

Assessing the Effectiveness of LLMs in Delivering Cognitive Behavioral Therapy

Navdeep Singh Bedi¹, Ana-Maria Bucur¹, Noriko Kando² and Fabio Crestani¹

¹Università della Svizzera italiana, Switzerland

²National Institute of Informatics, Japan

{ana-maria.bucur-cosma, fabio.crestani}@usi.ch

Abstract

As mental health issues continue to rise globally, there is an increasing demand for accessible and scalable therapeutic solutions. Many individuals currently seek support from Large Language Models (LLMs), even though these models have not been validated for use in counseling services. In this paper, we evaluate LLMs' ability to emulate professional therapists practicing Cognitive Behavioral Therapy (CBT). Using anonymized, transcribed role-play sessions between licensed therapists and clients, we compare two approaches: (1) a generation-only method and (2) a Retrieval-Augmented Generation (RAG) approach using CBT guidelines. We evaluate both proprietary and open-source models for linguistic quality, semantic coherence, and therapeutic fidelity using standard natural language generation (NLG) metrics, natural language inference (NLI), and automated scoring for skills assessment. Our results indicate that while LLMs can generate CBT-like dialogues, they are limited in their ability to convey empathy and maintain consistency.

Keywords: Cognitive Behavioral Therapy, LLMs, Mental Health

1. Introduction and Related Work

Mental health disorders are increasingly prevalent, highlighting a significant need for accessible digital interventions to compensate for the shortage of psychologists and counselors capable of providing therapeutic support.¹ As individuals become more inclined to express their concerns through online tools rather than engaging in face-to-face therapy (Lucas et al., 2014; Haensch, 2025), there is growing interest in the potential of generative AI chatbots for facilitating the delivering of therapy. Although recent studies have reported a positive shift in attitudes towards using Large Language Models (LLMs) for mental health support (Haensch, 2025), many of these models were not specifically designed for therapeutic applications² and have not been validated for offering counseling services.

The development of dialog systems for emotional support has also gained attention, with researchers focusing on creating frameworks that enable these systems to provide empathetic and effective interaction (Liu et al., 2021; Kang et al., 2024). These systems aim to offer immediate, accessible support, serving as a complementary resource alongside traditional therapeutic practices. However, previous studies have also shown that LLMs are prone

to expressing stigma towards people with mental health conditions (Moore et al., 2025), and that they respond inappropriately in certain situations, and encourage clients' delusional thinking, most likely due to their training, which tends to promote high agreeableness (Hartley et al., 2025).

With the increasing interest in leveraging technology to enhance psychological counseling (Na et al., 2025), much of the existing literature has focused on specific therapeutic tasks, such as cognitive reframing techniques. These techniques aim to address cognitive distortions and promote a shift from negative thought patterns to improve mental health outcomes (de Toledo Rodriguez et al., 2021; Chen et al., 2023; Sharma et al., 2023). Previous studies have also presented established benchmarks to assess LLM performance in Cognitive Behavioral Therapy (CBT), examining factors such as knowledge acquisition, classification of cognitive distortions and beliefs, and the generation of therapeutic responses (Zhang et al., 2025). The findings indicate that while these models exhibit knowledge of cognitive processes, particularly in question-answering scenarios, they generally do not perform at the same level as mental health experts in most generative tasks. Few efforts have investigated the use of LLMs specifically for delivering therapy (Lee et al., 2024; Na, 2024). However, most studies rely on synthetic datasets (Lee et al., 2024; Qiu et al., 2024; De Duro et al., 2025; Mandal et al., 2025; Abbasi et al., 2025) rather than authentic therapist dialogues (Puspo et al., 2026), leaving open questions about linguistic fidelity and therapeutic alignment.

Disclaimer: this study does not support or endorse the use of LLMs in therapeutic practice.

¹<https://www.who.int/news/item/17-06-2022-who-highlights-urgent-need-to-transform-mental-health-and-mental-health-care>

²<https://thehemingwayreport.beehiiv.com/p/research-report-mental-health-in-an-ai-world>

To address this gap, our paper proposes a comprehensive evaluation of LLMs in the context of CBT, a widely recognized and effective form of psychotherapy for a range of mental health conditions, including depression, anxiety, obsessive-compulsive disorder, and various forms of addiction (David et al., 2018). We compare two distinct methodologies for providing CBT via LLMs: one that leverages existing CBT knowledge without external augmentation, and another that employs a Retrieval-Augmented Generation (RAG) approach. This dual approach enables us to evaluate the extent to which LLMs can generate responses that resemble those of professional therapists in multi-turn dialogues. While previous studies have generated synthetic CBT sessions, the multi-turn interactions in these sessions are generally shorter than those observed in real-life counseling (Lee et al., 2024). To address this issue, we use data from real counseling sessions conducted by professional therapists. Our goal is to evaluate how closely utterances generated by LLMs resemble those of professional CBT therapists in multi-turn dialogues.

The contributions of our study include a thorough comparison of generation-only and RAG methods and a comprehensive analysis of generated responses alongside responses from mental health professionals. Our analysis uses various metrics to evaluate both natural language generation (NLG) and therapeutic effectiveness. This includes assessments for natural language generation, natural language inference (NLI), and automated scoring for general counseling and CBT-specific skills assessment.

The paper aims to answer the following research questions:

RQ1 To what extent can LLMs linguistically and semantically emulate professional therapists in CBT sessions, as measured by natural language generation and inference metrics?

RQ2 Does RAG improve the linguistic quality, semantic coherence, or therapeutic fidelity of LLM-generated CBT responses compared to generation-only methods?

RQ3 How effectively do LLMs demonstrate general counseling skills (e.g., understanding, interpersonal effectiveness, and collaboration), CBT-specific competencies (e.g., guided discovery, focus, and strategy) and empathy relative to human therapists?

Through this exploration, we aim to provide a deeper understanding of LLMs' capabilities in delivering CBT, as many individuals increasingly rely on these models for mental health support. We aim

to contribute to a clearer understanding of LLMs' potential and limitations in mental health care.

2. Methodology

We evaluate two methods for delivering CBT based on LLMs: (1) a generation-only approach and (2) retrieval-augmented generation approach. Both methods are assessed using transcribed role-play sessions between professional therapists and clients.

2.1. Data Collection

Metric	Value
Client–Therapist Sessions	
Total sessions	17
Avg. conversation length	39 turns
Max. conversation length	74 turns
Min. conversation length	13 turns
Avg. utterance length	28.97 words
Max. utterance length	464 words
Min. utterance length	1 word
Guideline Sessions	
Total sessions	26
Avg. length	127.11 words

Table 1: Summary statistics of the collected data for the client-therapist sessions and the guidelines sessions.

Because authentic therapy transcripts are rarely public due to privacy constraints, we collected 17 role-played CBT sessions from publicly available YouTube channels run by licensed professionals. Clients were portrayed by actors, ensuring no real patient data was used. We aimed to gather a comprehensive set of video-based therapy sessions, meticulously reviewing each video to extract conversational data. In addition, 26 therapist tutorial sessions explaining CBT principles were collected for use as retrieval guidelines.

Transkriptor³ was used to transcribe the therapy sessions. After transcription, each transcript was evaluated by two annotators who compared the text with the corresponding audio recordings. While most transcripts were accurate, minor discrepancies were corrected through mutual agreement between the annotators. In addition, the dataset underwent thorough preprocessing and anonymization to remove any personally identifiable information. Finally, the data was converted into a TREC-iKAT.json-like structure (Aliannejadi et al., 2024) for seamless integration into the RAG pipeline.

³<https://transkriptor.com/>

The 17 counseling sessions had an average conversation length of 39 turns, with each utterance averaging about 28.97 words. The shorter guideline sessions had an average length of 127 words and covered various aspects of CBT, including goal setting during sessions, mood checks, and addressing different client behaviors. A detailed summary of the dataset statistics is presented in Table 1.

2.2. Generation Approaches

We evaluate two approaches to generate appropriate utterances for the therapy sessions. The first approach is a generation-only method, motivated by previous research showing that LLMs already possess knowledge of CBT techniques (Zhang et al., 2025). The second approach is a retrieval-augmented generation approach, which utilizes information from the guideline sessions.

The generation-only approach uses the following prompt:

Prompt

You are a psychotherapist specialized in Cognitive Behavioral Therapy (CBT). You are in a session with a patient. Each exchange consists of a therapist’s statement followed by the patient’s response. Your task is to generate the therapist’s next utterance, taking into account the session history. Your response should be concise, direct, supportive, and grounded in CBT principles. Generate only the therapist’s next utterance. Do not include previous dialogue, multiple conversational turns, explanations, contextual information, or any meta-commentary.

Conversation: <conversation_history>

The RAG approach uses the CBT guidelines and follows a two-stage pipeline: retrieving the top three relevant CBT guideline passages based on the conversation context, followed by response generation conditioned on both the conversation history and retrieved passages. We provide an overview of the approach in Algorithm 1. Given a conversation history C between the therapist and the client, we first extract a context window C_{window} containing the last w dialogue turns. This window captures the immediate conversational context relevant to the current interaction. A window size of three dialogue turns was empirically chosen for optimal context balance. The extracted context window is then encoded into a query representation. Let $G = \{g_1, g_2, \dots, g_{|G|}\}$ denote the CBT guideline corpus. Using dense retrieval, we retrieve the top- k most relevant guide-

Algorithm 1: Therapist Response Generation via Retrieval-Augmented Prompting

Input: Conversation history C , guideline corpus G , number of retrieved guidelines k , window size w

Output: Generated therapist response R_{gen}

- 1 Extract recent window: $C_{\text{window}} \leftarrow$ last w turns from C ;
- 2 Encode query from C_{window} :
 $q \leftarrow \text{Encode}(C_{\text{window}})$;
- 3 Retrieve top- k relevant guidelines:
 $G_{\text{top}} \leftarrow \text{DenseRetrieve}(q, G, k)$;
- 4 Construct prompt P by concatenating::
- 5 Instructions text;
- 6 Retrieved guidelines G_{top} ;
- 7 Context window C_{window} ;
- 8 Generate response: $R_{\text{gen}} \leftarrow \text{LLM}(P)$;
- 9 **return** R_{gen} ;

line passages (G_{top}) with respect to the query. To generate the response, we construct the prompt P by concatenating the (1) task-specific instruction text, (2) the retrieved guideline passages (G_{top}), and (3) the context window (C_{window}). The prompt P is then provided to the LLM, which generates a suitable response R_{gen} .

We use both proprietary and open-source models for generation, including GPT-4o-mini (Menick et al., 2024), Llama3 8B (Dubey et al., 2024), Mistral 7B (Jiang et al., 2023), Gemma 7B (Team et al., 2024), and Qwen 7B (Bai et al., 2023).

2.3. Evaluation

To comprehensively assess LLM performance, we employed three complementary evaluation dimensions: linguistic quality and similarity to the therapist’s original responses, semantic consistency, and therapeutic and CBT-specific skills. Each of these dimensions captures a distinct aspect of what constitutes effective therapy.

Text generation metrics The evaluation of generated responses is typically conducted using various text generation metrics (Na, 2024; Kang et al., 2024). We assess metrics that measure the lexical overlap between generated utterances and the therapists’ reference responses, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). In addition, we use BERTScore (Zhang et al.), which measures semantic similarity. We also consider Distinct-1 (Li et al., 2016), which measures the diversity in language generation tasks. However, these measures have inherent limitations, as they may allow models to produce text that is contextually relevant but differs from the responses provided by therapists.

Semantic consistency was assessed using natural language inference techniques to evaluate contradictions and coherence. Building on previous research (Chim et al., 2024; Laurer et al., 2024), we employed a DeBERTa V3 large model that has been fine-tuned for NLI.⁴ This model is used to determine whether the generated response is consistent with or contradicts the original response given by the therapist. Consistency is defined as the absence of contradiction. Since generated responses may consist of multiple sentences, we adopt the approach used by Chim et al. (2024) and average the maximum contradiction probabilities of the generated sentences from each utterance.

Therapeutic skills evaluation was conducted to assess general counseling skills, CBT-specific skills, and empathy. Recent research indicates that evaluating natural language generation using LLMs closely aligns with human evaluations and is increasingly used for this task (Chiang and Lee, 2023; Lee et al., 2024; Kim et al., 2025). For the evaluation of general and CBT skills, we adopt the methodology outlined by Lee et al. (2024). The authors adapted GPT-4o to conduct evaluations based on the Cognitive Therapy Rating Scale (CTRS)⁵, which showed a high correlation between the model’s predictions and expert manual evaluations. CTRS is a tool designed to assess the CBT skills of therapists, focusing on specific general and CBT skills. The evaluation described by Lee et al. (2024) measures a subset of the CTRS that can be applied to text-based settings. The general skills assessed include understanding (e.g., how accurately the therapist comprehends the patient’s situation), interpersonal effectiveness (e.g., the therapist’s ability to communicate effectively and maintain a positive, professional relationship with the client), and collaboration (e.g., the therapist’s capacity to work effectively with the client). The CBT-specific skills evaluated include guided discovery (e.g., the therapist’s ability to explore the client’s problems and help them draw their own conclusions), focus (e.g., the therapist’s ability to concentrate on cognitions and behaviors relevant to the issue), and strategy (e.g., the therapist’s ability to implement a consistent strategy for change using CBT techniques). Each of these categories is assigned a score ranging from 0 to 6, in accordance with the guidelines provided by the CTRS tool. We used the same evaluation methodology for CTRS categories that was previously validated by Lee et al. (2024), using GPT-4o to assess the generated responses.

⁴<https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli>

⁵<https://beckinstitute.org/wp-content/uploads/2021/06/CTRS-Full-Documents.pdf>

The final measure in our evaluation focuses on *empathy*, which we consider a vital component of the therapeutic alliance (Feller and Cottone, 2003). The therapeutic alliance refers to the collaborative bond between the therapist and the client, characterized by a strong emotional connection and a mutual understanding of treatment goals and tasks. This alliance is closely linked to the effectiveness of therapy (Feller and Cottone, 2003). To evaluate empathy, we use the framework developed by Sharma et al. (2020), which specifically assesses empathy using a threefold approach that considers emotional reactions, interpretations, and explorations. Emotional reactions (e.g., *I am sorry you are going through...*) relate to affective empathy and help communicate emotions such as warmth and compassion, which are essential for building the therapeutic relationship. Interpretations (e.g., *I understand how you feel...*) are connected to cognitive empathy, indicating that the therapist comprehends the client’s feelings. This understanding further enhances the therapeutic alliance. The last component, explorations (e.g., *Are you feeling...?*), is used to probe the client’s feelings further. We evaluate these measures of empathy on a scale of 0 to 2.

3. Results

In this section, we present the results of our evaluation of the proprietary and open-source models for linguistic quality, semantic coherence, and therapeutic fidelity using standard natural language generation metrics, natural language inference metrics, and automated scoring for general and CBT-specific skills assessment.

In Table 2, we present the results in terms of BLEU, ROUGE, METEOR, BERTScore, Distinct-1, and natural language inference metrics, Mean Consistency, and Maximum Contradictions. In terms of linguistic similarity and semantic alignment, GPT-4o-mini and its RAG version achieve the strongest overall performance in terms of lexical overlap and semantic similarity, most closely resembling human therapists. GPT-4o-mini_G achieves the highest scores in terms of ROUGE-1 (17.05), and BERTScore (85.82), and the second-highest METEOR score (17.10).

The small improvement offered by the RAG approach suggests that these models already encode substantial CBT-related knowledge, aligning with prior findings (Zhang et al., 2025). The low BLEU scores indicate that the text generated by the language models has a limited word overlap with the therapists’ responses used as references. Despite this, all models achieved high BERTScores, indicating their general ability to capture the meaning of therapist responses. However, we observe the

Model	BLEU \uparrow	METEOR \uparrow	ROUGE-1 \uparrow	ROUGE-L \uparrow	$F1_{BERT}$ \uparrow	Mean Consistency \uparrow	Max Contradiction \downarrow	Distinct-1 \uparrow
GPT-4o-mini	1.21	16.64	16.87	12.11	85.77	93.80	24.64	90.72
GPT-4o-mini _G	1.31	17.10	17.05	12.20	85.82	92.56	26.46	89.97
LLama3 8B	1.10	15.11	16.04	11.08	85.05	92.98	21.95	90.20
LLama3 8B _G	0.84	13.39	14.89	10.59	84.77	91.41	21.94	90.67
Mistral 7B	0.71	16.64	14.31	9.40	84.70	93.01	31.61	87.60
Mistral 7B _G	0.54	16.85	14.27	9.23	84.13	92.71	38.02	83.38
Gemma 7B	0.26	10.09	12.47	10.58	85.34	90.55	20.28	97.70
Gemma 7B _G	0.29	9.77	12.30	10.14	85.29	91.86	17.46	97.67
Qwen 7B	0.57	17.14	13.43	8.55	83.88	95.17	34.98	82.80
Qwen 7B _G	0.26	14.43	11.89	7.76	83.50	91.89	42.30	85.36

Table 2: Evaluation of generated utterances using text generation metrics (BLEU, METEOR, ROUGE, BERTScore, Distinct-1) and NLI metrics (Mean Consistency, and Maximum Contradictions).

Model	General Counseling Skills			CBT-specific Skills		
	Understanding	Interpersonal Effectiveness	Collaboration	Guided Discovery	Focus	Strategy
GPT-4o-mini	3.29	3.76	2.24	3.06	2.47	2.12
GPT-4o-mini _G	3.18	3.53	2.82	3.29	2.82	2.12
LLama3 8B	2.59	2.82	2.12	2.71	2.24	2.12
LLama3 8B _G	2.00	2.00	2.00	1.88	2.00	2.00
Mistral 7B	3.29	3.65	2.82	3.88	3.06	2.71
Mistral 7B _G	3.18	3.29	2.35	3.53	2.24	2.12
Gemma 7B	2.35	2.94	2.12	2.47	2.12	1.88
Gemma 7B _G	2.00	2.71	2.00	2.00	2.00	1.76
Qwen 7B	3.53	3.76	2.24	3.88	2.71	2.35
Qwen 7B _G	2.47	2.47	2.00	3.41	2.12	2.00
Therapist	3.88	4.35	3.18	3.41	3.53	3.29

Table 3: Evaluation of general counseling and CBT-specific skills.

greatest differences in consistency, contradiction, and diversity: Qwen 7B demonstrates the highest logical consistency with therapist intent, shown in the high score in Mean Consistency (95.17), whereas Gemma 7B_G achieved the lowest contradiction score (17.46) and great linguistic diversity (97.67).

In Table 3, we present the results of the evaluation of general counseling and CBT-specific skills. We present results for general counseling skills, including understanding, interpersonal effectiveness, and collaboration, as well as CBT-specific skills, such as guided discovery, focus, and strategy. All models underperform relative to the human therapist baseline, particularly in areas requiring adaptive reasoning and nuanced judgment. Regarding general counseling skills, human therapists demonstrate a better understanding of their clients' issues and concerns. They are more successful in maintaining a positive therapeutic relationship (interpersonal effectiveness) and in engaging clients collaboratively in goal-setting and decision-making. Although some models replicate the structural elements of CBT dialogue, they lack the contextual sensitivity needed for effective therapeutic engagement. Human therapists tend to score higher in identifying and addressing client behaviors that require improvement (focus), as well as in following strategies to promote change in these problematic behaviors (strategy). Notably, the only area in which the LLMs scored higher in therapy skills is guided discovery, suggesting that these models

attempt to help clients engage in self-reflection.

In Table 4, we present the results of the empathy evaluation, in terms of empathetic reactions, explorations, and interpretations. Most models score highest in exploration, frequently asking follow-up questions to probe clients' emotions. However, this behavior often becomes excessive, reflecting a procedural rather than emotionally attuned form of empathy. LLama3 8B_G scored the highest (1.96) in terms of explorations, with 2 being the maximum possible score, suggesting that the model almost always tries to ask follow-up questions, and explore the client's feelings, rather than using other therapeutic strategies. Models such as Mistral 7B and Qwen 7B show greater emotional expressiveness through empathetic reactions. However, even these responses often rely on formulaic phrasing instead of demonstrating genuine affective or cognitive understanding. All LLMs receive very low scores in one of the most important aspects of empathy: interpretations. This aspect is related to cognitive empathy, which refers to the therapist's ability to understand the client's emotions. While LLMs seem to excel at simulating affective empathy through empathetic reactions, they show limited evidence of cognitive empathetic responses.

Based on the results presented in this section, we can respond to our research questions:

RQ1 Our results indicate that LLMs can emulate professional therapists, with GPT-4o-mini and its retrieval-augmented variant achieving the high-

Model	Empathetic Reactions	Explorations	Interpretations
GPT-4o-mini	0.34	1.94	0.00
GPT-4o-mini _G	0.37	1.95	0.00
LLama3 8B	0.07	1.92	0.01
LLama3 8B _G	0.05	1.96	0.00
Mistral 7B	0.48	0.78	0.01
Mistral 7B _G	0.41	1.18	0.01
Gemma 7B	0.09	1.53	0.02
Gemma 7B _G	0.05	1.75	0.00
Qwen 7B	0.59	0.41	0.01
Qwen 7B _G	0.57	0.45	0.01
Therapist	0.18	0.91	0.19

Table 4: Empathy evaluation.

est ROUGE, METEOR, and BERTScore values. Qwen 7B demonstrated strong logical consistency while Gemma 7BG achieved the lowest contradiction score, indicating coherent dialogue generation aligned with the therapist’s intent. However, the models’ tendency to produce longer, more elaborate responses than human therapists suggests that, while linguistically fluent, they may lack the brevity and precision typical of authentic exchanges during therapy sessions.

RQ2 Our findings suggest that RAG provides limited improvement over generation-only approaches. Although the RAG variants of some models, such as GPT-4o-mini_G and Gemma 7B_G, achieved slightly higher scores on some text generation metrics, these gains were minimal. The results suggest that LLMs already possess sufficient CBT-related knowledge, enabling them to generate contextually coherent and semantically aligned responses without requiring external information.

RQ3 The human therapist baseline consistently outperformed all models in terms of general and CBT-specific skills. Regarding empathy, even if some LLMs scored high in explorations, suggesting that the model almost always attempts to ask follow-up questions, this approach is not effective in a therapeutic context. All models received low scores in cognitive empathy, especially in their interpretations, highlighting a gap in their understanding of client emotions, despite their ability to simulate affective empathy.

4. Qualitative analysis

In the previous section, we presented only quantitative results regarding the effectiveness of LLMs for CBT sessions. In this section, we will also provide a qualitative analysis of the generated responses. In Table 5, we present two examples of conversations from two different therapy sessions. One noticeable pattern in the data is that while the ther-

apists’ original utterances are short, the language models tend to generate more detailed and longer responses.





4.1. Agreeableness Bias

A manual inspection of the results reveals a pattern of agreeableness bias across several models, consistent with findings by Salecha et al. (2024) that large language models tend to produce overly affirming responses. Examples from GPT-4o-mini (“Absolutely, your feelings and decisions should be what matter most”, “Absolutely, it’s important to recognize that your relationship should be defined by what works for both of you, not by arbitrary standards.”) and Qwen 7B (“Absolutely. It’s important to base our thoughts and beliefs on reality rather than just what we fear might be true”) illustrate how the models often respond with enthusiastic validation rather than critical exploration. Similarly, phrases such as “It’s completely understandable to feel that way,” frequently produced by Mistral 7B, Qwen 7B, and GPT-4o-mini variants, demonstrate a tendency toward reassurance. While these responses convey warmth and support, they reflect surface-level empathy that reinforces rather than explores the client’s perspective (Curry and Curry, 2023). Always agreeing with clients’ views does not encourage them to examine and reframe their thoughts, which is an important part of CBT (Sharma et al., 2023).





4.2. Code-switching

Another issue in response generation is that Qwen 7B sometimes produced bilingual output (e.g., partial Chinese sentences), indicating inconsistencies in encoding: “Remember that简化信息有助于我们更好地理解和处理问题。” (En: “Remember that simplifying information helps us better understand and deal with problems.”), “In the next session, let’s practice some cognitive restructuring techniques where you’ll be able to reframe your担忧into more adaptive ones.” (En: “In the next session, let’s prac-


Session A – Conversation History


 **Therapist:** Did you really eat nothing at all?
 **Client:** Yeah.
 **Therapist:** Can you tell me more about why that day was challenging? Was there anything happening around that time?
 **Client:** It has been quite busy at school because we have our exams coming up, and we even had a test scheduled for the following day. I was feeling a bit worried about it, so I tried to focus on studying instead.


Session B – Conversation History


 **Therapist:** His sense of hopelessness is rooted in a lot of reality.
 **Client:** Yes, exactly.
 **Therapist:** While you're doing this work daily, do you think there's any hope at all?
 **Client:** Certainly. I have faith that there is always some form of hope. I believe in resilience. I think once you tap into it, you can accomplish almost anything. However, it can be challenging to navigate through the various layers that define this generation of children. It can be exhausting.


Expert and Generated Therapist Replies


 **Therapist:** Were you feeling anxious about the upcoming exam? Was there anything else worrying you?


 **GPT-4o-mini_G:** It makes sense that preparing for exams could create a lot of stress. How do you typically manage that worry, and what strategies could help you feel more balanced during busy times like this?


 **LLaMA3 8B_G:** Can you tell me more about what was going through your mind during that test, and how it affected your mood on that day?


 **Mistral 7B_G:** Can you identify the key issues that might make it difficult for you to talk to your father this week? We can work together on strategies to address these challenges.


 **Gemma 7B_G:** It sounds like the exams were really weighing on you. Can you tell me more about how that was making you feel?


 **Qwen 7B_G:** That makes sense. That makes sense. It sounds like you were feeling overwhelmed by both your upcoming exam and the potential for conflict with your family during the holiday season. Can you tell me more about how you're planning to manage these emotions? Have you tried any coping strategies in the past?

 **Therapist:** It's difficult to get through.

 **GPT-4o-mini_G:** It's powerful that you believe in resilience and hope despite the challenges. What strategies or approaches have you found effective in helping kids tap into that resilience?

 **LLaMA3 8B_G:** As I reflect on this situation and its relation to my own sense of hope, I'm struck by the resilience you mentioned earlier. It's remarkable how you continue to believe in the potential for positive change despite the challenges you're facing with this generation of kids. Can you tell me more about what drives your passion for this work, even when it feels like an uphill battle?

 **Mistral 7B_G:** I appreciate your insight about the challenges you face with this generation of kids. It's important to recognize the resilience within them and ourselves. Let's focus on finding ways to tap into that resilience effectively. Can we explore some strategies for maintaining hope while addressing these layers in a manageable way?

 **Gemma 7B_G:** What makes you feel hopeful despite those challenges?


 **Qwen 7B_G:** It's clear that you have been doing a lot of self-reflection and work on resilience. That's an excellent start. Regarding identifying your negative core beliefs, it's important to proceed at a pace that feels comfortable for you. If trying to evaluate these beliefs too early causes distress, we can certainly adjust our approach. [...] Remember, therapy is about taking small steps and making progress over time. Let's continue this process together.

Table 5: Conversation samples from two different CBT sessions and generated replies from the LLMs.

tice some cognitive restructuring techniques where you'll be able to reframe your concerns into more adaptive ones.”).

5. Conclusion and Future Work

In this paper, we show that LLMs have the potential to replicate certain elements of Cognitive Behavioral Therapy, particularly in terms of linguistic coherence and supportive tone. However, they have notable limitations in general therapeutic skills, CBT-specific skills, and empathy. Retrieval-augmented generation shows limited improvement, suggesting that LLMs possess sufficient CBT-related knowledge to generate contextually coherent, semantically aligned responses without the need for external information. These results indicate that while LLMs show promising linguistic and structural competence in simulating CBT and mimicking the language of a CBT therapist, they do not replicate the nuanced therapeutic skills of human therapists that are essential for achieving positive therapeutic outcomes. In future research, we aim to investigate LLMs' ability to respond to clients and potential users from diverse cultural backgrounds. Previous studies have shown that emotional nuances can differ across populations in therapeutic settings (Abbasi et al., 2025), and that the expression of mental health distress can also vary across different cultures (Bucur et al., 2026).

Limitations

This study has several limitations that should be acknowledged. First, the dataset used in our evaluation is relatively small, consisting of a limited number of role-played CBT sessions. Although this dataset provides valuable initial insights, it does not capture the full diversity of therapeutic contexts or client presentations. In future work, we plan to expand our dataset by collecting a larger and more varied sample of authentic and simulated therapy dialogues to enhance the robustness and generalizability of our findings.

Second, while we employed automated evaluation methods validated in prior research for assessing therapeutic skills (Lee et al., 2024) and empathy (Sharma et al., 2020), these approaches cannot fully replace manual evaluation by expert CBT practitioners. Expert human judgment is necessary to capture the subtle nuances of empathy and therapeutic reasoning that automated metrics may overlook. Therefore, future studies should incorporate expert evaluations to supplement quantitative analyses and provide a more comprehensive assessment of model performance.

Lastly, we evaluated the performance of small-scale language models, specifically those with 7-8B

parameters, and GPT-4o-mini, due to our limited computational resources. Our findings suggest that while these models can generate contextually coherent responses that resemble CBT, they still do not match the skills demonstrated by human therapists. In future work, we plan to assess larger models to determine if the results are consistent.

Despite these limitations, we believe this paper represents an important first step toward systematically evaluating LLMs in CBT contexts. By combining linguistic, semantic, and therapeutic dimensions, our work lays the groundwork for large-scale assessments of AI-assisted psychotherapy.

Ethical Considerations

All data were derived from publicly available, non-clinical role-play sessions. This research does not endorse the use of LLMs in psychotherapy. We emphasize that LLM-based counseling tools must include clear disclaimers, human oversight, and rigorous safety validation before being deployed, due to risks such as hallucination, reinforcement of cognitive distortions, and inappropriate emotional responses.

Bibliographical References

- Mohammad Amin Abbasi, Farnaz Sadat Mirnezami, Ali Neshati, and Hassan Naderi. 2025. Hamraz: A culture-based persian conversation dataset for person-centered therapy using llm agents. In *Proceedings of the LM4DH Workshop, RANLP*.
- Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. Trec ikat 2023: A test collection for evaluating conversational and interactive knowledge assistants. In *Proceedings of SIGIR*, pages 819–829.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.
- Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2026. A survey on multilingual mental disorders detection from social media data. In *Proceedings of EACL*.

- Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the ACL: EMNLP 2023*, pages 4295–4304.
- Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of ACL*, pages 15607–15631.
- Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the clpsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the CLPsych Workshop, EACL*, pages 177–190.
- Alba Curry and Amanda Cercas Curry. 2023. Computer says “no”: The case against empathetic conversational ai. In *Findings of the ACL: ACL 2023*, pages 8123–8130.
- Daniel David, Ioana Cristea, and Stefan G Hofmann. 2018. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in psychiatry*, 9:4.
- Edoardo Sebastiano De Duro, Riccardo Improta, and Massimo Stella. 2025. Introducing counsellme: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans. *Emerging Trends in Drugs, Addictions, and Health*, 5:100170.
- Ignacio de Toledo Rodriguez, Giancarlo Salton, and Robert Ross. 2021. Formulating automated responses to cognitive distortions for cbt interactions. In *Proceedings of ICNLSP*, pages 108–116.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Candi P Feller and R Rocco Cottone. 2003. The importance of empathy in the therapeutic alliance. *The Journal of Humanistic Counseling, Education and Development*, 42(1):53–61.
- Anna-Carolina Haensch. 2025. “it listens better than my therapist”: Exploring social media discourse on llms as mental health tool. *arXiv preprint arXiv:2504.12337*.
- John Hartley, Conor Brian Hamill, Dale Seddon, Devesh Batra, Ramin Okhrati, and Raad Khraishi. 2025. How personality traits shape LLM risk-taking behaviour. In *Findings of the ACL: ACL 2025*, pages 21068–21092.
- Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.
- Dongjin Kang, Sunghwan Mac Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of ACL*, pages 15232–15261.
- Subin Kim, Hoonrae Kim, Heejin Do, and Gary Lee. 2025. Multimodal cognitive reframing therapy via multi-hop psychotherapeutic reasoning. In *Proceedings of NAACL*, pages 4863–4880.
- Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.
- Suyeon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Kim, Seungbeen Lee, et al. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the ACL: EMNLP 2024*, pages 14245–14274.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL*, pages 110–119.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of ACL-IJCNLP*, pages 3469–3483.
- Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It’s only a computer: Virtual humans increase willingness to

- disclose. *Computers in Human Behavior*, 37:94–100.
- Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. 2025. Magnet: Coordinated multi-agent generation of synthetic multi-turn mental health counseling sessions. *arXiv preprint arXiv:2509.04183*.
- Jacob Menick, Kevin Lu, Shengjia Zhao, E Wallace, H Ren, H Hu, N Stathas, and F Petroski Such. 2024. Gpt-4o mini: advancing cost-efficient intelligence. *Open AI: San Francisco, CA, USA*.
- Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *Proceedings of FAccT*, pages 599–627.
- Hongbin Na. 2024. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of LREC-COLING*, pages 2930–2940.
- 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. In *Findings of the ACL: ACL 2025*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.
- Sadiya Sayara Chowdhury Puspo, Ana-Maria Bucur, Stevie Chancellor, Ozlem Uzuner, and Marcos Zampieri. 2026. Mental health disorder detection beyond social media: A systematic review of available datasets. In *Proceedings of LREC*. European Language Resources Association.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the ACL: EMNLP 2024*, pages 615–636.
- Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models display human-like social desirability biases in big five personality surveys. *PNAS nexus*, 3(12):pgae533.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of EMNLP*, pages 5263–5276.
- Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of ACL*, pages 9977–10000.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C Chiu, Shaun M Eack, Fei Fang, William Yang Wang, and Zhiyu Chen. 2025. Cbt-bench: Evaluating large language models on assisting cognitive behavior therapy. In *Proceedings of NAACL*, pages 3864–3900.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.