted question and advice generation anchoring to expert-authored content for safety, future versions could allow more personalized generations with additional validation mechanisms to ensure safety and adherence to ethical standards. This setting also enhanced efficiency and adaptability. Training LLMs to freely generate expert-quality therapeutic questions and advice would require extensive domain-specific data and frequent model retraining. Findings from Study 1 have evidenced that prompting provides a more efficient alignment for expertise integration than fine-tuning. By contrast, SSAG maintained alignment with expert guidance while enabling flexible generation and dynamic dialogue flow management, reducing the need for continuous model retraining and making it more scalable.

The prompt templates used for the LLM-powered chatbots are provided in Appendix B.2 and Appendix B.3

5.2.3 Self-developed application. We developed a mobile app for Study 2 (see Fig 9) to support chatbot interaction and physical activity tracking. The app included: (1) a Menu Interface for navigation; (2) a Chat Interface for psychotherapy sessions, and (3) an Activity Tracking Interface to monitor daily steps and walking distance. This integrated design of the app enabled a seamlessly interactive experience for chatbot interventions and evaluations.
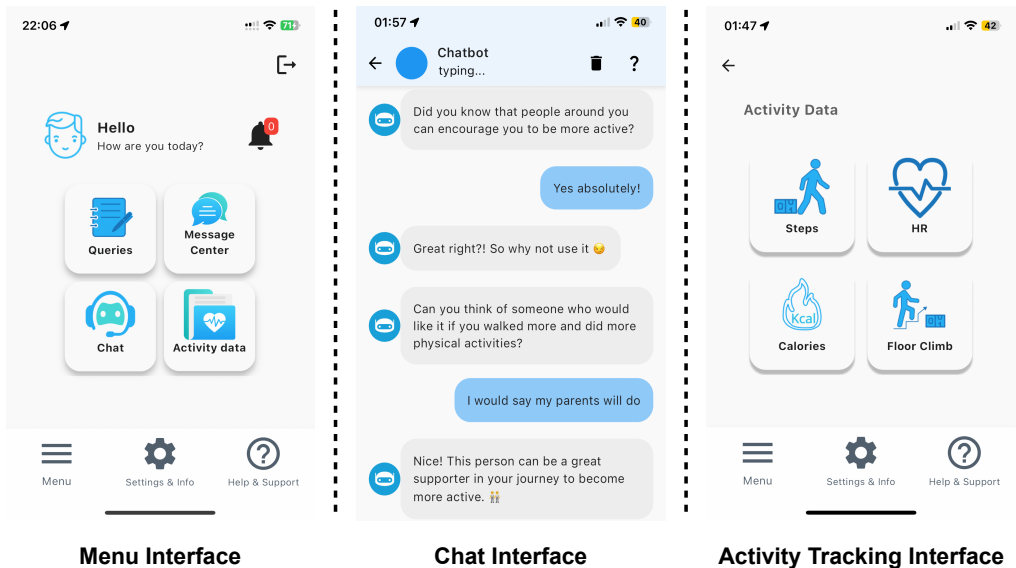

Menu Interface   **Chat Interface**   **Activity Tracking Interface**

Fig. 9. A mobile application was developed for a field study to enable long-term chatbot interactions and the tracking of PA levels.

5.2.4 Measures. To ensure consistency with Study 1, we used the same evaluation metrics, including linguistic quality, engagement, perceived empathy and MI level, therapeutic alliance,

and usability. These were assessed daily after each chatbot interaction. Intrinsic motivation to change was measured every two days (pre-, Day 2, Day 4, and post-intervention) using a validated scale adapted from [57]. Physical activity (i.e., daily steps and distance) was continuously tracked throughout the study to evaluate the effectiveness of physical activity promotion.

5.2.5 Participants. Based on a G*Power analysis [30], a minimum of 16 participants was required to detect a medium effect size ($d = 0.25$) with 80% power and $\alpha = 0.05$. To ensure robustness, 21 participants were recruited (N = 21). All participants were 18 or older, fluent in English, and recruited through institutional channels and social media. They received monetary compensation aligned with the guidelines for completing the participation over 10 days. The study was approved by the institutional ethics committee. Participant demographics are shown in Table 5.

Demographic	Categories	Numbers of Participants (%)
Gender	Female	13 (61.9%)
	Male	8 (38.1%)
Age	18-24	9 (42.9%)
	25-34	10 (47.6%)
	35-44	1 (4.8%)
	45-54	1 (4.8%)
Education	High school degree or equivalent	2 (9.5%)
	Bachelor's degree	5 (23.8%)
	Master's degree	10 (47.6%)
	Doctorate or higher	4 (19.0%)
Professional Domain	Health and Medical Science	3 (14.3%)
	Science, Technology, Engineering, Mathematics (STEM)	5 (23.8%)
	Business, Economics, and Law	3 (14.3%)
	Communication, Arts, Culture and Entertainment	4 (19.0%)
	Education and Social Science	3 (14.3%)
	Government and Public Sector	2 (9.5%)
	Other	1 (4.8%)

Table 5. Characteristics of participants in Study 2.

5.2.6 Data analysis. We first checked data suitability using the Shapiro-Wilk test [79] for normality and Bartlett's test [5] for homogeneity of variance. As data were not fully normally distributed, we applied Generalized Estimating Equations (GEE) [38], a statistical method robust to non-normality, to compare participants' perception and therapeutic effectiveness across rule-based, LLM-SAG (Prompt), and LLM-SSAG chatbot conditions.

5.3 Findings

Dataset	AnnoMI		BiMISC	
Metric	Accuracy	F1	Accuracy	F1
Flan-T5 [17]	46.2	77.6	19.1	17.4
Vicuna-13B [94]	44.7	76.2	10.5	18.8
GPT-4	**50.0**	**78.9**	**33.6**	**27.9**
GPT-4o (FT)	**63.6**	**81.4**	**47.2**	**36.9**

Table 6. Therapeutical strategy prediction on dataset AnnoMI (single-strategy prediction) and BiMISC (multiple-strategy prediction).

5.3.1 Pre-validation of therapeutic strategy prediction in SSAG. To support the feasibility of SSAG, we first validated LLMs' ability to predict therapeutic strategies (Step 1 of SSAG as defined in Section 5.1), which serve as the basis for response generation in Step 2 of SSAG. Building on prior work [15, 89], we benchmarked LLMs using two open-sourced MI datasets: AnnoMI [106] and BiMISC [87] for evaluating this prediction task.

As shown in Table 6, GPT-4 (zero-shot setting) achieved strong performance, while fine-tuning on data further improved accuracy (GPT-4o (FT)). Notably, performance was lower on the BiMISC dataset due to its multi-strategy complexity (i.e., one utterance has multiple MI strategies), compared to the single-strategy in the AnnoMI dataset. Although the overall accuracy was modest, this is acceptable within SSAG, where multiple strategies may be contextually appropriate and no single strategy is universally optimal. These results confirmed that LLMs can effectively support SSAG's Step 1 for therapeutic strategy prediction and serve as a basis for guiding dialogue generation in Step 2 of SSAG.

Comparison	Measure	Mean (SD)	Coefficient	Effect ($Std.B$)	p-value
Rule-based vs. LLM-SAG (Prompt)					
	Linguistic Quality	3.26 (0.79) vs. 3.84 (0.63)	0.57	0.72 (large)	0.01 **
	Dialogue Relevance	3.75 (0.70) vs. 4.00 (0.42)	0.22	0.36 (medium)	0.18
	Empathy	2.81 (1.13) vs. 3.54 (0.84)	0.75	0.70 (large)	0.00 **
	Engagement	2.69 (0.89) vs. 3.18 (0.98)	0.48	0.54 (large)	0.07
	Perceived MI	3.38 (0.74) vs. 3.83 (0.52)	0.41	0.62 (large)	0.04 *
	Therapeutic Alliance	3.28 (0.65) vs. 3.87 (0.52)	0.49	0.62 (large)	0.02 *
	Usability	3.50 (0.60) vs. 4.04 (0.59)	0.51	0.71 (large)	0.03 *
Rule-based vs. LLM-SSAG					
	Linguistic Quality	3.35 (0.83) vs. 3.83 (0.73)	0.50	0.63 (large)	0.06 −
	Dialogue Relevance	3.89 (0.74) vs. 3.95 (0.23)	0.23	0.37 (medium)	0.21
	Empathy	2.67 (0.95) vs. 3.70 (1.00)	1.00	0.94 (large)	0.01 **
	Engagement	2.74 (0.69) vs. 3.23 (0.93)	0.50	0.56 (large)	0.10
	Perceived MI	3.59 (0.69) vs. 3.90 (0.56)	0.35	0.53 (medium)	0.09
	Therapeutic Alliance	3.56 (0.49) vs. 3.83 (0.63)	0.37	0.53 (large)	0.10
	Usability	3.56 (0.85) vs. 4.09 (0.72)	0.56	0.78 (large)	0.04 *
LLM-SAG (Prompt) vs. LLM-SSAG					
	Linguistic Quality	3.84 (0.63) vs. 3.83 (0.73)	0.01	0.01 (small)	0.97
	Dialogue Relevance	4.00 (0.42) vs. 4.09 (0.55)	0.09	0.15 (small)	0.58
	Empathy	3.54 (0.84) vs. 3.70 (1.00)	0.17	0.16 (small)	0.63
	Engagement	3.18 (0.98) vs. 3.23 (0.93)	0.05	0.06 (small)	0.89
	Perceived MI	3.83 (0.52) vs. 3.90 (0.56)	0.07	0.10 (small)	0.73
	Therapeutic Alliance	3.87 (0.52) vs. 3.83 (0.63)	0.04	0.10 (small)	0.73
	Usability	4.04 (0.59) vs. 4.09 (0.72)	0.05	0.07 (small)	0.86
Motivation Change Day-1 vs. Day-5					
	Rule-based	3.74 (0.70) vs. 4.07 (0.35)	0.11	0.20 (medium)	0.04 *
	LLM-SAG (Prompt)	3.74 (0.70) vs. 4.13 (0.33)	0.39	0.67 (large)	0.00 **
	LLM-SSAG	3.74 (3.70) vs. 4.14 (0.65)	0.44	0.75 (large)	0.00 **

Table 7. Results of generalized estimating equations (GEE) [38] comparing rule-based and LLM-powered chatbots via full alignment (SAG) and partial alignment (SSAG). (**p<.01, *p<.05, "Coefficient" represents the unstandardized regression coefficient.)

5.3.2 Comparing chatbot types: rule-based vs. LLM-SAG vs. LLM-SSAG. We evaluated three chatbot types: rule-based, LLM-SAG (Prompt), and LLM-SSAG in Study 2 using consistent measures from

Study 1. As shown in Table 7, LLM-powered chatbots aligned with partial expert-crafted scripts (LLM-SSAG) performed comparably to those aligned with full scripts (LLM-SAG) in both conversational quality and therapeutic effectiveness. Besides, both of LLM-powered chatbots significantly outperformed rule-based chatbots in delivering psychotherapy across various assessing dimensions.

To address RQ2, we compared LLM-SAG (Prompt) and LLM-SSAG to evaluate whether LLM-SSAG chatbot could achieve comparable therapeutic effectiveness and conversational quality to those aligned with full pre-scripted dialogues by SAG. Results from the GEE analysis (Table 7) showed no significant differences in performance between LLM-SSAG and LLM-SAG across key metrics, including linguistic quality (p=0.97), dialogue relevance (p=0.58), empathy (p=0.63), engagement (p=0.89), MI adherence (p=0.73), therapeutic alliance (p=0.73), and usability (p=0.86). The findings that no significant performance differences exist between LLM-SSAG and LLM-SAG across most metrics supported the effectiveness of the partial alignment approach in LLM-SSAG, maintaining similar levels of conversational quality and therapeutic adherence while offering a more flexible alternative to the fully aligned LLM-SAG chatbots for psychotherapy. Radar plots in Fig 10 explicitly visualized their comparable performances.
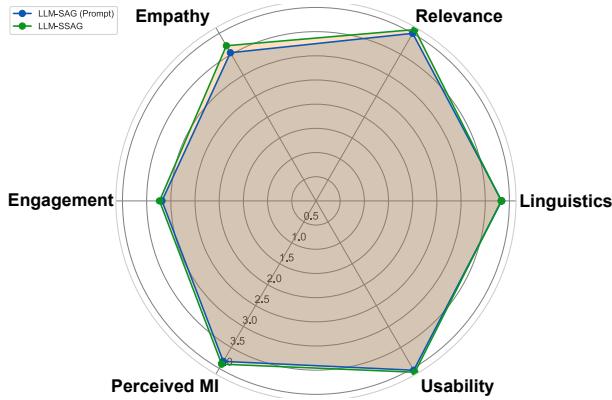

Fig. 10. Comparison (mean value) between the LLM-powered chatbots employing strict SAG through prompting (LLM-SAG (Prompt)) vs. our proposed flexible alignment (LLM-SSAG).

As shown in Table 7 and Fig 11 (a), linguistic quality (LLM-SSAG: M=3.83; LLM-SAG: M=3.84) and dialogue relevance (LLM-SSAG: M=4.09; LLM-SAG: M=4.00) of both LLM-powered chatbots were rated higher than the rule-based chatbot baseline, indicating better conversational quality. Empathy (LLM-SSAG: M=3.70; LLM-SAG: M=3.54), engagement (LLM-SSAG: M=3.23; LLM-SAG: M=3.18), MI adherence (LLM-SSAG: M=3.90; LLM-SAG: M=3.83), and usability (LLM-SSAG: M=4.09; LLM-SAG: M=4.04) were also higher than the rule-based baseline, with LLM-SSAG slightly outperforming LLM-SAG. Both LLM-powered chatbots scored higher in therapeutic alliance (LLM-SSAG: M=3.83; LLM-SAG: M=3.87) than rule-based chatbots, reflecting stronger therapeutic rapport.

In terms of the actual behavioral outcomes, as shown in Fig 11 (b), both LLM-SAG and LLM-SSAG chatbots showed sustained increases in motivation change and daily step counts throughout the intervention period, while the rule-based chatbot showed limited impact on motivation change over time. As for the daily step counts, while all conditions started with similar step counts on Day 1, the trajectories diverged significantly. LLM-SAG maintained a steady increase, peaking on Day 4. LLM-SSAG followed a similar upward trend, showing consistent increases and achieving the highest average steps by the end of the study. In contrast, the rule-based chatbot exhibited a downward

(a) Medians of measures in human evaluation across chatbot conditions.

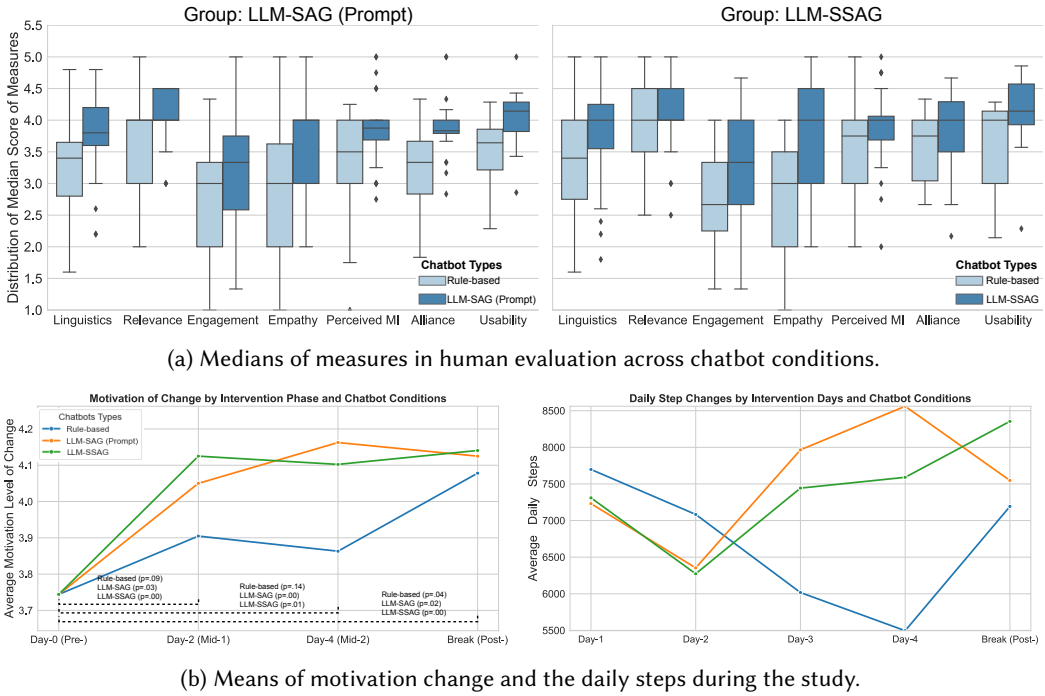(b) Means of motivation change and the daily steps during the study.

Fig. 11. Analysis results of chatbot evaluation, motivation change, and daily steps across intervention days and chatbot conditions.

trend, reaching its lowest point on Day 4 before a slight recovery. These results demonstrated that LLM-SSAG chatbot can offer comparable performance to fully aligned LLM-SAG chatbot.

5.3.3 Comparison of reliance on expert-crafted content: full alignment (SAG) vs. partial alignment (SSAG). To address RQ3, SSAG supported efficient co-development where experts pre-script only essential dialogue elements (i.e., psychotherapeutic topics, key therapeutic questions, health advice (optional)), and LLMs handle response generation or paraphrase as well as dynamic dialogue flow management based on therapeutic strategies and dialogue context, which reduces the need for fully expert-scripted content and structured flows, thereby could lower expert authoring workload.

To quantify this reduction, we compared the number of pre-authored utterances required for SSAG versus SAG. The LLM-SAG (Prompt) chatbot utilized the full expert-crafted dialogue scripts, consisting of 1,876 utterances and complete tree-structured flows (see Fig 2). In contrast, LLM-SSAG operated with only 501 therapeutic questions and 203 general advice utterances, covering only 37.5% of the total utterances in the dataset. A detailed breakdown by psychotherapeutic topic is included in Appendix A.1. These results showed that SSAG requires less than two-fifths of the expert-authored dialogue content needed for SAG. While this suggested a potential reduction in expert authoring workload, we did not directly measure time or workload. Thus, these findings should be interpreted as indirect evidence of efficiency, not a conclusive workload evaluation.

6 Discussion

6.1 Expert-Crafted Dialogues Remain Essential for Chatbot-Delivered Psychotherapy in the Era of LLMs

Chatbots for psychotherapy [41, 42, 85] have traditionally relied on rule-based systems structured around expert-crafted dialogue scripts [20, 82, 93], valued for their safety, controllability, and

strict adherence to evidence-based psychotherapeutic techniques like MI and CBT. While reliable, these systems often produce rigid, non-adaptive conversations [58], limiting their ability to tailor responses to nuanced client needs. The emergence of LLMs introduces new possibilities for more personalized, empathetic, flexible, and engaging dialogue [43, 71, 73]. However, when deployed without domain alignment, pure LLMs, such as ChatGPT [66], lack the domain-specific knowledge, structure, and safety controls needed for high-stakes contexts like psychotherapy. They struggle to generate sequential, goal-directed therapeutic questions and can deviate from evidence-based psychotherapeutic techniques, making experts in such domains hesitate in relying on them alone.

Study 1 compared four chatbot types: rule-based, pure LLM, and two LLMs aligned with expert-crafted dialogue scripts using fine-tuning or prompting (SAG). Results showed that LLM-aligned chatbots significantly outperformed both rule-based and pure LLMs across assessing metrics. These findings addressed **RQ 1** that *expert-crafted dialogue scripts remain essential in the era of LLMs for effective digital psychotherapy.* Specifically, findings from Study 1 revealed that participants rated LLM-powered chatbots higher than rule-based ones in linguistic quality, empathy, and engagement, highlighting the conversational fluency and personalization that generative models bring. While rule-based chatbots are reliable, their rigidity and reliance on pre-authored scripts limit their ability to adapt to diverse conversational contexts [93]. Participants noted a lack of contextual awareness and personalization, both essential for empathetic and engaging psychotherapy interactions. However, the flexibility of LLMs came with tradeoffs. Pure LLMs were perceived as more natural and engaging but often lacked therapeutic structure and relevance. Participants described them as "human-like" yet vague or unfocused with qualitative insights, resulting in lower scores for dialogue relevance and MI adherence compared to rule-based and aligned LLMs. In contrast, rule-based chatbots maintained therapeutic objectives and structure but were often seen as rigid and less engaging. Aligned LLMs (SAG) effectively bridged this gap by preserving the structure of expert scripts while introducing empathetic and adaptive dialogue generation. Participants appreciated this combination, describing aligned LLM-powered chatbots as "professional yet warm" and capable of delivering structured guidance with emotional resonance, an essential balance in psychotherapy [42, 45]. The qualitative insights reinforced the quantitative findings: aligned LLMs offered a more effective tradeoff between therapeutic rigor and conversational naturalness than either rule-based or unaligned pure LLMs.

Although LLMs are increasingly capable of generating empathetic and engaging dialogue [51, 53], such AI-generated empathy should be interpreted with caution because AI perceives warmth, empathy differently than humans [90]. For instance, LLMs may over-express empathy [75, 100], which can lead to discomfort or unintended emotional effects, and AI often fails to express empathy in positive circumstances while humans do [75]. A study [19] shows conversational agents adapt empathetic responses to certain identities, which can be potentially exploitative. While our findings show that participants perceived LLM-generated responses as more empathetic and engaging than rule-based ones, such empathy and engagement need further control. Prior work suggests that human-AI collaboration can enhance perceived empathy [80], supporting our approach of aligning LLMs with expert guidance to balance conversational attributes (e.g., empathy) with therapeutic structure. Given the relatively short duration of our study, future research should examine how human-perceived chatbot's empathy evolves over longer-term use.

In sum, expert-crafted dialogue scripts remain essential in the era of LLMs, not as rigid blueprints but as scaffolding that enables LLMs to deliver therapeutically-structured, safe, and engaging dialogues. SAG-aligned LLMs guided by the expert-crafted scripts outperformed both rule-based and pure LLMs by effectively managing the tradeoff, offering a practical path for engaging and therapeutically effective chatbot-delivered psychotherapeutic interventions.

6.2 Effective Alignment of LLMs with Domain Expertise for Psychotherapy

In psychotherapy, aligning LLMs with expert knowledge is essential to ensure conversations remain therapeutically relevant and effective. Prior work on instructing and aligning LLMs [43, 45, 89] has focused primarily on two alignment approaches: fine-tuning and prompting, each with distinct trade-offs. Fine-tuning [105] embeds domain knowledge directly into model parameters but is resource-intensive and requires large, curated datasets, often scarce in sensitive fields like psychotherapy. In contrast, prompting [52] offers a cost-effective and flexible alternative by guiding LLMs at inference time using structured inputs, avoiding retraining while encoding therapeutic goals through prompt design. Prompting has been shown to use less than one-tenth the computational cost of fine-tuning for achieving similar alignment [67], making it a more scalable solution for psychotherapy where domain-specific data is limited but adherence to therapeutic principles is critical. It also allows quick adaptation to new therapeutic techniques, supporting flexible deployment in evolving psychotherapeutic settings.

To explore RQ1.2, Study 1 compared these two alignment approaches for their effectiveness in delivering health interventions using the same expert-crafted dialogue scripts. The findings revealed that the prompted chatbot consistently outperformed the one by fine-tuning across assessing measures. Prompting preserved the structure of expert-authored dialogue flows while offering greater adaptability in response generation, better supporting context-aware, engaging, and goal-directed therapeutic conversations. This difference in our work partly arises from how each approach handles dialogue structure: fine-tuning embeds the tree structure into the model, reducing adaptability, while prompting uses tree-structured scripts [101] during inference, enabling more flexible, diverse conversational paths while preserving therapeutic coherence and staying on the scripted therapeutic path, as noted by our participants, "It (LLM-SAG (Prompt)) can understand my specific questions and give correct answers, then it can go back to previous talk after my questions, interesting." This tradeoff is particularly important in psychotherapy, as prior work [45, 88] indicates that adherence to therapeutic principles should not come at the expense of empathy and conversational engagement.

However, both alignment approaches (i.e., SAG through either prompting and fine-tuning) rely heavily on extensive expert input, making scalability and efficiency a challenge. We thereby proposed Script-Strategy Aligned Generation (SSAG), a more flexible alignment that reduces scripting overhead by leveraging only core components (i.e., therapeutic topics, key questions, and optional advice) while allowing LLMs to support the dialogue generation and manage dialogue flow dynamically through therapeutic strategy prediction, inspired by prior CSCW research that uses therapeutic strategies to enhance human counseling effectiveness [78, 100] as well as NLP studies [88, 89] that instruct LLMs to follow evidence-based therapeutic principles. Results from Study 2 showed that LLM-SSAG performed comparably to LLM-SAG (Prompt) across most evaluation metrics, addressing our RQ2. This demonstrated that SSAG can maintain therapeutic structure by anchoring each response to a predicted MI strategy (e.g., asking a question, offering advice, or providing a reflection), ensuring alignment with evidence-based therapeutic behaviors while allowing flexible generation. By guiding LLMs through strategy prediction, especially reflective listening, SSAG tried to mitigate the over-reliance on therapeutic "method" (e.g., CBT) seen in prior work [45] and supports more engaging, empathetic dialogues. This structured-yet-adaptive design mirrors techniques like Retrieval-Augmented Generation (RAG) [35], but with greater controllability for expert oversight.

The choice of MI and CBT balanced LLMs by combining MI's emphasis on empathy and engagement with CBT's structured, goal-oriented interventions. However, SSAG is not limited to these two focused frameworks. Its stepwise, strategy-level design allows integration of other therapeutic

techniques by substituting proper expert-defined strategies. Because SSAG operates at the level of therapeutic strategies rather than fixed language patterns, it is also adaptable to multilingual and culturally diverse contexts by accommodating different language norms and counseling styles.

6.3 Implication: LLM-Supported Co-Development of Psychotherapy Chatbots

Psychotherapy chatbots, as socially impactful tools, hold promise for expanding access to mental health support, particularly in underserved communities. However, their development has traditionally relied on labor-intensive expert scripting, which limits scalability. Our work addressed this challenge by first evaluating alignment approaches and then introducing Script-Strategy Aligned Generation (SSAG), a method designed to reduce dependence on fully expert-scripted dialogues and enable more efficient co-development of scalable and explainable LLM-powered psychotherapy chatbots between AI (i.e., LLM), developers, and domain experts.

From a technical perspective, Study 1 confirmed that prompting is a more scalable and resource-efficient alignment method than fine-tuning. Prompting preserved therapeutic adherence while allowing for flexible, human-like interaction, making it suitable for conversational therapies such as MI and CBT, where structured intervention must be delivered through adaptable dialogue. Building on this, SSAG introduced a two-stage strategy-level alignment method that relies on partial expert input (i.e., therapeutic topics, key questions, and general advice), while delegating dialogue generation and flow management to the LLM via predicted MI strategies. Results from Study 2 showed that SSAG matched the performance of fully aligned SAG chatbots across key therapeutic metrics while requiring less than 40% of expert-authored dialogues, offering strong evidence of reduced reliance on expert-crafted dialogue content, answering our RQ3. While we did not directly measure reductions in expert workload by SSAG, its design eliminates the need for domain experts to manually construct full dialogue trees or write structured prompts. This suggests potential for easing expert involvement during the chatbot development process. For instance, experts can focus on defining high-level strategies and therapeutical questions, while the LLM handles context-specific generation, reducing the need for technical prompt engineering. We acknowledge that this only partially demonstrates collaborative development. While SSAG facilitates division of labor between domain experts and AI, the degree of collaboration was not empirically studied. Future CSCW research should more directly evaluate expert-AI workflows using participatory design methods, task-based workload assessments, or longitudinal co-development trials to rigorously assess reductions in human workload of both expert and developer.

From a CSCW and human-AI collaboration perspective, SSAG advances the vision of cooperative development in expert-driven and sensitive domains. SSAG positions LLMs as flexible yet controllable collaborators that support experts in constructing adaptive, therapeutically grounded dialogues, operating within the bounds of expert-guided therapeutic strategies. This draws on prior CSCW work on human-AI co-creation frameworks [103] and domain-specific chatbot co-development [86], demonstrating how structured human-AI collaboration can scale innovation in mental health technologies. Although we evaluated it in a semi-controlled setting, deployment in unsupervised settings still introduces ethical and safety concerns. Real-world deployment requires additional safeguards, including human-in-the-loop oversight [63, 84], real-time validation mechanisms, and adherence to ethical standards. These measures are especially critical in unsupervised or high-risk settings like psychotherapy.

Taken together, SSAG aligns with CSCW's mission to foster computer-supported cooperative work in socially impactful domains. By reducing development barriers and enabling scalable, explainable, and expert-aligned AI, SSAG opens pathways for delivering cost-effective and accessible mental health support at scale, particularly in underserved or resource-limited fields. Its design facilitates a scalable and efficient development between experts and developers, illustrating how

AI can be harnessed in ways that are explainable, domain-aligned, and co-developed. Through this lens, this work serves as both a technical advancement and a CSCW-relevant contribution, representing a promising step toward building cooperative, expert-informed and socially impactful AI technologies that advance mental health support systems and the broader goals of CSCW.

7 Limitations and Future Work

Despite the contributions of this work, we acknowledge several limitations that remain for future exploration.

First, while the expert-crafted dialogue scripts used in this study are valuable, the reliance on such data introduces potential biases, which could impact the quality of LLM alignment. Its limited size and sole focus on MI and CBT, may not fully capture the diversity of real-world therapeutic interactions or generalize well to other more intensive forms of psychotherapy that require different conversational structures and techniques. Future work should expand datasets to cover a wider range of therapeutic scenarios enhancing SSAG's generalizability and robustness. Moreover, although SSAG demonstrated reduced reliance on full scripting, requiring fewer than 40% of expert-authored utterances compared to SAG, we did not directly measure expert workload or time savings. Thus, the efficiency claim should be interpreted as indirect. Future work should include task-based workload assessments or participatory design studies to evaluate the reduction of human workload more rigorously.

Second, although Study 2 included a 10-day field evaluation, the sample size was modest. Given the personalized nature of psychotherapy, user evaluations may vary widely and the reliance on self-reported feedback limits generalizability of our findings. Future work should involve more diverse samples and incorporate expert assessments for more robust evaluations. While the study duration was relatively short, prior HCI and digital mental health research has employed comparable intervention periods (e.g., 2 weeks) to assess the effectiveness of such tools or applications [28, 32, 47]. Nonetheless, human-perceived AI empathy and warmth may evolve over time. Longitudinal studies are needed to examine how these perceptions and therapeutic outcomes develop with continued use in real-world deployments.

Third, despite efforts to ensure control and explainability, deploying LLM-powered chatbots in psychotherapy raises ongoing ethical and safety concerns. Even when aligned with expert-crafted dialogues, LLMs may produce unpredictable or inappropriate responses that pose risks in sensitive mental health contexts. Addressing these concerns requires safeguards such as real-time monitoring, fallback mechanisms for high-risk situations, and stronger data privacy and explainability protocols.

Fourth, participants' prior exposure to chatbots or similar AI-driven services may have influenced their perceptions and interactions. Those with more experience might have had different expectations or levels of comfort, potentially affecting their evaluations of the chatbots. While we did not explicitly control the prior exposure for participant recruitment, the counterbalanced study design helps mitigate systematic bias. Future work could explore this variable more directly with stratified analyses or larger samples to better understand how prior exposure shapes user perceptions in chatbots in psychotherapeutic contexts.

While this work advances the alignment of LLMs with expert-crafted dialogues, addressing these limitations is essential to ensure the scalability, generalizability and ethical deployment of LLM-powered psychotherapy chatbots.

8 Conclusion

Our work provided initial evidence that aligning LLMs with expert-crafted dialogues may significantly enhance their performance and user perceptions in chatbot-delivered psychotherapy for behavioral interventions. By creating a dataset of expert-crafted dialogue scripts for Motivational

Interviewing (MI) and Cognitive Behavioral Therapy (CBT), we conducted empirical studies comparing rule-based, pure LLM, and LLM-aligned chatbots to evaluate their conversational quality and therapeutic effectiveness. Our proposed approach, Script-Strategy Aligned Generation (SSAG), integrated human expertise with LLMs, resulting in controllable and engaging chatbots. This approach mitigated data scarcity, reduced reliance on expert input, and enhanced the controllability of LLMs in delivering psychotherapy for behavioral interventions. The findings underscored the importance of human-guided LLMs for controllable and explainable digital mental health tools and contributed to the field of CSCW by advancing scalable, cost-effective and LLM-supported development of psychotherapy chatbots, a promising path for future research in digital mental healthcare and health behavioral intervention.

Acknowledgment

References

[1] 2023. "Should" Statements. https://www.psychologytools.com/resource/should-statements. Accessed: 2024-10-28.

[2] 2024. ATLAS.Ti. https://atlasti.com. Accessed: 2024-9-11.

[3] Selby P. Almusharraf F, Rose J. 2020. Engaging Unmotivated Smokers to Move Toward Quitting: Design of Motivational Interviewing-Based Chatbot Through Iterative Interactions. *J Med Internet Res* 63(6) (2020), 776–777. doi:10.2196/20251.

[4] Dion Alperstein and Louise Sharpe. 2016. The Efficacy of Motivational Interviewing in Adults With Chronic Pain: A Meta-Analysis and Systematic Review. *The Journal of Pain* 17, 4 (2016), 393–403. doi:10.1016/j.jpain.2015.10.021

[5] Hossein Arsham and Miodrag Lovric. 2011. *Bartlett's Test.* Springer Berlin Heidelberg, Berlin, Heidelberg, 87–88. doi:10.1007/978-3-642-04898-2_132

[6] Erkan Basar, Divyaa Balaji, Linwei He, Iris Hendrickx, Emiel Krahmer, Gert-Jan de Bruijn, and Tibor Bosse. 2023. HyLECA: A Framework for Developing Hybrid Long-term Engaging Controlled Conversational Agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 56, 5 pages. doi:10.1145/3571884.3604404

[7] Erkan Basar, Iris Hendrickx, Emiel Krahmer, Gert-Jan Bruijn, and Tibor Bosse. 2024. To What Extent Are Large Language Models Capable of Generating Substantial Reflections for Motivational Interviewing Counseling Chatbots? A Human Evaluation. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, Nikita Soni, Lucie Flek, Ashish Sharma, Diyi Yang, Sara Hooker, and H. Andrew Schwartz (Eds.). ACL, TBD, 41–52. doi:10.18653/v1/2024.hucllm-1.4

[8] Erkan Basar, Xin Sun, Iris Hendrickx, Jan de Wit, Tibor Bosse, Gert-Jan De Bruijn, Jos A. Bosch, and Emiel Krahmer. 2025. How Well Can Large Language Models Reflect? A Human Evaluation of LLM-generated Reflections for Motivational Interviewing Dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 1964–1982. https://aclanthology.org/2025.coling-main.135/

[9] J S Beck. 2011. *Cognitive behavior therapy: Basics and beyond.* Guilford Press.

[10] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3, null (March 2003), 1137–1155.

[11] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open Source Language Understanding and Dialogue Management. arXiv:1712.05181 [cs.CL] https://arxiv.org/abs/1712.05181

[12] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank van der Velde, Gunay Tariverdiyeva, Divyaa Balaji, and Alan Chamberlain. 2022. The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing* 26 (02 2022), 1–25. doi:10.1007/s00779-021-01582-9

[13] Anouck Braggaar, Christine Liebrecht, Emiel van Miltenburg, and Emiel Krahmer. 2024. Evaluating Task-oriented Dialogue Systems: A Systematic Review of Measures, Constructs and their Operationalisations. arXiv:2312.13871 [cs.CL] https://arxiv.org/abs/2312.13871

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[15] Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. 5599–5611. doi:10.18653/v1/P19-1563

[16] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of Text Generation: A Survey. arXiv:2006.14799 [cs.CL] https://arxiv.org/abs/2006.14799

[17] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416 [cs.LG]

[18] European Commission. [n. d.]. Intellectual-Property (IP) Management in EU-funded Projects — intellectual-property-helpdesk.ec.europa.eu. https://intellectual-property-helpdesk.ec.europa.eu/ip-management-and-resources/ip-eu-funded-projects_en.

[19] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F. Jung, Nicola Dell, Deborah Estrin, and James A. Landay. 2024. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *Proceedings of*

the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 446, 18 pages. doi:10.1145/3613904.3642336

[20] Jan de Wit and Anouck Braggaar. 2023. Tilbot: A Visual Design Platform to Facilitate Open Science Research into Conversational User Interfaces. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 55, 5 pages. doi:10.1145/3571884.3604403

[21] Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology* 2, 11 (01 Nov 2023), 688–701. doi:10.1038/s44159-023-00241-5

[22] Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations. arXiv:2305.10172 [cs.CL]

[23] Justin Dieter, Tian Wang, Arun Tejasvi Chaganty, Gabor Angeli, and Angel X. Chang. 2019. Mimic and Rephrase: Reflective Listening in Open-Ended Dialogue. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Mohit Bansal and Aline Villavicencio (Eds.). Association for Computational Linguistics, Hong Kong, China, 393–403. doi:10.18653/v1/K19-1037

[24] Sandra Julia Diller and Jonathan Passmore. 2023. Defining digital coaching: a qualitative inductive approach. *Frontiers in Psychology* 14 (2023). https://api.semanticscholar.org/CorpusID:263716095

[25] Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A Survey of Natural Language Generation. *Comput. Surveys* 55, 8 (Dec. 2022), 1–38. doi:10.1145/3554727

[26] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. doi:10.1145/3411764.3445188

[27] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of advanced nursing* 62, 1 (2008), 107–115.

[28] Laura Eltahawy, Todd Essig, Nils Myszkowski, and Leora Trub. 2024. Can robots do therapy?: Examining the efficacy of a CBT bot in comparison with other behavioral intervention technologies in alleviating mental health symptoms. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100035. doi:10.1016/j.chbah.2023.100035

[29] OpenAI et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[30] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 2 (May 2007), 175–191.

[31] Amanda Ferguson, Lê-Anh Dinh-Williams, and Zindel Segal. 2021. Mindfulness-based cognitive therapy. In *Handbook of cognitive behavioral therapy: Overview and approaches (Vol. 1)*. American Psychological Association, Washington, 595–615.

[32] 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 (06 Jun 2017), e19. doi:10.2196/mental.7785

[33] Marsha E. Fonteyn, Benjamin Kuipers, and Susan J. Grobe. 1993. A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research* 3, 4 (1993), 430–441. arXiv:https://doi.org/10.1177/104973239300300403 doi:10.1177/104973239300300403

[34] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 3816–3830. doi:10.18653/v1/2021.acl-long.295

[35] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL]

[36] Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. arXiv:1703.09902 [cs.CL] https://arxiv.org/abs/1703.09902

[37] Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P. Bigham. 2022. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. arXiv:2205.12673 [cs.CL] https://arxiv.org/abs/2205.12673

[38] James W Hardin and Joseph M Hilbe. 2012. *Generalized estimating equations, second edition* (2 ed.). Chapman & Hall/CRC, Philadelphia, PA.

[39] S C Hayes, K D Strosahl, and K G Wilson. 2012. *Acceptance and commitment therapy: The process and practice of mindful change.* The Guilford Press.

[40] Winston Haynes. 2013. *Bonferroni Correction*. Springer New York, New York, NY, 154–154. doi:10.1007/978-1-4419-9863-7_1213

[41] Linwei He, Erkan Başar, Reinout Wiers, Marjolijn Antheunis, and Emiel Krahmer. 2022. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health* (04 2022). doi:10.1186/s12889-022-13115-x

[42] Linwei He, Erkan Basar, Reinout W Wiers, Marjolijn L Antheunis, and Emiel Krahmer. 2022. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health* 22, 1 (April 2022), 726.

[43] Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess Z. Griffin, and Nicholas C. Jacobson. 2025. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI* 2, 4 (2025), AIoa2400802. arXiv:https://ai.nejm.org/doi/pdf/10.1056/AIoa2400802 doi:10.1056/AIoa2400802

[44] Philip Henson, Pamela Peck, and John Torous. 2019. Considering the therapeutic alliance in digital mental health interventions. *Harv. Rev. Psychiatry* 27, 4 (2019), 268–273.

[45] Zainab Iftikhar, Sean Ransom, Amy Xiao, and Jeff Huang. 2024. Therapy as an NLP Task: Psychologists' Comparison of LLMs and Human Peers in CBT. arXiv:2409.02244 [cs.HC] https://arxiv.org/abs/2409.02244

[46] Muhammad Rafsan Kabir, Rafeed Mohammad Sultan, Ihsanul Haque Asif, Jawad Ibn Ahad, Fuad Rahman, Mohammad Ruhul Amin, Nabeel Mohammed, and Shafin Rahman. 2024. Beyond Labels: Aligning Large Language Models with Human-like Reasoning. arXiv:2408.11879 [cs.CL] https://arxiv.org/abs/2408.11879

[47] Stanisław Karkosz, Robert Szymański, Katarzyna Sanna, and Jarosław Michałowski. 2024. Effectiveness of a Web-based and Mobile Therapy Chatbot on Anxiety and Depressive Symptoms in Subclinical Young Adults: Randomized Controlled Trial. *JMIR Form Res* 8 (20 Mar 2024), e47960. doi:10.2196/47960

[48] Arshia Kermani, Veronica Perez-Rosas, and Vangelis Metsis. 2025. A Systematic Evaluation of LLM Strategies for Mental Health Text Analysis: Fine-tuning vs. Prompt Engineering vs. RAG. arXiv:2503.24307 [cs.CL] https://arxiv.org/abs/2503.24307

[49] Dexter C. Kozen. 1992. *Depth-First and Breadth-First Search*. Springer New York, New York, NY, 19–24. doi:10.1007/978-1-4612-4400-4_4

[50] Jin Myung Kwak, Minseon Kim, and Sung Ju Hwang. 2023. Context-dependent Instruction Tuning for Dialogue Response Generation. arXiv:2311.07006 [cs.CL] https://arxiv.org/abs/2311.07006

[51] Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. 2024. Large Language Models Produce Responses Perceived to be Empathic. arXiv:2403.18148 [cs.CL] https://arxiv.org/abs/2403.18148

[52] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Comput. Surveys* 55 (2021), 1 – 35.

[53] Man Luo, Christopher J. Warren, Lu Cheng, Haidar M. Abdul-Muhsin, and Imon Banerjee. 2024. Assessing Empathy in Large Language Models with Real-World Physician-Patient Interactions. arXiv:2405.16402 [cs.CL] https://arxiv.org/abs/2405.16402

[54] Michael B Madson, Richard S Mohn, Julie A Schumacher, and Alicia S Landry. 2015. Measuring Client Experiences of Motivational Interviewing during a Lifestyle Intervention. *Meas Eval Couns Dev* 48, 2 (April 2015), 140–151.

[55] Carol Maher, Courtney Davis, Rachel Curtis, Camille Short, and Karen Murphy. 2019. A physical activity and diet program delivered by artificially-intelligent virtual health coach: a case-controlled proof-of-concept study (Preprint). *JMIR mHealth and uHealth* 8 (12 2019). doi:10.2196/17558

[56] Renata Martins and Daniel McNeil. 2009. Review of Motivational Interviewing in promoting health behaviors. *Clinical psychology review* 29 (03 2009), 283–93. doi:10.1016/j.cpr.2009.02.001

[57] Hiroshi Matsumoto, Ayano Taniguchi, and Junichi Nishida. 2021. A revised Self-determined Motivation Scale for Exercise with integrated regulation inclusion. *Journal of Health Psychology Research* 34 (05 2021). doi:10.11560/jhpr.200608129

[58] Michael McTear. 2021. *Rule-Based Dialogue Systems: Architecture, Methods, and Tools*. Springer International Publishing, Cham, 43–70. doi:10.1007/978-3-031-02176-3_2

[59] William Miller and Stephen Rollnick. 2002. Motivational Interviewing: Preparing People for Change, 2nd ed. *Journal For Healthcare Quality* 25 (05 2002), 46. doi:10.1097/01445442-200305000-00013

[60] William Miller and Stephen Rollnick. 2002. Motivational Interviewing: Preparing People for Change, 2nd ed. *Journal For Healthcare Quality* 25 (05 2002), 46. doi:10.1097/01445442-200305000-00013

[61] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2002. Manual for the motivational interviewing skill code (MISC). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico* (2002).

[62] Do Min, Veronica Perez-Rosas, Ken Resnicow, and Rada Mihalcea. 2023. VERVE: Template-based ReflectiVE Rewriting for MotiVational IntErviewing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 10289–10302.

[63] Eduardo Mosqueira Rey, Elena Hernandez Pereira, David Alonso Rios, Jose Bobes Bascaran, and Angel Fernandez Leal. 2022. Human-in-the-loop machine learning: a state of the art. (08 2022). doi:10.1007/s10462-022-10246-w

[64] Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A Survey of Large Language Models in Psychotherapy: Current Landscape and Future Directions. arXiv:2502.11095 [cs.CL] https://arxiv.org/abs/2502.11095

[65] natolambert Nathan Lambert; lcastricato Louis Castricato; lvwerra Leandro von Werra; Dahoas1 Alex Havrilla. 2022. Illustrating Reinforcement Learning from Human Feedback (RLHF). (12 2022). https://huggingface.co/blog/rlhf

[66] OpenAI. [n. d.]. https://openai.com/chatgpt/

[67] OpenAI. 2024. Fine-tuning. https://platform.openai.com/docs/guides/fine-tuning [Accessed 17-09-2024].

[68] Long Ouyang, Jeff Wu, Xu Jiang, Carroll L. Wainwright Diogo Almeida, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. (04 2022). doi:10.48550/arXiv.2203.02155

[69] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39. doi:10.1016/j.ijhcs.2018.01.004

[70] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a Chatbot for a Brief Motivational Interview on Stress Management: Qualitative Case Study. *J Med Internet Res* 21, 4 (16 Apr 2019), e12231. doi:10.2196/12231

[71] Corina Pelau, Dan-Cristian Dabija, and Irina Ene. 2021. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior* 122 (2021), 106855. doi:10.1016/j.chb.2021.106855

[72] Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13370–13388. doi:10.18653/v1/2023.findings-emnlp.892

[73] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 5370–5381. doi:10.18653/v1/P19-1534

[74] Denise Rey and Markus Neuhäuser. 2011. *Wilcoxon-Signed-Rank Test*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1658–1659. doi:10.1007/978-3-642-04898-2_616

[75] Mahnaz Roshanaei, Rezvaneh Rezapour, and Magy Seif El-Nasr. 2024. Talk, Listen, Connect: Navigating Empathy in Human-AI Interactions. arXiv:2409.15550 [cs.HC] https://arxiv.org/abs/2409.15550

[76] Rebecca Rubin and Matthew Martin. 1994. Development of a measure of interpersonal competence. *Communication Research Reports* 11 (06 1994), 33–44. doi:10.1080/08824099409359938

[77] Anna Sell, Christer, Miralem Helmefalk, and Leif Marcusson. 2019. Digital coaching to support university students' physical activity. (2019).

[78] Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E. Kraut, and Diyi Yang. 2022. Modeling Motivational Interviewing Strategies on an Online Peer-to-Peer Counseling Platform. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 527 (nov 2022), 24 pages. doi:10.1145/3555640

[79] S. S. SHAPIRO and M. B. WILK. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3-4 (dec 1965), 591–611. doi:10.1093/biomet/52.3-4.591

[80] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (01 Jan 2023), 46–57. doi:10.1038/s42256-022-00593-2

[81] Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-Style Reflection Generation Using Generative Pretrained Transformers with Augmented Context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, 10–20. https://aclanthology.org/2020.sigdial-1.2

[82] Geovana Ramos Sousa Silva and Edna Dias Canedo. 2022. Towards User-Centric Guidelines for Chatbot Conversational Design. *International Journal of Human–Computer Interaction* 40, 2 (Sept. 2022), 98–120. doi:10.1080/10447318.2022.2118244

[83] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (01 Aug 2023), 172–180. doi:10.1038/s41586-023-06291-2

[84] Xin Sun, Jos A. Bosch, Jan De Wit, and Emiel Krahmer. 2023. Human-in-the-Loop Interaction for continuously Improving Generative Model in Conversational Agent for Behavioral Intervention. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 99–101. doi:10.1145/3581754.3584142

[85] Xin Sun, Dimosthenis Casula, Arathy Navaratnam, Anna Popp, Franziska Knopp, Giovanni Busini, Jan Wesołowski, Marie Van Reeth, Elke Reich, Reinout Wiers, and Jos A. Bosch. 2023. Virtual Support for Real-World Movement: Using Chatbots to Overcome Barriers to Physical Activity. In *HHAI 2023: Augmenting Human Intellect (Frontiers in Artificial Intelligence and Applications, Vol. 368)*. 201–214. doi:10.3233/FAIA230084

[86] Xin Sun, Emiel Krahmer, Jan De Wit, Reinout Wiers, and Jos A. Bosch. 2023. Plug and Play Conversations: The Micro-Conversation Scheme for Modular Development of Hybrid Conversational Agent. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 50–55. doi:10.1145/3584931.3606998

[87] Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos T.P. Dobber, and Jos A. Bosch. 2024. Eliciting Motivational Interviewing Skill Codes in Psychotherapy with LLMs: A Bilingual Dataset and Analytical Study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 5609–5621. https://aclanthology.org/2024.lrec-main.498

[88] Xin Sun, Xiao Tang, Abdallah El Ali, Zhuying Li, Xiaoyu Shen, Pengjie Ren, Jan de Wit, Jiahuan Pei, and Jos A. Bosch. 2024. Chain-of-Strategy Planning with LLMs: Aligning the Generation of Psychotherapy Dialogue with Strategy in Motivational Interviewing. arXiv:2408.06527 [cs.CL] https://arxiv.org/abs/2408.06527

[89] Xin Sun, Xiao Tang, Abdallah El Ali, Zhuying Li, Pengjie Ren, Jan de Wit, Jiahuan Pei, and Jos A.Bosch. 2025. Rethinking the Alignment of Psychotherapy Dialogue Generation with Motivational Interviewing Strategies. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 1983–2002. https://aclanthology.org/2025.coling-main.136/

[90] Sara Syed, Zainab Iftikhar, Amy Wei Xiao, and Jeff Huang. 2024. Machine and Human Understanding of Empathy in Online Peer Support: A Cognitive Behavioral Approach. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 875, 13 pages. doi:10.1145/3613904.3642034

[91] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2025. Moral Alignment for LLM Agents. arXiv:2410.01639 [cs.LG] https://arxiv.org/abs/2410.01639

[92] Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 308–319.

[93] Margaret Urban and Stephen Mailey. 2019. Conversation Design: Principles, Strategies, and Practical Application. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–3. doi:10.1145/3290607.3298821

[94] vicuna. [n. d.]. lmsys/vicuna-13b-v1.5 · Hugging Face — huggingface.co. https://huggingface.co/lmsys/vicuna-13b-v1.5.

[95] Changyu Wang, Siru Liu, Hao Yang, Jiulin Guo, Yuxuan Wu, and Jialin Liu. 2023. Ethical Considerations of Using ChatGPT in Health Care. *J Med Internet Res* 25 (11 Aug 2023), e48009. doi:10.2196/48009

[96] Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12047–12064. doi:10.18653/v1/2023.findings-emnlp.806

[97] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]

[98] Anuradha Welivita and Pearl Pu. 2023. Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for

Computational Linguistics.

[99] Bei Xu and Ziyuan Zhuang. 2020. Survey on psychotherapy chatbots. *Concurrency and Computation: Practice and Experience* (12 2020). doi:10.1002/cpe.6170

[100] Wenjie Yang, Anna Fang, Raj Sanjay Shah, Yash Mathur, Diyi Yang, Haiyi Zhu, and Robert E. Kraut. 2024. What Makes Digital Support Effective? How Therapeutic Skills Affect Clinical Well-Being. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 190 (apr 2024), 29 pages. doi:10.1145/3641029

[101] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 517, 14 pages.

[102] Hongbo Zhang, Chen Tang, Tyler Loakman, Bohao Yang, Stefan Goetze, and Chenghua Lin. 2024. CADGE: Context-Aware Dialogue Generation Enhanced with Graph-Structured Knowledge Aggregation. arXiv:2305.06294 [cs.CL] https://arxiv.org/abs/2305.06294

[103] Mingyuan Zhang, Zhaolin Cheng, Sheung Ting Ramona Shiu, Jiacheng Liang, Cong Fang, Zhengtao Ma, Le Fang, and Stephen Jia Wang. 2023. Towards Human-Centred AI-Co-Creation: A Three-Level Framework for Effective Collaboration between Human and AI. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 312–316. doi:10.1145/3584931.3607008

[104] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. arXiv:2210.03493 [cs.CL]

[105] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593 [cs.CL] https://arxiv.org/abs/1909.08593

[106] Wu Zixiu, Balloccu Simone, Kumar Vivek, Helaoui Rim, Reiter Ehud, Reforgiato Recupero Diego, and Riboni Daniele. 2022. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6177–6181. doi:10.1109/ICASSP43922.2022.9746035

Appendix A.1 Overview of the Expert-Crafted Dialogue Scripts

We provide a detailed overview of the dataset we created in this work and introduced in Section 3.

Technique	Psychotherapeutic Topic	Description of Each Topic	Num of Question / Advice / Total Utterance per Topic
Overall			501 / 203 / 1876
	Greeting	Establishes rapport and initiates a dialogue.	2 / 0 / 9
	Transition topic	Facilitates smooth shifts between therapeutic topics.	8 / 0 / 18
	General introduction	Introduces the purpose of the conversation.	14 / 11 / 54
	Status of physical activity	Explores current levels of physical activity to understand baseline behavior.	15 / 13 / 81
	Handling Bad Weather	Discusses strategies to maintain motivation during poor weather.	14 / 14 / 65
	Sleep Quality	Gathers information about sleep habits and explores links to energy/motivation.	3 / 12 / 32
	Setting a Goal	Helps the user identify a meaningful, self-directed behavioral goal.	8 / 4 / 36
MI	Setting a Goal follow up	Revisits previously set goals to reflect on progress, challenges, or adjustments.	27 / 2 / 118
	Motivational onboarding for energy	Educates and motivates the user to consider movement as a way to boost energy.	11 / 9 / 50
	Rating confidence of physical activity	Uses confidence scales to explore and enhance self-efficacy for being active.	10 / 11 / 68
	Evoking self-efficacy	Encourages the user to recognize their own strengths and capacity for change.	10 / 16 / 103
	Improving motivation	Identifies personal reasons for change to boost intrinsic motivation.	39 / 10 / 117
	Supportive social environment	Encourages reflection on positive social support for behavior change.	15 / 7 / 101
	Supportive social environment follow up	Assesses ongoing support and identifies ways to strengthen helpful connections.	22 / 5 / 79
	Mindfulness	Introduces present-moment awareness as a way to manage thoughts and emotions.	17 / 5 / 52
	Mindfulness follow up	Reviews practice experience and reinforces benefits of mindfulness.	7 / 8 / 27
	All or nothing thinking	Identifies black-and-white thinking patterns that can block progress.	47 / 1 / 122
	All or nothing thinking follow up	Applies alternative thinking strategies to reduce rigid thought patterns.	22 / 2 / 63
	Should-statements	Identifies unhelpful internal rules (e.g., "I should exercise every day").	31 / 5 / 73
CBT Practices	Should-statements follow up	Challenges and reframes unrealistic expectations or self-criticism.	15 / 1 / 49
	Implement Intentions	Helps translate intentions into action using planning and cue-based strategies.	45 / 11 / 130
	Stress management	Teaches techniques (e.g., reframing, breathing) to cope with stress effectively.	42 / 9 / 107
	Habit building	Supports creation of consistent, repeatable behaviors that lead to long-term motivation and change.	22 / 17 / 103
	Time management	Offers tools to prioritize and schedule healthy activities realistically.	18 / 12 / 78
	Unsupportive social environment	Identifies social barriers and develops strategies to handle them constructively.	25 / 13 / 91
	Small wins	Encourages small, achievable actions to build momentum and confidence.	12 / 5 / 50

Table 8. Overview of expert-crafted dialogue scripts. "Num of Question / Advice / Total Utterance per Topic" indicates the number of therapeutic questions, general advice, and total utterances included in the dialogue script for each psychotherapeutic topic.

Appendix A.2   Example of Expert-Crafted Dialogue in Motivational Interviewing (MI) with the Topic "Physical Activity Level"
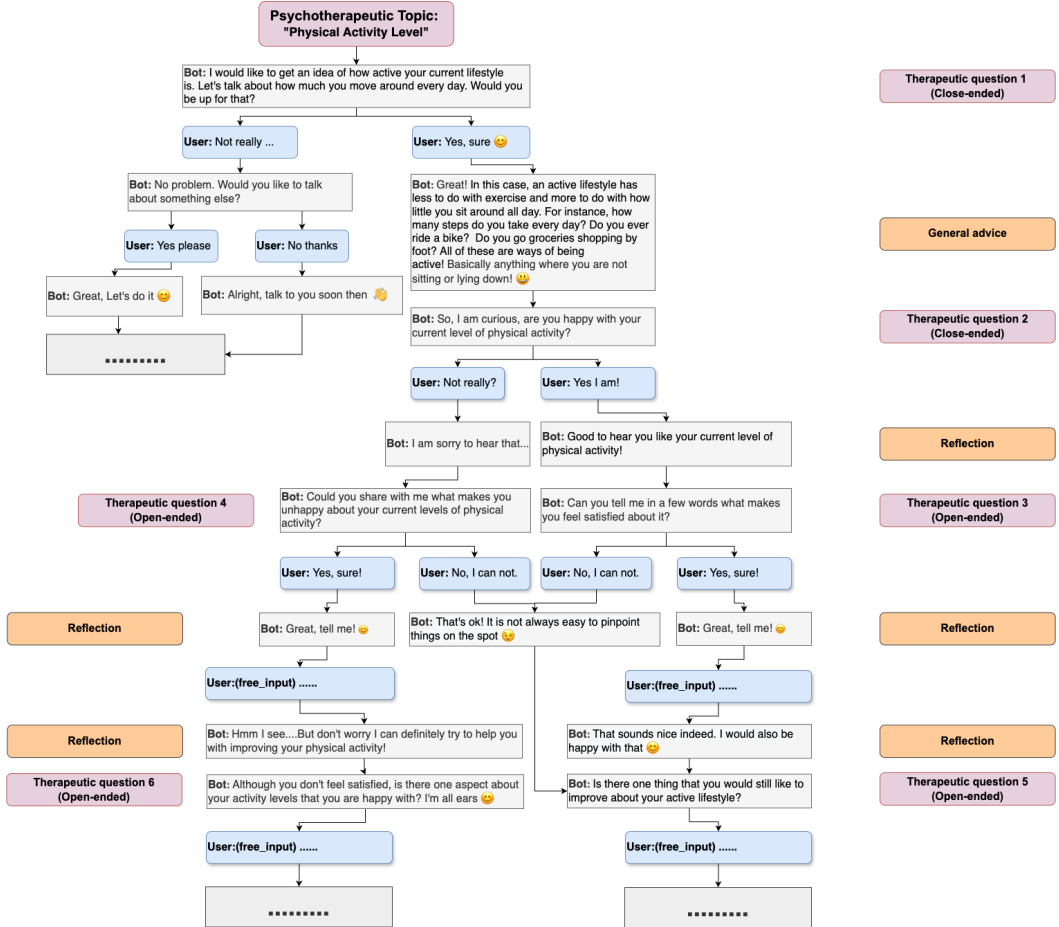

Fig. 12. Examples of expert-crafted dialogues in motivational interviewing (MI) with psychotherapeutic topic "physical activity level". The expert-crafted dialogues contain 1) psychotherapeutic topics, 2) therapeutic questions (both open-/close-ended questions), 3) reflections, 4) general advice, 5) dialogue flows pre-designed by experts.

Appendix A.3   Example of Expert-Crafted Dialogue in Cognitive Behavioral Therapy (CBT) with the Topic "Should Statement"
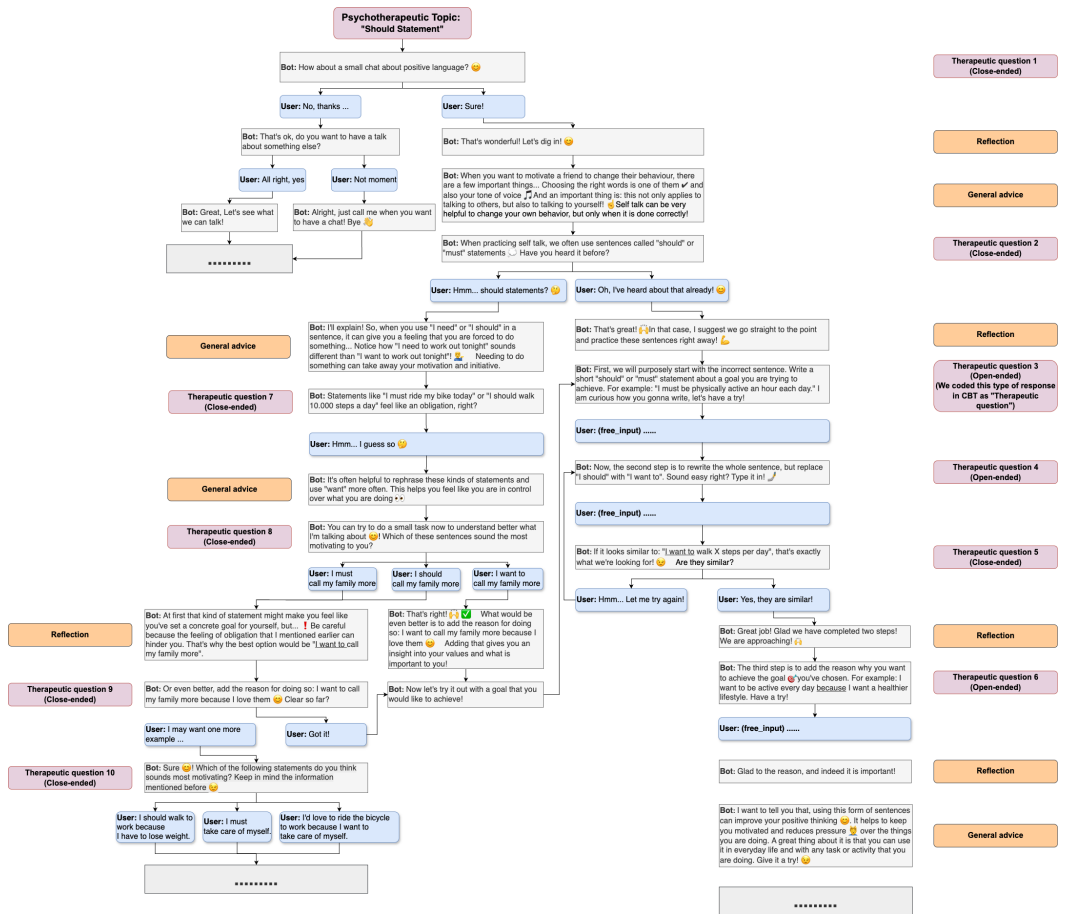

Fig. 13. Examples of expert-crafted dialogues in cognitive behavioral therapy (CBT) with psychotherapeutic topic "should statement". The expert-crafted dialogues contain 1) psychotherapeutic topics, 2) therapeutic questions (both open-/close-ended questions), 3) reflections, 4) general advice, 5) dialogue flows pre-designed by experts.

Appendix B.1   Prompt Template Used in the Study: Pure LLM & LLM-SAG (FT)

Prompt Components	Content
Conversation context	**[The continuously accumulated context of the conversation]** Conversation context: Therapist: [...] Client: [...] Therapist: [...] Client: [...] [......]
Task instruction	**[The base instructions to explain the generation task]** Task: You are a psychotherapist conducting a session to promote healthier behavior using Motivational Interviewing (or Cognitive Behavioral Therapy). The current therapeutic topic is [given topic]. The definition of this topic is [description of the given topic].

Table 9. The prompt template used in the Study 1 for pure LLM chatbots and LLM-SAG (FT) chatbots.

Appendix B.2   Prompt Template Used in the Study: LLM-SAG (Prompt)

Prompt Components	Content
Conversation context	**[The continuously accumulated context of the conversation]** Conversation context: Therapist: [...] Client: [...] Therapist: [...] Client: [...] [......]
Expert-crafted dialogue script	**[The expert-crafted dialogue script (including both dialogue content and flow) for the specific psychotherapeutic topic of the current session.]** All dialogue flows in expert-crafted dialogue scripts of [given topic]: Flow 1: [Pairs of Therapist-Client Dialogues] Flow 2: [Pairs of Therapist-Client Dialogues] Flow 3: [Pairs of Therapist-Client Dialogues] Flow 4: [Pairs of Therapist-Client Dialogues] [......]
Task instruction	**[The base instructions to explain the generation task]** Task: You are a psychotherapist conducting a session to promote healthier behavior using Motivational Interviewing (or Cognitive Behavioral Therapy) following tree-structured dialogue flows. The current therapeutic topic is [given topic]. The definition of this topic is [description of the given topic]. Rules to follow: 1. Identify the relevant dialogue tree based on the current conversation context and user input. 2. At each conversational turn, please use the Tree-of-Thought technique about the decision point and use Breadth-First Search to go level-by-level in the dialogue tree. Do not explore deeper branches unless the current path is chosen. 3. If no exact match is found, please intelligently adapt while staying close to the current dialogue flow/tree. 4. You can give tailored responses to user input and provide dialogue diversity, but please keep staying strictly to the given dialogue flow/tree. 5. Follow the dialogue tree structure and generate possible next responses based on relevance and coherence and with a clear format. 6. If the ongoing conversational topic is finished, please say goodbye to the user and ask the user to go click the button on the left to evaluate this chatbot.

Table 10. The prompt template used in Study 1 and 2 for the LLM-SAG (Prompt) chatbots.

Appendix B.3   Prompt Template Used in the Study: LLM-SSAG

Step	Prompt Components	Content
Step 1 of SSAG		
	Conversation context	**[The continuously accumulated context of the conversation]**
		Conversation context:
		Therapist: [...] Client: [...] Therapist: [...] Client: [...] [......]
	Therapeutic strategy	**[The definition of the therapeutic strategy]**
		The definition of the therapeutic strategy:
		"reflection": [...] "question": [...] "advice": [...]
	Task instruction	**[The base instructions to explain the generation task]**
		Task:
		As a psychotherapist of Motivational Interviewing, please predict the next appropriate therapeutic strategy only from the set ["reflection", "question", "advice"] based on the conversation context. The next therapeutic strategy is (are):
Step 2 of SSAG		
	Conversation context	**[The continuously accumulated context of the conversation]**
		Conversation context:
		Therapist: [...] Client: [...] Therapist: [...] Client: [...] [......]
	Next therapeutic strategy	**[The next therapeutic strategy of the chatbot response]**
		The next therapeutic strategy is (or strategies are): "Reflection" or "Question" or "Advice" or their combination based on the prediction from Step 1.
		The definition of the therapeutic strategy: [definition of the predicted therapeutic strategy].
	Task instruction	**[The base instructions to explain the generation task]**
		Task:
		You are a psychotherapist conducting a session to promote healthier behavior using Motivational Interviewing (or Cognitive Behavioral Therapy). The current therapeutic topic is [given topic].
		The definition of this topic is [description of the given topic].
		You should generate the response strictly aligned with the next therapeutic strategy above.

Table 11. The prompt template used in Study 2 for the LLM-SSAG chatbots.

Appendix C   Implementation Details of Generations with LLMs

We utilized the August 2024 edition of GPT-4o, coded as `gpt-4o-2024-08-06` [1]. We used `openai` Python library to generate with GPT-4. We opted for default hyperparameters, including the temperature as default to control the randomness of generation. The models were used in compliance with their respective licenses and terms at the time of the study. OpenAI provides a Terms of Use[2].

---

[1]https://platform.openai.com/docs/models/gpt-4o
[2]https://openai.com/policies/terms-of-use