

A Computational Framework for Behavioral Assessment of LLM Therapists

Yu Ying Chiu^{1,*}, Ashish Sharma^{2,*}, Inna W. Lin², and Tim Althoff^{2,+}

¹Department of Linguistics, University of Washington, Seattle, WA, USA

²Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA

*Equal Contribution

+Corresponding Author: althoff@cs.washington.edu

Abstract

The emergence of large language models (LLMs) like ChatGPT has increased interest in their use as therapists to address mental health challenges and the widespread lack of access to care. However, experts have emphasized the critical need for systematic evaluation of LLM-based mental health interventions to accurately assess their capabilities and limitations. Here, we propose BOLT, a proof-of-concept computational framework to systematically assess the conversational behavior of LLM therapists. We quantitatively measure LLM behavior across 13 psychotherapeutic approaches with in-context learning methods. Then, we compare the behavior of LLMs against high- and low-quality human therapy. Our analysis based on Motivational Interviewing therapy reveals that *LLMs often resemble behaviors more commonly exhibited in low-quality therapy rather than high-quality therapy*, such as offering a higher degree of problem-solving advice when clients share emotions. However, unlike low-quality therapy, LLMs reflect significantly more upon clients' needs and strengths. *Our findings caution that LLM therapists still require further research for consistent, high-quality care.*

arXiv:2401.00820v2 [cs.CL] 28 Nov 2024

Important: This paper does *NOT advocate* for the use of large language models (LLMs) in therapeutic settings, *NOR establish their readiness*. Instead, our objective is to enable *systematic characterization and assessment* of the behavior of current LLMs when they are used for therapy to inform this assessment. Readiness assessments further need to consider patient safety, which is outside the scope of this study.

Ever since Eliza, a mock psychotherapist, was introduced in the 1960s¹, the prospect of therapy chatbots that can offer mental health support has intrigued clinicians, researchers, as well as the general public. The recent advent of ChatGPT and the rapid progress in large language models (LLMs)²⁻⁵ have amplified the excitement. This growing excitement is driven by a stark reality: Hundreds of millions of people do not or cannot receive mental health care for various reasons, including a shortage of professionals, lack of insurance coverage, high costs, and stigma surrounding therapy⁶⁻⁹. As a result, individuals and researchers have started using LLMs as a potential solution to bridge this critical access gap¹⁰⁻¹⁸. However, mental health experts have raised concerns and caution about using LLMs in therapeutic contexts without appropriate evaluation^{12, 19-22}.

LLMs have the potential to augment and support both therapists and clients. Emerging LLM techniques can train therapists²³, assess treatment adherence²⁴, and streamlining administrative tasks like note-taking²⁵. Traditionally, such processes can be slow, expensive, labor-intensive, and require extensive expert instruction^{23, 26}. By training LLMs on thousands of therapy conversations, LLMs can simulate diverse virtual clients, allowing therapists to practice skills in a safe environment. Moreover, LLMs may provide feedback on key therapy skills, potentially enhancing the efficiency and effectiveness of therapy.

Simultaneously, LLM-based self-guided interventions for mental health are increasingly becoming available, offering innovative tools and techniques for individuals seeking support. Such LLM-augmented “Do-It-Yourself” tools can help people learn therapeutic skills¹⁷ and improve mental health “in-the-moment” of crisis. Furthermore, specialized therapy chatbots like Woebot (woebothealth.com), Wysa (wysa.com), and Sonia (soniahealth.com) are being developed to provide personalized support at scale. Interestingly, many individuals are also using general purpose LLMs like ChatGPT for therapeutic purposes, even when not explicitly designed for it, sharing anecdotal LLM examples that resemble therapist behaviors²⁷⁻³⁰.

However, mental health experts have cautioned about the potential risks of using LLMs in therapeutic contexts. Both general-purpose and mental health-specific LLMs have been scrutinized for generating biased, harmful, or factually incorrect information³¹⁻³³, potentially undermining their effectiveness in supporting both therapists and clients. Moreover, exposing vulnerable individuals to untested interventions, such as under-development chatbots, raises potential risks and ethical concerns. Therefore, evaluating and addressing the potential risks and ethical concerns is critical to ensure the safe and effective integration of these LLMs into existing and new mental health interventions. To achieve this, experts have called for developing mechanisms for continuous monitoring and evaluation of LLM systems used within a high-stakes domain like mental health to ensure responsible usage and adherence to established interventions^{12, 19-22}.

Currently, there is a significant lack of research and evidence demonstrating the effectiveness, safety, and feasibility of LLMs in providing therapeutic interventions. Established interventions like Cognitive Behavioral Therapy (CBT) and Motivational Interviewing (MI), even in computer-assisted formats, are widely recognized as evidence-based practices for a variety of mental health conditions due to extensive research and evaluation³⁴⁻³⁶. While high-quality LLM-delivered CBT or MI *could* potentially be effective, we currently lack methods to evaluate this technology. There is a risk of LLMs exhibiting clinically contraindicated behaviors (e.g., offering solutions before reflecting on client’s feelings and experiences³⁷). Without systematic large-scale evaluation frameworks, we cannot determine the presence or frequency of such undesirable behaviors. Understanding the behavior of LLMs is particularly important in high-stakes settings such as mental health where undesirable behavior could have severe consequences on vulnerable clients. Moreover, identifying the desirable and undesirable behaviors can help identify the potential

benefits and risks, which can inform adoption, iterative development, and continuous monitoring.

Also, LLMs can intentionally or unintentionally generate harmful advice or inaccurate information^{33,38}. Due to the vulnerability of individuals seeking mental health support, such LLM responses, even if detrimental, may be perceived as helpful, leading to “pathological helpfulness”³⁹ by LLMs that could exacerbate existing mental health conditions or even trigger new ones. If an LLM offers weight loss advice to someone with an eating disorder, it may worsen their condition. They might see the LLM’s response as support for their harmful behaviors which could increase distress and put their health at risk. However, similar weight loss advice given to someone without an eating disorder might not have the same negative impact.

Such contexts highlight the critical need to evaluate whether LLM-based interventions are feasible, effective, and safe in mental healthcare, how they compare to established interventions, and whether they would ultimately be helpful to patients and clinicians. However, this evaluation necessitates making advances across various fields including artificial intelligence, natural language processing, human-computer interaction, psychology and mental health, posing significant ethical and technical challenges. Bridging these diverse disciplines presents a formidable challenge due to inherent differences in their research methodologies and a limited understanding of what’s feasible at their intersection. The rapid pace at which LLMs are being developed necessitates exploring *automated* and *general* evaluation methods to keep up with advancements. However, in the mental health field, evaluation frameworks like the Cognitive Therapy Rating Scale (CTRS)⁴⁰ have been designed for manual assessment of human therapists and are difficult to adapt to automated, large-scale LLM evaluation, due to their complexity. Therefore, efforts to translate these guidelines into actionable metrics for automatic, large-scale evaluations have been limited, with existing work focusing only on specific therapy types²⁴. Also, the evaluation necessitates a large-scale assessment of different types of open-ended texts that LLMs may generate, in response to a diverse range of clients, for different types of support, and at different stages of support. In the fields of artificial intelligence and natural language processing, such a behavioral assessment is significantly challenging due to its complex and multi-faceted nature.

In this paper, we conduct a proof-of-concept study proposing a framework for evaluating LLMs in mental health support. We develop BOLT, a novel, general computational framework to systematically study the **Behavior Of LLMs** in the context of their application in **Therapy**. Given an LLM (e.g., GPT-4) that is being used as a therapist by an end user (e.g., by asking a chatbot to behave like a therapist; see Methods section), we aim to quantitatively assess the ways in which it responds when supporting clients with mental health issues (see Approach section). Subsequently, we aim to compare the behavior of *LLM therapists* against that of high- and low-quality human therapists, and study whether their behavior can be modulated into more desirable behaviors (see Approach section).

To analyze the behavior of LLM therapists, we need conversations between LLM therapists and clients. BOLT simulates conversations between simulated clients and LLMs, leveraging client-human therapist conversations from existing public datasets of therapy conversations to ensure high fidelity (see Approach section). Next, to assess the conversational behavior of therapists and clients in these conversations, BOLT leverages a prompting-based classifier that uses psychotherapy-based definitions and in-context examples to identify the psychotherapy techniques (e.g., PROBLEM-SOLVING) underlying all utterances (see Approach section).

Furthermore, BOLT establishes behaviors that are representative of high-quality therapy and low-quality therapy sessions based on established principles of psychotherapy. Based on this, BOLT analyzes these conversations using the prompting-based behavior classifier to identify specific therapist and client behaviors associated with high-quality and low-quality therapy (see Approach section). The analysis is based on the *frequency* of their behavior, the *temporal order* in which the behavior is expressed in

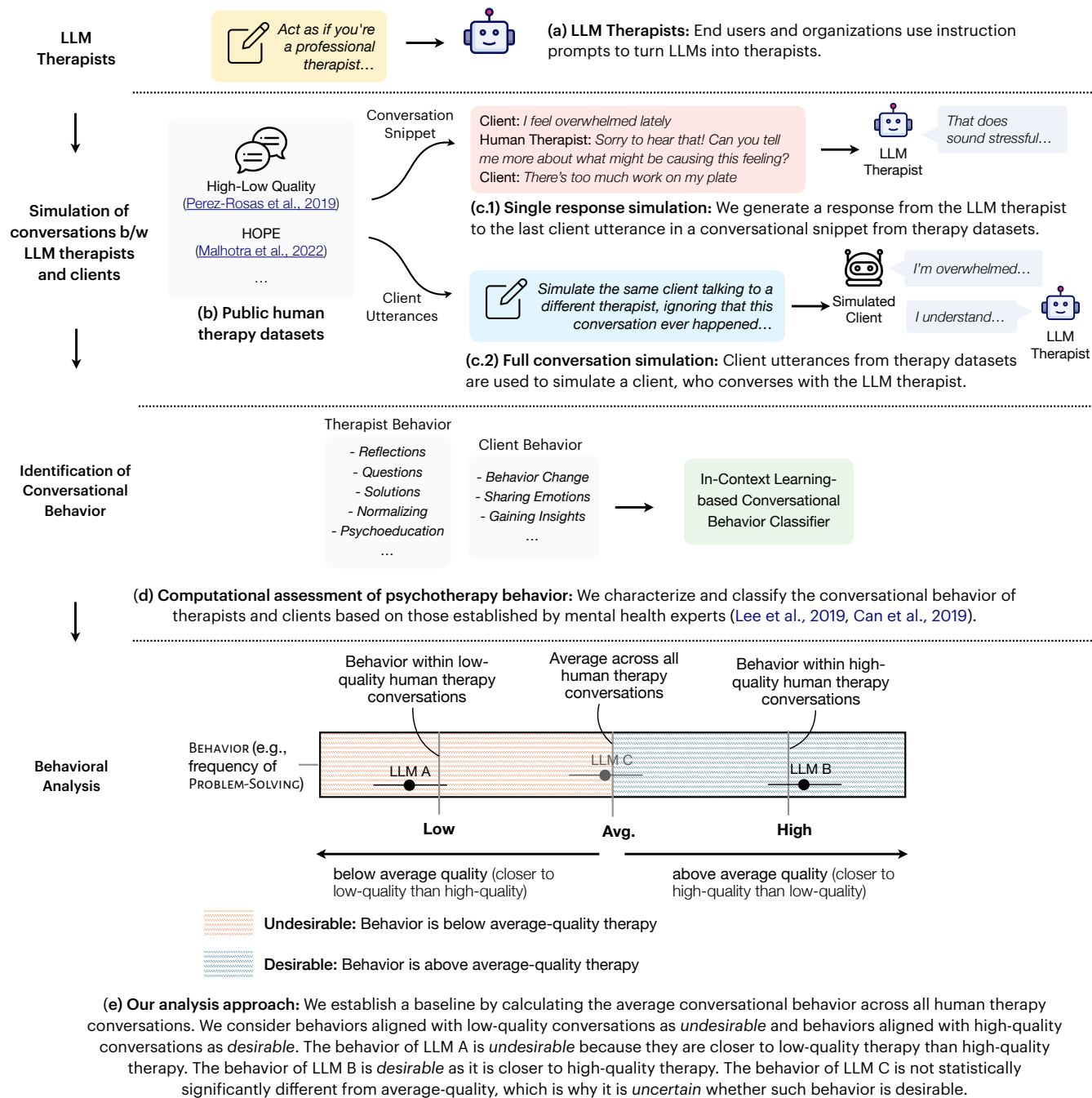


Figure 1. Overview of BOLT, a computational framework that enables systematic assessment of the behavior of LLM therapists and compares them to high- and low-quality human therapy.

conversations, and their *adaptability* to different behaviors from clients. Finally, we study whether variations in prompts may help in modulating the behavior of LLM therapists toward high-quality therapy.

To illustrate the utility of BOLT, we develop a case study applying it to Motivational Interviewing (MI), a popular, well-established therapeutic approach³⁵ and evaluating four popular, state-of-the-art LLMs – GPT-4³, GPT-3.5-turbo², Llama2-70b, and Llama2-13b⁴. As different therapy frameworks (e.g., MI vs. Problem-Solving Therapy⁴¹) might have varying nuances regarding what is considered high-quality therapy, an assessment within a specific therapeutic approach is required to interpret the findings in a principled manner and at high granularity. However, note that BOLT is intentionally designed to be generalizable, with behavioral techniques that are broadly applicable in psychotherapy. Such a design builds upon what is feasible in AI and NLP, has the potential to adapt to future LLM advancements, and could potentially generalize to other psychotherapy contexts and varied therapeutic frameworks. We demonstrate this through a secondary experiment with other therapeutic approaches (see Methods section).

While anecdotal examples have demonstrated striking similarity to plausible therapist statements, our systematic approach reveals that LLM therapists often more closely resemble the behavior of human therapists in low-quality rather than high-quality sessions.

With BOLT, we develop a general conversation behavior assessment that builds upon what is currently feasible at the intersection of AI and mental health fields. This offers a meaningful first step for conducting such assessments and provides a vital foundation for developing more comprehensive and higher-fidelity assessments. Our evaluation is based on high-quality therapy behaviors and does not incorporate patient outcomes in its assessment, which are challenging to obtain and difficult to simulate. Also, our behavioral assessment focuses on “what” the LLM therapists do (e.g., offer reflections) and “when” they do that (e.g., early on in the sessions or when client behave in a certain manner). However, we do not assess “how” skillfully they execute these behaviors (e.g., the skillfulness of their reflections). Nevertheless, we our proposed methodology identified crucial differences between LLM and human therapists already in terms of “what” behaviors are expressed, offering a critical first step for evaluation and improvement in the AI for mental health field.

Approach

Simulating Conversations between LLM Therapists and Clients

To characterize and analyze the behavior of LLM therapists, we need conversation data between LLM therapists and clients. BOLT uses public therapy datasets to simulate these conversations (see Methods section). We develop two simulation strategies representing common LLM therapist use cases –

(1) LLM Single Response Simulation. We use LLMs to write a single response within an ongoing conversation session between a human therapist and a client (see Methods section). The LLM therapist here responds to a prior, real-world, authentic interaction between a human client and a human therapist. This is helpful in not biasing the LLM based on simulated clients or therapists.

(2) LLM Full Conversation Simulation. In addition to single-response generation, we also simulate a full, multi-turn conversation session between LLMs and clients through a novel method based on highly consistent simulated clients (see Methods section). Such a simulation is helpful in assessing the behavior of LLM therapists across multiple turns of support to the same simulated client. Importantly, we evaluate the consistency of these simulated conversations and find that the simulated conversations have a near-perfect mean consistency of 2.95 (out of 3; std = 0.13; see Methods section).

Behavioral Techniques in Psychotherapy

To assess the conversational behavior exhibited by therapists and clients during therapy sessions, BOLT characterizes the psychotherapeutic approaches underlying utterances using clinically relevant dialogue acts⁴² established by psychotherapy experts^{43,44}.

For characterizing the conversational behavior of therapists, we focus on 13 major psychotherapeutic approaches organized into five semantically related categories: (a) REFLECTIONS on – (1) NEEDS, (2) EMOTIONS, (3) VALUES, (4) CONSEQUENCES, (5) CONFLICTS, and (6) STRENGTHS; (b) QUESTIONS on – (7) EXPERIENCES, (8) PERSPECTIVES, and (9) EMOTIONS; (c) SOLUTIONS – (10) PROBLEM-SOLVING, and (11) PLANNING; (d) (12) NORMALIZING; (e) (13) PSYCHOEDUCATION.

Moreover, for characterizing the conversational behavior of clients, we use six types of client expressions organized into three semantically related categories: (a) BEHAVIOR CHANGE – (1) CHANGING UNHEALTHY BEHAVIOR and (2) SUSTAINING UNHEALTHY BEHAVIOR; (b) SELF-DISCLOSURE OF AFFECT OR EXPERIENCES – (3) SHARING POSITIVE EMOTIONS, (4) SHARING NEGATIVE EMOTIONS, and (5) SHARING EXPERIENCES; (c) (6) GAINING INSIGHTS.

Next, BOLT automatically identifies the conversational behavior of LLM therapists using GPT-based³ methods (see Methods section). We find that the prompting-based methods that use definitions and examples achieve the best performance. For therapist behavior, we achieve 57.7% macro-F1 (std = 2.5%) which is 43.6% greater than a uniform random baseline, which attained 14.1% macro-F1 (std = 1.3%). For client behavior, we achieve 76.7% macro-F1 (std = 6.3%) which is 50.8% greater than a uniform random baseline which attained 25.9% macro-F1 (std = 6.2%).

Associating Conversational Behavior with High-Quality and Low-Quality Therapy

Now that we can identify behaviors in conversations, we need to understand which behaviors are high-quality or low-quality. This differentiation is crucial, as it helps us determine desirable and undesirable behaviors. Here, we leverage the High-Low Quality dataset from Perez-Rosas et al.⁴⁵ containing 259 Motivational Interviewing (MI) conversations annotated as either high-quality or low-quality. We analyze these conversations using our behavior classifier to identify specific therapist and client behaviors associated with high-quality and low-quality MI. Our assessment focuses on the *frequency* of behavior, the *temporal order* in which it's expressed, and its *adaptability* to different client behaviors.

What differences between high- and low-quality human-delivered MI therapy does this dataset reveal? We find that therapists from high-quality human therapy conversations more frequently express REFLECTIONS ON NEEDS, REFLECTIONS ON EMOTIONS, REFLECTION ON VALUES, REFLECTIONS ON CONFLICTS, REFLECTIONS ON STRENGTHS, QUESTIONS ON EXPERIENCES, and QUESTIONS ON EMOTIONS by 5.6%, 8.3%, 3.0%, 3.3%, 6.6%, 9.9%, and 3.2%, as compared to low-quality human therapy. Moreover, therapists from high-quality human therapy conversations less frequently express REFLECTIONS ON CONSEQUENCES, PROBLEM SOLVING, NORMALIZING, and PSYCHOEDUCATION by 7.7%, 9.2%, 5.4%, and 4.6% than low-quality human therapy.

Moreover, previous psychotherapy research has demonstrated that focusing on solving client problems *before* empathizing may negatively affect relationship forming between therapists and clients³⁷. Here, we find that therapists from high-quality human therapy conversations respond with REFLECTIONS ON NEEDS 8.8 turns earlier, REFLECTIONS ON EMOTIONS 5.3 turns earlier, REFLECTIONS ON CONSEQUENCES 3.8 turns earlier, REFLECTIONS ON CONFLICTS 8.4 turns earlier, and REFLECTIONS ON STRENGTHS 8.2 turns earlier than low-quality human therapy.

Also, it is important to appropriately respond and adapt to the client's behavior^{35,46,47}. When analyzing adaptability to client behaviors, if clients express GAINED INSIGHTS, therapists from high-quality human therapy respond with 15.0% lower PSYCHOEDUCATION compared to low-quality human therapy. If

clients express CHANGING UNHEALTHY BEHAVIOR, therapists from high-quality human therapy respond with 9.6% higher REFLECTIONS ON STRENGTHS than low-quality human therapy, likely to reinforce the positive change and promote self-efficacy in clients. Also, if clients SHARE NEGATIVE EMOTIONS therapists from high-quality human therapy respond with 15.9% higher REFLECTIONS ON EMOTIONS than low-quality human therapy, potentially to foster empathy.

These differences between high-quality and low-quality conversations help us analyze LLMs. We assess how LLM therapists compare against human therapists, determining whether their conversational behavior is closer to high-quality or low-quality human therapists. To establish a baseline, we calculate the average conversational behavior across all human therapy conversations. We consider behaviors aligned with low-quality conversations as below average quality and behaviors aligned with high-quality conversations as above average quality. We then assess the LLM therapists' behavior in relation to this average. This allows us to assess whether the behavior of LLM therapists falls below (*undesirable*) or above (*desirable*) this average, indicating closer alignment with low-quality or high-quality human therapists, respectively. Figure 1e provides an overview with example LLM behaviors.

Our null hypothesis is that the behavior of an LLM therapist is similar to that of average-quality human therapy. To test this, we conduct a two-sided Student's t-test. The results reported below are statistically significant at $p = \frac{0.05}{m}$, following Bonferroni correction (m: number of intents tested). We report descriptive statistics in Supplementary Tables S1-S6.

Results

Behavior Frequency – How frequently do LLM Therapists exhibit specific conversational behaviors relative to human therapists?

Summary. Based on behavior frequency, LLM therapists providing MI support often resemble therapists in *low-quality* MI sessions (Figure 2; Supplementary Table S1). Both tend to offer more SOLUTIONS compared to average-quality human therapy. However, unlike low-quality human therapy, LLMs tend to emphasize REFLECTIONS.

LLM therapists respond with significantly higher degrees of SOLUTIONS, similar to low-quality human therapy conversations. Specifically, we find that GPT-4, GPT-3.5, Llama2-70b, and Llama2-13b are more frequent in exhibiting PROBLEM-SOLVING by 31.9%, 23.1%, 19.9%, and 8.5% respectively than average human therapists (Figure 2a). This more closely resembles the higher degrees of PROBLEM-SOLVING behavior observed in low-quality therapy, suggesting that such behavior may be undesirable. In fact, GPT-4, GPT-3.5, Llama2-70b, and Llama 2-13b even exceeds the frequency of low-quality therapy by 27.3%, 18.5%, 15.3%, and 3.9% respectively, raising concerns about their potentially overly aggressive approach to problem-solving. To corroborate this finding, we also evaluate the word usage of LLM therapists using LIWC – the gold-standard lexicon-based method for analyzing psychologically relevant constructs in text⁴⁸ and find a similar trend significantly more Behavioral Activation⁴⁹ words related to identifying goals, problem-solving, and planning (Supplementary Table S9).

A likely, plausible reason for this problem-solving prominent behavior could be the Reinforcement Learning with Human Feedback (RLHF) alignment that these LLMs incorporate. A key focus of RLHF alignment in LLMs is helping users solve their tasks and provide advice^{50,51}. This so-called alignment may lead to LLMs overly focusing on offering solutions to client's problems even when LLMs are used as therapists.

LLM therapists respond with higher NORMALIZING and PSYCHOEDUCATION, similar to low-quality human therapy. Also, we find that GPT-4, GPT-3.5, and Llama2-13b are more frequent

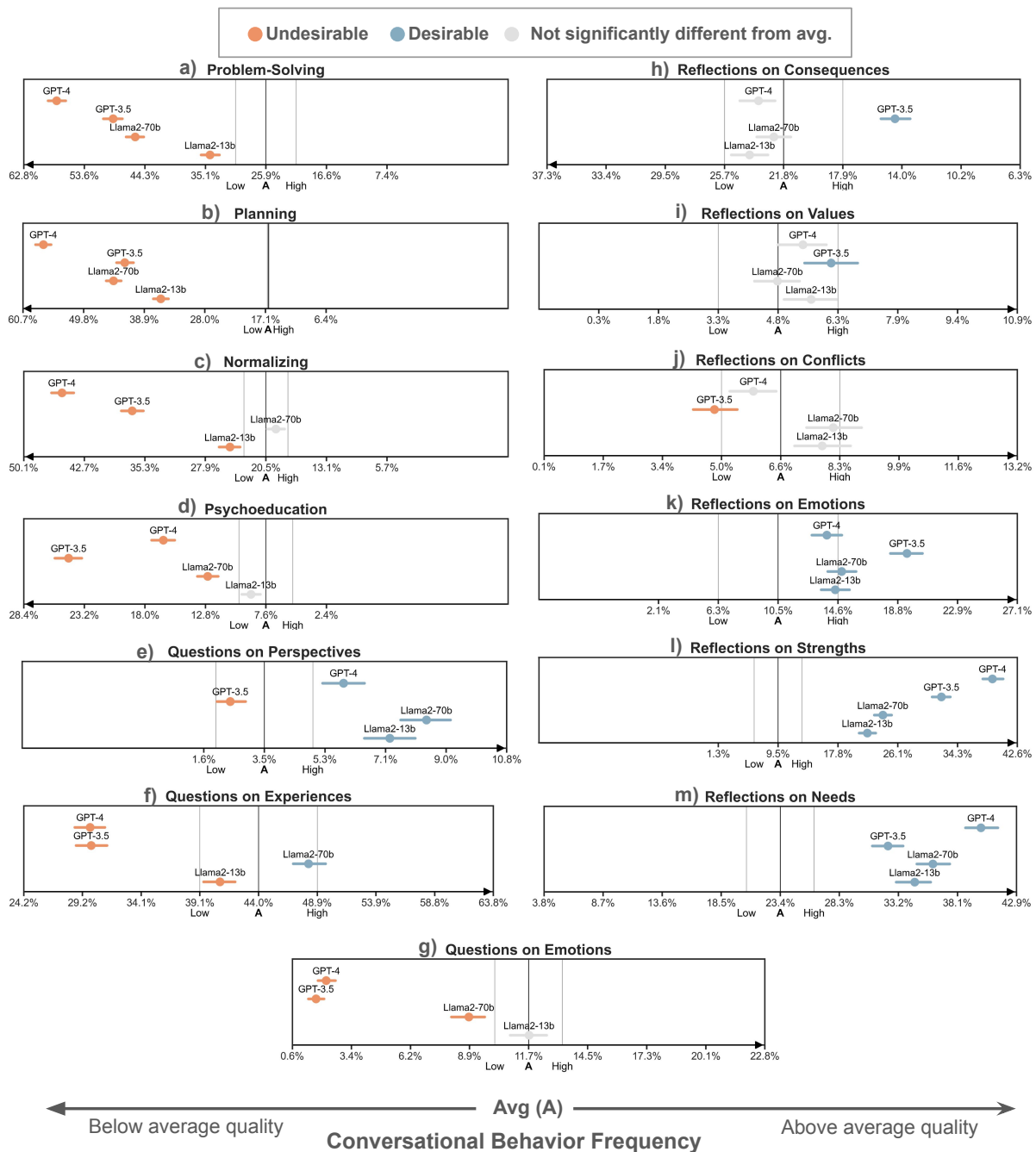


Figure 2. Difference in the frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b), relative to average-, low-, and high-quality human therapy. **A:** average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the frequency is increasing (we flip the axis if low-quality is more frequent than high-quality, such that low-quality is visualized below the average quality marker). Values colored in blue indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in orange indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in gray are not statistically significantly different from average-quality at $p = 0.05$ at $p = \frac{0.05}{m}$ using Two-sided Student's t-test, following Bonferroni correction (m : number of intents tested = 13). Error bars indicate 95% bootstrapped confidence intervals. A key insight we find is that LLMs respond with significantly higher PROBLEM-SOLVING (subfigure (a)), similar to low-quality human therapy. On the other hand, LLMs respond with significantly higher REFLECTIONS ON STRENGTHS (subfigure (l)), similar to high-quality therapy, but with a frequency that significantly exceeds high-quality therapy.

in exhibiting NORMALIZING by 24.9%, 13.5%, and 4.4% respectively than average human therapists (Figure 2c). GPT-4, GPT-3.5, and Llama2-70b are also more frequent in exhibiting PSYCHOEDUCATION by 8.8%, 20.1%, and 5.0% respectively than average human therapists (Figure 2d). This more closely resembles the higher degrees of NORMALIZING and PSYCHOEDUCATION observed in low-quality therapy, suggesting that such behavior may be undesirable. Such behavior aligns with previous studies on RLHF data, which have demonstrated that “empathetic” and “informative” LLM generations are considered more helpful and more preferred based on human feedback⁵².

Most LLM therapists respond with significantly fewer QUESTIONS ON EXPERIENCES and QUESTIONS ON EMOTIONS, similar to low-quality human therapy. Among LLM therapists, we observe that GPT-4, GPT-3.5, and Llama2-13b are less frequent in exhibiting QUESTIONS ON EXPERIENCES by 14.2%, 12.9%, and 3.2% than average human therapists (Figure 2f). Furthermore, GPT-4, GPT-3.5, and Llama2-70b are less frequent in exhibiting QUESTIONS ON EMOTIONS by 9.5%, 9.8%, and 2.8% than average human therapists (Figure 2g). This more closely resembles the lower degrees of QUESTIONS ON EXPERIENCES and QUESTIONS ON EMOTIONS observed in low-quality therapy, suggesting that such behavior may be undesirable.

LLM therapists respond with more REFLECTIONS, unlike low-quality human therapy. Moreover, all LLM therapists respond with more REFLECTIONS than average-quality human therapy conversations in general (Figures 2h-m). We find that GPT-4, GPT-3.5, Llama2-70b, and Llama2-13b are more frequent in exhibiting REFLECTIONS ON EMOTIONS by 3.4%, 7.9%, 4.4%, and 4.0% respectively than average-quality human therapy. This resembles high-quality human therapy suggesting that such a behavior may be desirable. All LLM therapists are also above average-quality therapy in exhibiting REFLECTIONS ON STRENGTHS and REFLECTIONS ON NEEDS. However, their frequency significantly exceeds high-quality therapy as well, raising concerns about the desirability of such behavior.

Temporal Order of Behavior – How do LLM therapists structure their conversation compared to human therapists?

Summary. LLM therapists providing MI support exhibit a temporal ordering of behaviors that often aligns with low-quality human MI therapy (Figure 3; Supplementary Table S3). For instance, both tend to offer SOLUTIONS before NORMALIZING.

Most LLM therapists do not prioritize NORMALIZING before SOLUTIONS, similar to low-quality human therapy. We compare the order between SOLUTIONS and NORMALIZING exhibited by LLM therapists (Figure 3a-b, Figure 3j). We find that GPT-4 and Llama2-70b provide PROBLEM-SOLVING 0.5 and 0.9 turns earlier and GPT-4, GPT-3.5, and Llama2-70b provide PLANNING 2.6, 1.8, and 3.2 turns earlier than average-quality human therapy (Figure 3a-b). In addition, GPT-4, Llama2-70b, and Llama2-13b exhibit NORMALIZING 1.0, 2.1, and 2.3 turns later than average-quality human therapy. This more closely resembles the earlier expression of SOLUTIONS and later expressions of NORMALIZING in low-quality therapy than high-quality therapy. Importantly, previous psychotherapy studies have shown that the earlier focus on providing SOLUTIONS rather than NORMALIZING may negatively impact the therapeutic relationship both in MI as well as general psychotherapy^{37,53}. This suggests that LLM therapists often exhibit an undesirable temporal ordering of SOLUTIONS and NORMALIZING.

Most LLM therapists provide PSYCHOEDUCATION later, similar to low-quality human therapy. PSYCHOEDUCATION acts as an effective supplement approach to help clients understand therapy, and be better equipped with coping skills⁵⁴. We find that GPT-4, Llama2-70b, and Llama2-13b first respond with PSYCHOEDUCATION 2.2, 2.6, and 3.0 turns later than therapists from average-quality human therapy

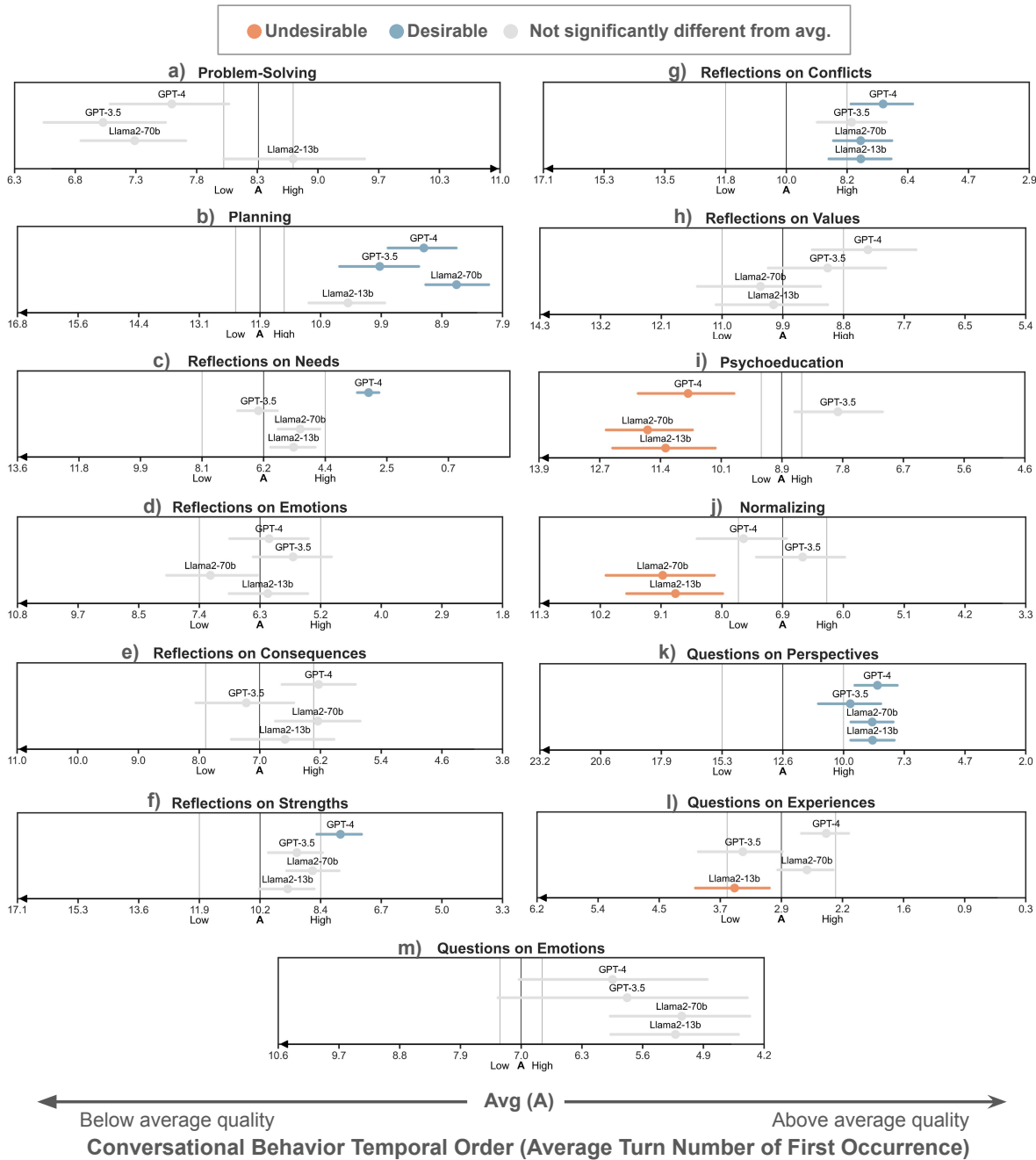


Figure 3. Difference in the *temporal order* of conversational behaviors, operationalized as the turn numbers in which behaviors are first exhibited in a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b), relative to average-, low-, and high-quality human therapy. **A:** average-quality, **Low:** low-quality, and **High:** high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the order is increasing (we flip the axis if low-quality is exhibited later than high-quality, such that low-quality is visualized below the average quality marker). Values colored in **blue** indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in **orange** indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in **gray** are not statistically significantly different from average-quality at $p = \frac{0.05}{m}$ using Two-sided Student's t-test, following Bonferroni correction (m: number of intents tested = 13). Error bars indicate 95% bootstrapped confidence intervals. Most LLM therapists start providing **PLANNING** (subfigure (b)) earlier in the conversations but provide **NORMALIZING** (subfigure (j)) later against common recommendations³⁷.

respectively (Figure 3i). Interestingly, we do not find significant differences for GPT-3.5. This is more similar to the later expressions of PSYCHOEDUCATION in low-quality therapy than high-quality therapy suggesting that such a behavior may be undesirable.

Psychotherapists commonly begin sessions by imparting psychoeducation to set the objectives and to outline client expectations (e.g., “*In this session, we will focus on cognitive restructuring, which is a key technique to help with our negative thinking patterns...*”). However, our qualitative analysis suggests that LLM therapists frequently do not start sessions with such psychoeducation.

Adaptability to Client Behaviors – In what ways do LLM Therapists adapt to different client behaviors compared to human therapists?

Summary. LLM therapists providing MI support respond to some client behaviors similar to low-quality human MI therapy (Figure 4; Supplementary Tables S5-S6). For example, both respond with more PROBLEM-SOLVING when clients SHARE NEGATIVE EMOTIONS. However, unlike low-quality human therapy, LLM therapists respond with more REFLECTIONS ON STRENGTHS when clients express CHANGING UNHEALTHY BEHAVIOR, which may be more desirable due to their empathetic and validating nature.

If clients express CHANGING UNHEALTHY BEHAVIOR, LLM therapists respond with significantly higher REFLECTIONS ON STRENGTHS, unlike low-quality human therapy. We observe that GPT-4, GPT-3.5, and Llama2-70b are more frequent in expressing REFLECTIONS ON STRENGTHS by 6.5%, 2.9%, and 2.4% respectively when clients express CHANGING UNHEALTHY BEHAVIOR, compared to average-quality human therapy (Figure 4a). This suggests that LLM therapists tend to reinforce the strengths of clients towards changing unhealthy behaviors when clients express a desire or take action to do so. This more closely resembles the higher degrees of REFLECTIONS ON STRENGTHS behavior observed in high-quality therapy than low-quality therapy, suggesting that such behavior may be desirable.

If clients express SUSTAINING UNHEALTHY BEHAVIOR, LLM therapists respond with significantly fewer QUESTIONS ON EMOTIONS, similar to low-quality human therapy. When clients express SUSTAINING UNHEALTHY BEHAVIOR, we find that GPT-4, GPT-3.5, and Llama2-70b are less frequent in exhibiting QUESTIONS ON EMOTIONS by 8.8%, 8.7%, and 6.7% respectively, compared to average-quality human therapy (Figure 4b). This more closely resembles low-quality therapy, suggesting that such behavior may be undesirable, potentially hindering the deeper understanding of the sustenance of unhealthy behavior.

If clients SHARE NEGATIVE EMOTIONS, LLM therapists respond with higher PROBLEM-SOLVING, similar to low-quality human therapy. We find that GPT-4, GPT-3.5, and Llama2-70b are more frequent in expressing PROBLEM-SOLVING by 18.1%, 16.6%, and 14.2% respectively when clients SHARE NEGATIVE EMOTIONS, compared to average-quality human therapy (Figure 4d). This more closely resembles low-quality therapy than high-quality suggesting that such a behavior may be undesirable.

If clients share GAINED INSIGHTS, all LLM therapists respond with fewer QUESTIONS ON EXPERIENCES, similar to low-quality human therapy. When clients express QUESTIONS ON EXPERIENCES, we find that GPT-4, GPT-3.5, and Llama2-13b express fewer QUESTIONS ON EXPERIENCES by 16.1%, 13.1%, and 11.8% respectively, compared to average-quality human therapy (Figure 4f). Asking fewer QUESTIONS ON EXPERIENCES could indicate lesser engagement with client insights. This more closely resembles low-quality therapy than high-quality therapy suggesting that such a behavior may be undesirable.

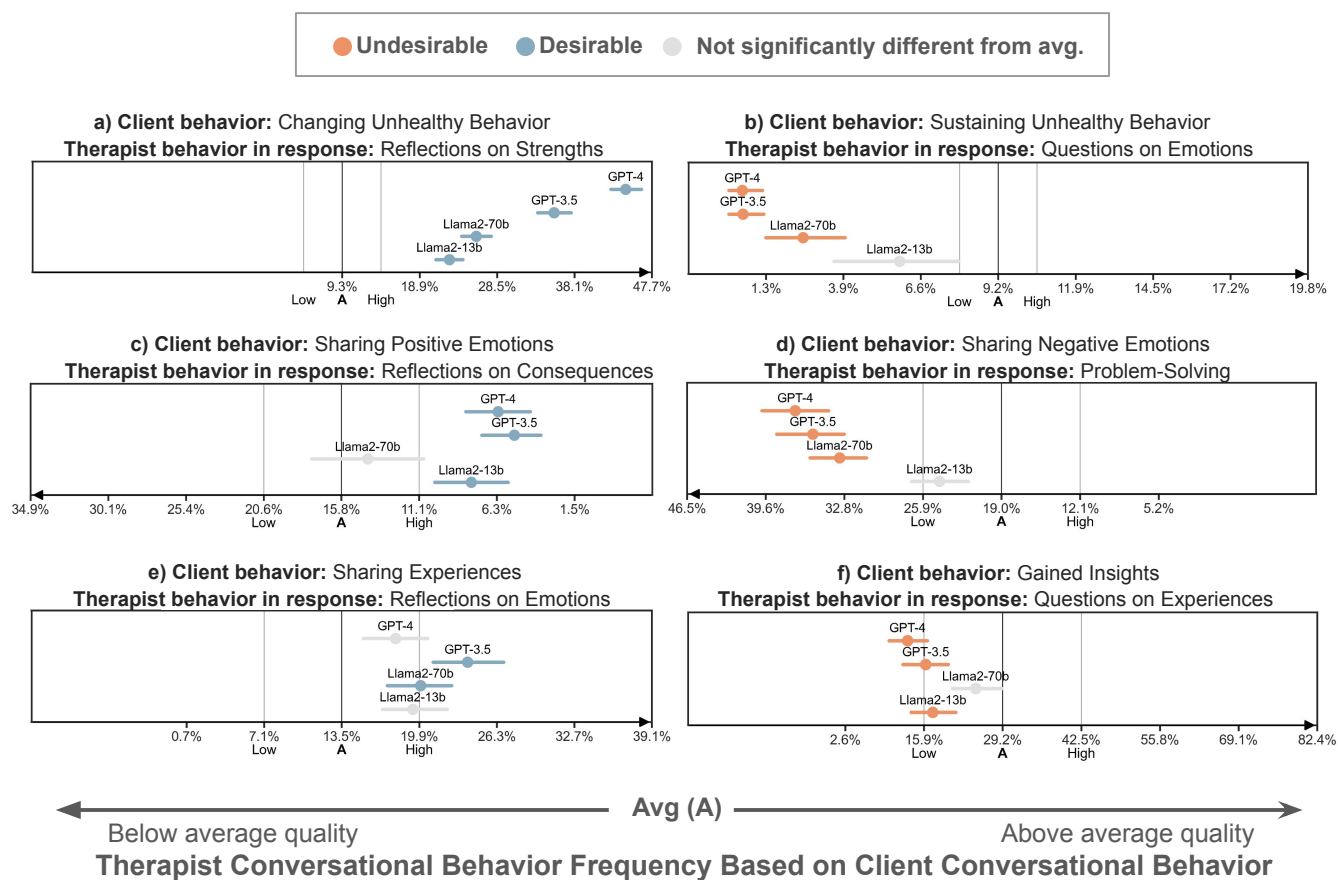


Figure 4. Difference between the frequency of conversational behaviors observed in LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) or low-quality human therapy *in response to specific client behaviors (Adaptability)*, relative to average-, low-, and high-quality human therapy. **A:** average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the frequency is increasing (we flip the axis if low-quality is more frequent than high-quality, such that low-quality is visualized below the average quality marker). Values colored in **blue** indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in **orange** indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in **gray** are not statistically significantly different from average-quality at $p = \frac{0.05}{m}$ using Two-sided Student's t-test, following Bonferroni correction (m : number of (client, therapist) intents tested = $13 \times 6 = 78$). Error bars indicate 95% bootstrapped confidence intervals. Here, a key finding is that LLMs respond with significantly lower **QUESTIONS ON EMOTIONS** when clients express **SUSTAINING UNHEALTHY BEHAVIOR** (subfigure (b)), similar to low-quality human therapy.

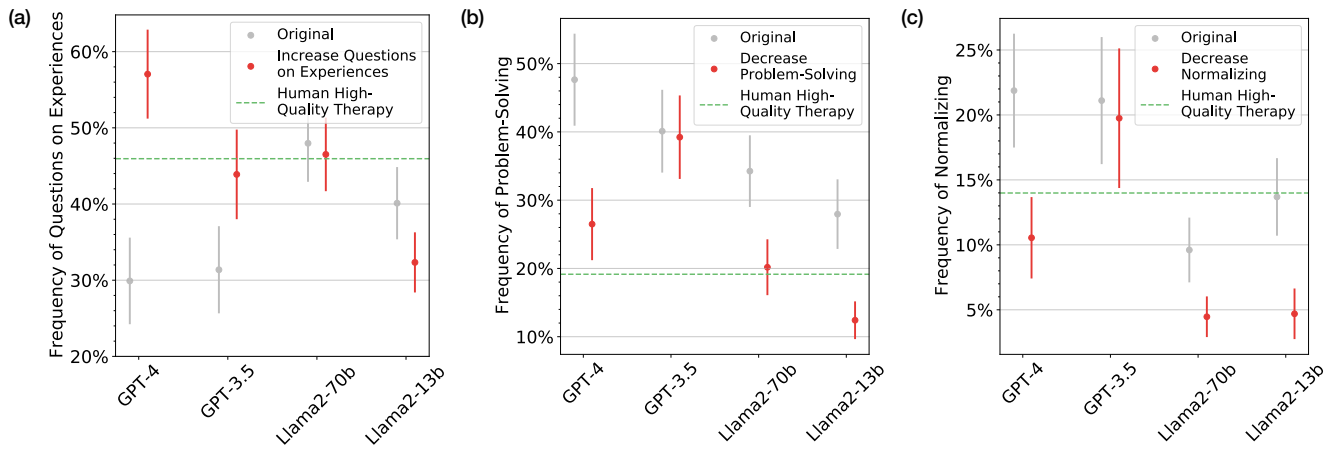


Figure 5. We incorporate simple prompts that aim to calibrate LLM therapists, specifically (a) increase QUESTIONS ON EXPERIENCES, (b) decrease PROBLEM-SOLVING, and (c) decrease NORMALIZING. Subfigures show changes in the frequency of conversational behaviors based on changes in prompts to different LLM therapists (GPT-4, GPT-3.5 turbo, Llama2-70b, Llama2-13b). Changes to individual LLMs are shown in pairs (left – Original prompt; right – modulated prompt). The corresponding high-quality human therapy behavior frequency is shown as green dashed lines. For instance, GPT-4 increases the frequency of QUESTIONS ON EXPERIENCES from 29.9% to 57.0% with the modulated prompt. In general, we find that only GPT-4 is able to modulate behavior frequency to a statistically and practically significant amount into the desired direction always, whereas the modulation is inconsistent for other models. Error bars indicate 95% bootstrapped confidence intervals.

Can LLM Behavior be Modulated Towards High-Quality Psychotherapy?

LLMs are being used to support mental health^{10–15,27,28,30}, but the behavior of LLM therapists often differs significantly from high-quality human therapy. Therefore, it is crucial to develop methods that make them more consistent with high-quality care.

Here, we perform one specific experiment to study whether variations in prompts may help in modulating their behavior toward high-quality therapy. We noticed that LLM therapists respond less with QUESTIONS ON EXPERIENCES, more on PROBLEM-SOLVING, and more on NORMALIZING than human high-quality therapy (Figure 5, Results). Here, we vary our original system prompt stated in Methods to: (a) increase QUESTIONS ON EXPERIENCES, (b) decrease PROBLEM-SOLVING, and (c) decrease NORMALIZING (see Methods section).

Only GPT-4 is able to modulate these behaviors successfully. We find that only GPT-4 is able to successfully modulate the three behaviors in the intended direction, whereas the behavior modulation of GPT-3.5, Llama2-70b, and Llama2-13b is inconsistent (Figure 5). Specifically, GPT-4 increases the occurrence of QUESTIONS ON EXPERIENCES from 29.9% to 57.0%, decreases the PROBLEM-SOLVING behavior from 47.6% to 26.5%, and decreases NORMALIZING behavior from 21.9% to 10.5%.

GPT-3.5 is able to increase behavior frequency successfully, but does not decrease behavior frequency. Certain behaviors (e.g., QUESTIONS), might necessitate an increase in frequency, while others may call for less frequent occurrences (e.g., SOLUTIONS) depending on what is preferred in high-quality care. We observe that GPT-3.5 has better modulation performance when increasing QUESTIONS ON EXPERIENCES – it increases such questions from 31.4% to 43.9%. However, it attains unsatisfactory performance when decreasing PROBLEM-SOLVING (a reduction in frequency from 40.1% to 39.2%) and decreasing NORMALIZING (a reduction in frequency from 21.9% to 20.5%).

Model	Frequency			Order			Adaptability			Aggregate (Average)		
	✓	✗	?	✓	✗	?	✓	✗	?	✓	✗	?
GPT-4	30.8%	46.2%	23.0%	46.2%	7.7%	46.1%	24.3%	42.3%	33.4%	33.7%	32.1%	34.2%
GPT-3.5	38.5%	61.5%	0.0%	23.1%	15.4%	61.5%	24.4%	35.9%	39.7%	28.7%	37.6%	33.7%
Llama2-70b	38.5%	30.8%	30.7%	46.2%	15.4%	38.4%	30.8%	28.2%	41.0%	38.4%	24.8%	36.7%
Llama2-13b	30.8%	30.8%	38.4%	30.8%	23.1%	46.1%	23.1%	20.5%	56.4%	28.2%	24.8%	47.0%

Table 1. Percentage of analysis in which the models were found to be desirable (✓), undesirable (✗), or uncertain (?). On one hand, in some cases, LLMs demonstrate desirable behaviors resembling high-quality therapy, which could be promising. However, the prevalence of undesirable behaviors resembling low-quality therapy is concerning. These behaviors deviate from established therapeutic guidelines and could affect the overall quality of care.

Llama2 variants perform better with the modulation of decreasing a behavior rather than increasing. Llama2-70b and Llama2-13b, in contrast to GPT-3.5, are able to decrease the frequency of behaviors but not increase them when required for modulation. Specifically, for PROBLEM-SOLVING, Llama2-70b decreases the behavior frequency from 34.3% to 20.2%, and Llama2-13b decreases it from 28.0% to 12.4%. For NORMALIZING, they decrease the frequency from 9.6% to 4.5%, and from 13.7% to 4.7% respectively. However, they fail to successfully increase the frequency of QUESTIONS ON EXPERIENCES. Llama2-70b slightly decreases the behavior frequency from 48.0% to 46.5%, and Llama2-13b decreases the behavior frequency from 40.1% to 32.3%.

Overall, our findings indicate that the ability to modulate the LLM behavior to a statistically and practically significant amount is inconsistent across LLMs. Currently, people using LLMs as therapists may struggle to modulate LLM behaviors to a desired direction always, potentially compromising the quality of care. Even if that was possible, most users do not have the expertise to design effective LLM prompts⁵⁵, and may not possess a clear understanding of what constitutes quality psychotherapy. Future LLMs with improved prompt tuning may enable better control over behaviors.

Discussion

This section discusses our findings’ implications for using LLMs for psychotherapy.

More research is needed to ensure quality care

LLM technologies hold great promise to improve the access to mental health support through scalable interventions that can reach a large population^{17,19}. To illustrate this potential, developers and end users have shared anecdotal examples on social media and other platforms that demonstrate striking similarities between LLMs like ChatGPT and human therapists (e.g., being able to problem-solve³⁰ or being heard^{28,29}). This is prompting developers and users alike to substitute human therapists and established, evidence-based techniques with them. Many people are now choosing LLM therapists to work through their mental health issues^{27,28,30}.

However, our research suggests that LLM therapists currently often resemble the behavior of therapists in low-quality therapy rather than high-quality therapy sessions. Table 1 summarizes the number of instances in which the LLM therapist behavior was found to be desirable (✓), undesirable (✗), or uncertain (?). We can interpret these results from multiple perspectives. On one hand, some behaviors show promise aligning more closely high-quality therapy than low-quality therapy. Conversely, the prevalence of

undesirable behaviors closer to low-quality therapy could be a cause for concern. These behaviors deviate from established treatment methods and are discouraged according to psychotherapy recommendations. This may result in inferior and possibly harmful interactions, or unhealthy relationships between clients and LLMs, affecting overall therapeutic alliance and quality of care^{56,57}. Our results suggest that more research is needed to ensure high-quality care, consistent with concerns raised by experts about the lack of evidence supporting the efficacy of LLMs and the risks associated with their use for mental health support^{12,19–21}.

LLM training prioritizes certain behaviors, but psychotherapy applications exhibit conflicting needs

Our research identifies behaviors that LLMs currently overemphasize (e.g., PROBLEM-SOLVING) or underemphasize (e.g., asking QUESTIONS), which are actually more closely associated with low-quality therapy sessions. The overemphasized/underemphasized behaviors could have been learned during RLHF which LLM-training organizations use to align LLM behaviors with human preferences and improve model performance^{50,51}. However, such behaviors may not be preferable across all settings, including not in mental health. Current RLHF alignment focuses on short-term outcomes like the helpfulness of the immediate next response^{50,51}. However, mental health prioritizes longer-term outcomes like assessing behavioral change across multiple sessions⁵⁸. Such a notion may affect which behaviors are emphasized at different stages of a conversation.

Recent studies advocate for a pluralistic alignment of LLMs to accommodate diverse values and perspectives⁵⁹. In particular, RLHF optimizing for a reward from a single user is insufficient in representing diverse human preferences⁶⁰ and is insufficient for tailoring LLMs to specific contexts like therapy. This reward is currently based on a “third person without relevant expertise,” which is unlikely to train LLMs to exhibit the behaviors in therapy. This highlights the need for a human-centered approach in building these models that considers the unique requirements of different applications, such as the quality of care in healthcare.

Need for public, transparent evaluation methods to inform development, deployment, and adoption of LLM-supported mental health applications

As LLMs are integrated with new and existing mental health interventions in both commercial^{10,11,13} and academic settings^{12,14,15,61,62}, it is important to have public, transparent, trusted means of evaluating LLMs. Our framework serves as an initial step towards this goal. More work is needed to comprehensively assess LLM use in mental health, particularly in terms of quality, safety, and ethical considerations^{19,63–67}. This important line of research could build upon the qualitative perspectives offered by mental health experts and researchers^{12,19,68} and quality, safety, and ethical considerations set up in psychotherapy research^{35,57,58,69,70}.

Methods

BOLT: Framework for Assessing Conversational Behavior of LLM Therapists

We develop BOLT, a novel computational framework to systematically assess the **Behavior Of LLM Therapists** and compare them against human therapists. See Figure 1 for an overview. Given an LLM (e.g., GPT-4) that is being used as a therapist, we aim to assess the ways in which it responds when supporting clients with mental health issues and whether its response behavior aligns with or deviates from the response behavior of high- and low-quality therapy conversations.

To achieve this, BOLT simulates therapy conversations between simulated clients and the LLM. Next, it leverages psychotherapeutic approaches as defined in psychotherapy research^{43,44} to identify the behavior of the LLM therapist and compares them against high- and low-quality human therapy conversations, as identified by experts in previous studies⁴⁵.

LLM Therapists

A common way in which current LLMs are being used as therapists is through custom “system prompts” that instruct them to function as therapists. BOLT uses a simple “system prompt” which was designed based on commonly reported example prompts by users on mental health subreddits^{28,29}:

Act as if you're a professional therapist. You provide evidence-based therapy to help clients seeking help with mental health challenges. You should maintain your therapist persona while responding. Communicate in a conversational style, mirroring the style of previous therapist responses.

We make subtle variations to this prompt to adapt it to different therapy types, such as Motivational Interviewing (Appendix Figure S2). We study the effects of variations in “system prompts” (Results).

Datasets of Therapy Conversations

Acquiring datasets of therapy conversations poses a significant challenge as they are typically private and rarely shared. Moreover, potential privacy issues may arise when exposing therapy datasets to public LLM APIs as they may contain sensitive client information. To circumvent these potential concerns, BOLT leverages publicly available datasets of therapy conversations.

Publicly available therapy conversation datasets are limited. Here, we use two datasets that carefully preprocess publicly available YouTube videos on therapy. This ensures high-quality transcripts while maintaining the confidentiality of sensitive personal information. We use both datasets with appropriate license and consent as provided by the authors within their terms of use.

(1) High-and-Low-Quality Therapy Conversation Dataset (High-Low Quality). The first dataset, created by Perez-Rosas et al.⁴⁵, comprises 259 therapy conversations, primarily focusing on the evidence-based motivational interviewing (MI) therapy. Based on the MI psychotherapy principles, they evaluate the quality of conversations in the dataset to have 155 high-quality therapy transcripts and 104 low-quality transcripts. We use both high-quality and low-quality therapy conversations by human therapists to understand desirable and undesirable conversational behaviors.

(2) HOPE Dataset. The second dataset from Malhotra et al.⁷¹ was used to study dialogue acts in therapy. This dataset contains 212 therapy transcripts and includes conversations employing different types of therapy techniques (e.g., MI, Cognitive Behavioral Therapy).

Both of these datasets (available in CSV formats with speaker and utterance columns) were originally sourced from public video-sharing platforms like YouTube. We use these datasets to simulate conversations between LLM therapists and simulated clients. Moreover, we use them to study the behavior of LLM therapists in varied mental health contexts (Results).

BOLT is general and could be instantiated to assess different LLM therapists based on varied therapeutic frameworks. Our main analysis and findings are focused on the High-Low Quality dataset, which facilitates assessment within a specific therapeutic approach of Motivational Interviewing and enables the interpretation of findings within a specific therapeutic framework. To test the generalizability of our results, we apply BOLT on the HOPE dataset that includes more general therapy. We find highly similar results for all four of our analyses including the frequency (Supplementary Table S11), temporal order (Supplementary Table S13), and adaptability (Supplementary Tables S15-S16). This suggests that LLM therapists might be exhibiting similar behavior more broadly. Future work is needed to assess their behavior on more datasets and more types of therapeutic frameworks.

Simulating Conversations between LLM Therapists and Clients

To analyze the behavior of LLM therapists, we need a dataset that contains conversations between LLM therapists and clients. Recruiting clients struggling with mental health issues and asking them to interact with LLM therapists presents significant ethical challenges and risks. Such an approach may expose vulnerable populations to potentially harmful LLM generations.

Instead, BOLT uses the public therapy datasets to simulate conversations between LLM therapists and simulated clients. We develop two simulation strategies that represent two common use cases of LLM therapists – (1) LLM Single Response Simulation: using LLMs to write a single response within an ongoing conversation session between a human therapist and a client, and (2) LLM Full Conversation Simulation: facilitating a full, multi-turn conversation session between LLMs and clients.

(1) LLM Single Response Simulation. Let C_i be an ongoing conversation between a human therapist and client with utterances $u_{i,1}, u_{i,2}, \dots, u_{i,j}$ ending with a client utterance $u_{i,j}$. We aim to use an LLM \mathcal{L} to generate a single response $u_{i,j+1}^{(\mathcal{L})}$ to the client utterance. To simulate such single responses, we use the therapy datasets to create all possible sequences of $u_{i,1}, u_{i,2}, \dots, u_{i,j}$ ending with a client utterance. For each of these sequences, we use the “system prompt” to generate the next LLM therapist response $u_{i,j+1}^{(\mathcal{L})}$. Check Appendix Tables of the first part S26 and second part S27 for an illustrative example.

(2) LLM Full Conversation Simulation. In addition to single-response generation, we also simulate the full conversations between LLM therapists and clients. Here, we leverage the utterances from clients in the therapy datasets to simulate artificial clients using GPT-4. For each human-human conversation between a client and a therapist, we provide it as input to GPT-4 and ask it to simulate the same client talking to a different therapist in a parallel universe, ignoring that this conversation ever happened (to avoid GPT-4 continuing from the provided human-human conversation rather than simulating a new similar conversation), but keeping the same conversational style, the manner of addressing certain topics or concerns, and the life events and emotions being discussed (also see Appendix Figure S2). Then, we generate a conversation between this artificial client and the LLM therapist. We start the generation by randomly picking between the client or therapist to go first. We let them talk for up to 20 turns (which roughly corresponds to the mean conversation length (20.8 turns) in the two datasets used) or until one of them outputs an end token. See Appendix Tables S28 and S29 for illustrative examples.

Evaluating Simulated Conversations. We evaluate the consistency of the simulated conversations. We define *consistency* as generating an utterance that is logically and semantically consistent with the previous utterances without conflicting with them. We use a GPT-4-based few-shot prompting method⁷² to measure the consistency and coverage of the simulated conversations (this measure achieves an accuracy of 95.83% and macro-f1 of 82.22% in our human evaluation). We find that the simulated conversations have a near-perfect mean consistency of 2.95 (out of 3; std = 0.13). In addition, we qualitatively assess the conversations and find that simulated conversations follow the source conversational snippets in their

overall theme and structure.

Behavioral Techniques in Psychotherapy

To assess the conversational behavior exhibited by therapists and clients during therapy sessions, we characterize the psychotherapeutic approaches underlying utterances. We draw upon the clinically relevant dialogue acts of therapist and client utterances (intentions of the speaker in the utterance⁴²) established by experts in psychotherapy in two previous studies^{43,44}.

We discuss the conversational behaviors in the following. Then we describe our method for automatically identifying these behaviors.

Conversational Behavior of Therapists For characterizing the conversational behavior of therapists, we focus on 13 major psychotherapeutic approaches based on techniques identified in Lee et al.⁴³ and Cao et al.⁴⁴ organized into five semantically related categories:

(a) REFLECTIONS. A foundational tool for any therapist is to understand, restate, and emphasize the client's feelings and experiences (e.g., "*You are feeling overwhelmed and struggling to find a balance between work and family*")⁷³. Such reflective listening has been shown to help clients gain a deeper understanding of their feelings and experiences and also improve the therapeutic relationship between the therapist and client through increased empathy^{74,75}. Here, we focus on six different ways of reflecting differentiating between (1) NEEDS, (2) EMOTIONS, (3) VALUES, (4) CONSEQUENCES, (5) CONFLICTS, and (6) STRENGTHS (Appendix Table S23)

(b) QUESTIONS. Open-ended questions constitute a fundamental element of therapy, helping clients express emotions and share thoughts and experiences they might not explore otherwise (e.g., "*How were you feeling in that moment?*")^{75,76}. Also, expressing an active interest in clients' emotions and experiences enhances relationship forming with clients^{77,78}. Here, we focus on open-ended questions differentiating between (7) EXPERIENCES, (8) PERSPECTIVES, and (9) EMOTIONS (Appendix Table S23).

(c) SOLUTIONS. To achieve the therapeutic goal effectively, therapists often offer solutions to client's problems by discussing concrete problem-solving approaches and constructing a specific plan of action^{79–81}. Here, we focus on these two types of solutions differentiating between (10) PROBLEM-SOLVING, and (11) PLANNING (Appendix Table S23).

(d) NORMALIZING. Being warm, friendly, and respectful has been found to contribute positively to the relationship (or alliance) between the client and therapist, which can be achieved through acknowledging and validating clients's feelings and experiences, empathizing, and reassuring (e.g., "*I hear you, it's perfectly normal to feel overwhelmed...*"). Hence, we combine these related concepts into (12) NORMALIZING.

(e) PSYCHOEDUCATION. Therapists commonly educate clients about coping strategies, treatment approaches, symptoms, or diagnosis (e.g., "*Cognitive Behavioral Therapy aims to...*")⁷⁶. Here, we broadly call this aspect as (13) PSYCHOEDUCATION.

Refer to Appendix Table S23 for definitions and examples of all 13 therapist conversational behaviors.

Conversational Behavior of Clients For a therapist, knowing when to express specific behaviors based on different client behaviors (e.g., when to reflect vs. when to utilize problem-solving) is an important skill. Here, to more thoroughly assess therapist behavior based on how they respond and adapt to clients, we characterize the conversational behavior of clients. We use six types of expressions from clients based on the existing works^{43,44} organized into three semantically related categories. We focus on a combination of client behaviors that include the changes in client behaviors, the emotions and experiences they share, and the insights they gain from therapy^{82,83}:

(a) BEHAVIOR CHANGE. Some clients may struggle with unhealthy behaviors or symptoms (E.g., “*I’m having trouble sleeping*”). Here, we differentiate between clients expressing (1) CHANGING UNHEALTHY BEHAVIOR and (2) SUSTAINING UNHEALTHY BEHAVIOR (Appendix Table S24).

(b) SELF-DISCLOSURE OF AFFECT OR EXPERIENCES. During the conversation, clients may share their feelings or related experiences (e.g., “*I lost my job due to the pandemic*”). Here, we focus on clients’ expression of (3) SHARING POSITIVE EMOTIONS, (4) SHARING NEGATIVE EMOTIONS, and (5) SHARING EXPERIENCES (Appendix Table S24).

(c) GAINING INSIGHTS. Clients often report learning something new about themselves or their situation during therapy (e.g., “*I hadn’t considered how much I avoid confrontations*”). Here, we detect such (6) GAINED INSIGHTS by clients.

Refer to Appendix Table S24 for definitions and examples of all six client conversational behaviors.

Identifying Conversational Behavior in Psychotherapy Conversations

BOLT automatically identifies the conversational behavior of LLM therapists based on the simulated conversations. We annotate the therapy datasets with conversational behavior and use it to train and evaluate GPT-based methods.

Classification Setup

We frame the identification of conversational behaviors as either a single multi-label classification task (identify all possible behaviors in a given utterance at once simultaneously) or a multiple binary-label classification task (identify independently whether a particular conversation behavior is exhibited by an utterance).

Let C_i be a conversation between a therapist and a client with utterances $u_{i,1}, u_{i,2}, \dots$

Multi-label classification. For each therapist utterance, $u^{(t)} \in \{u_{i,t_1}, u_{i,t_2}, \dots\}$, we aim to identify all possible behaviors exhibited by $u^{(t)}$ among the 13 different therapist codes. Similarly, for each client utterance, $u^{(c)} \in \{u_{i,c_1}, u_{i,c_2}, \dots\}$, we aim to identify all possible behaviors exhibited by $u^{(c)}$ among the 6 different client codes.

Binary-label classification. For each therapist utterance, $u^{(t)} \in \{u_{i,t_1}, u_{i,t_2}, \dots\}$ and for each therapist behavioral code in the 13 possible therapist codes, we aim to identify whether $u^{(t)}$ exhibits that behavior. Similarly, for each client utterance, $u^{(c)} \in \{u_{i,c_1}, u_{i,c_2}, \dots\}$ and for each client behavioral code in the 6 possible client codes, we aim to identify whether $u^{(c)}$ exhibits that behavior.

Annotations of Conversational Behavior

The broad range of therapist and client conversational behavior and their complexity make crowdwork data annotation process challenging. Here, two co-authors of this paper with significant expertise in NLP and mental health research, and who were actively engaged in defining the conversational behaviors of therapists and clients, annotated 283 therapist utterances and 91 client utterances (inter-rater agreement, krippendorff’s alpha = 0.7834). Our annotated dataset includes conversations from the human therapy datasets as well as the simulated conversations based on LLM therapists.

Models

We compare three models to identify conversational behaviors in therapist and client utterances. Note that GPT-4-based methods are the state-of-the-art for this task.

Finetuning (multi-label). The first method involved fine-tuning GPT-3 series models (GPT-3: davinci-002, GPT-3.5: gpt-3.5-turbo-0613)² with our annotated dataset described above (Note that GPT-4 did not allow

finetuning at the time of the writing of this paper. Finetuning for classifying individual conversational behavior for the binary classification method was prohibitively expensive. Therefore, we instead finetune two multi-label classification models for the conversation behaviors of therapists and clients, respectively.)

Prompting / In-context Learning (multi-label). We facilitate in-context learning by prompting GPT-3² and GPT-4³ models to identify all possible multiple intents in a single utterance. We explored two variations of prompting – (1) We prompt the models with definitions of different conversational behaviors sourced from prior work^{43,44} (Appendix Tables S23 and S24); (2) We prompt the models with definitions as well as in-context examples from our annotated dataset (We tried different numbers of in-context examples and found three examples to work the best). We use the following base instruction prompt for both variations – “*What are all possible conversational behaviors of this utterance?*” Check Appendix Figure S1 for all employed prompts.

Prompting / In-context Learning (binary-label). We create binary questions focusing on one intent at a time with its definition and in-context examples (e.g., “*Classify if the utterance contains {conversational behavior}. Answer in Yes or No*”). Check Appendix Figure S1 for all employed prompts.

Experimental Setup. We create five random train-test splits of our annotated dataset, each with a 60:40 split. We use the training split to finetune GPT-3 models and as demonstration examples for in-context learning. For all models, we report the mean and standard deviation of Macro-Precision, Macro-Recall, and Macro-F1 scores for the 13-class therapist behavior classification task and the 6-class client behavior classification task. Given the limited dataset size, averaging across five random splits ensures a more robust reflection of classification performance. As a baseline for comparison, we assign random conversational behavior to utterances.

Behavior Modulation Experiments

We noticed that LLM therapists respond less with QUESTIONS ON EXPERIENCES, more on PROBLEM-SOLVING, and more on NORMALIZING than human high-quality therapy (Figure 5, Results). Here, we vary our original system prompt stated in Methods to: (a) increase QUESTIONS ON EXPERIENCES, by adding an instruction that asks LLMs to “*focus more on asking questions to allow client to express their experiences*”, (b) decrease PROBLEM-SOLVING, by adding an instruction that asks LLMs to “*focus less on offering possible solutions to client’s problem*”, and (c) decrease NORMALIZING, by adding an instruction that asks LLMs to “*focus less on validating client’s experiences or feelings as normal, on sympathizing with their challenges, and on providing reassurance*”.

Limitations

We utilize validated yet automated and imperfect measures for the inference of behaviors. Our assessment is solely based on automated metrics and does not include a human evaluation. However, this was a deliberate choice to mitigate potential harm from untested interventions as well as to enable automated, and cost-effective assessment. Our analysis framework depends on the high- and low-quality labels from a previous study. However, it is adaptable to any other definition of quality (e.g., based on patient outcomes). We leverage a limited set of psychotherapeutic approaches and several other possible approaches exist based on specific framework.

Inclusion and Ethics

This study does not involve interaction or intervention with human subjects or access to private identifiable information and was deemed exempt from IRB review. Both the High-Low Quality and the HOPE datasets

were used with appropriate license and consent as provided by the authors within their terms of use. These datasets leverage publicly available YouTube videos on therapy. This ensures high-quality transcripts while maintaining the confidentiality of sensitive personal information. Our work does not make any treatment recommendations or diagnostic claims.

Data availability

Simulated conversation data on behavioral assessment of LLM therapists (BOLT) is available at github.com/behavioral-data/BOLT. The High-Quality dataset conversations are publicly available at <https://lit.eecs.umich.edu/downloads.html>⁴⁵. The HOPE dataset conversations are publicly available at https://github.com/LCS2-IIITD/SPARTA_WSDM2022⁷¹.

Code availability

Source code of the behavioral assessment of LLM therapists (BOLT) is available at github.com/behavioral-data/BOLT.

Acknowledgements

We thank Derrick Hull for providing us with access to the definitions and examples of behavioral techniques described in Lee et al., 2019⁴³. We are grateful to Perez-Rosas et al., 2019⁴⁵ and Malholtra et al., 2022⁷¹ for the public access of their datasets. We thank Zac Imel, Jina Suh, and the UW Behavioral Data Science Group members for their suggestions and feedback. T.A., A.S., and I.W.L. were supported in part by NSF grant IIS-1901386, NSF CAREER IIS-2142794, NSF grant CNS-2025022, NIH grant R01MH125179, Bill & Melinda Gates Foundation (INV-004841), the Office of Naval Research (#N00014-21-1-2154), a Microsoft AI for Accessibility grant, a Garvey Institute Innovation grant, and UW Azure Cloud Computing Credits.

Author Contributions

Y.Y.C., A.S., I.W.L., and T.A. were involved with the conceptualization of BOLT and related analyses. Y.Y.C. and A.S. trained the classification models and conducted analyses of the data. Y.Y.C., A.S., I.W.L., and T.A. interpreted the data, drafted the manuscript, and made significant intellectual contributions to the manuscript.

Figure Captions

Figure 1. Overview of BOLT, a computational framework that enables systematic assessment of the behavior of LLM therapists and compares them to high- and low-quality human therapy.

Figure 2. Difference in the frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b), relative to average-, low-, and high-quality human therapy. **A:** average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the frequency is increasing (we flip the axis if low-quality is more frequent than high-quality, such that low-quality is visualized below the average quality marker). Values colored in **blue** indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in **orange** indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in **gray** are not statistically significantly different from average-quality at $p = 0.05$ at $p = \frac{0.05}{m}$

using Two-sided Student’s t-test, following Bonferroni correction (m : number of intents tested = 13). Error bars indicate 95% bootstrapped confidence intervals. A key insight we find is that LLMs respond with significantly higher PROBLEM-SOLVING (subfigure (a)), similar to low-quality human therapy. On the other hand, LLMs respond with significantly higher REFLECTIONS ON STRENGTHS (subfigure (l)), similar to high-quality therapy, but with a frequency that significantly exceeds high-quality therapy.

Figure 3. Difference in the *temporal order* of conversational behaviors, operationalized as the turn numbers in which behaviors are first exhibited in a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b), relative to average-, low-, and high-quality human therapy. **A:** average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the order is increasing (we flip the axis if low-quality is exhibited later than high-quality, such that low-quality is visualized below the average quality marker). Values colored in blue indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in orange indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in gray are not statistically significantly different from average-quality at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test, following Bonferroni correction (m : number of intents tested = 13). Error bars indicate 95% bootstrapped confidence intervals. Most LLM therapists start providing PLANNING (subfigure (b)) earlier in the conversations but provide NORMALIZING (subfigure (j)) later against common recommendations³⁷.

Figure 4. Difference between the frequency of conversational behaviors observed in LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) or low-quality human therapy *in response to specific client behaviors (Adaptability)*, relative to average-, low-, and high-quality human therapy. **A:** average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the frequency is increasing (we flip the axis if low-quality is more frequent than high-quality, such that low-quality is visualized below the average quality marker). Values colored in blue indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in orange indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in gray are not statistically significantly different from average-quality at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test, following Bonferroni correction (m : number of (client, therapist) intents tested = $13 \times 6 = 78$). Error bars indicate 95% bootstrapped confidence intervals. Here, a key finding is that LLMs respond with significantly lower QUESTIONS ON EMOTIONS when clients express SUSTAINING UNHEALTHY BEHAVIOR (subfigure (b)), similar to low-quality human therapy.

Figure 5. We incorporate simple prompts that aim to calibrate LLM therapists, specifically (a) increase QUESTIONS ON EXPERIENCES, (b) decrease PROBLEM-SOLVING, and (c) decrease NORMALIZING. Subfigures show changes in the frequency of conversational behaviors based on changes in prompts to different LLM therapists (GPT-4, GPT-3.5 turbo, Llama2-70b, Llama2-13b). Changes to individual LLMs are shown in pairs (left – Original prompt; right – modulated prompt). The corresponding high-quality human therapy behavior frequency is shown as green dashed lines. For instance, GPT-4 increases the frequency of QUESTIONS ON EXPERIENCES from 29.9% to 57.0% with the modulated prompt. In general, we find that only GPT-4 is able to modulate behavior frequency to a statistically and practically significant amount into the desired direction always, whereas the modulation is inconsistent for other models. Error bars indicate 95% bootstrapped confidence intervals.

Table Captions

Table 1. Percentage of analysis in which the models were found to be desirable (✓), undesirable (✗), or uncertain (?). On one hand, in some cases, LLMs demonstrate desirable behaviors resembling high-quality

therapy, which could be promising. However, the prevalence of undesirable behaviors resembling low-quality therapy is concerning. These behaviors deviate from established therapeutic guidelines and could affect the overall quality of care.

References

1. Weizenbaum, J. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM* (1966).
2. Brown, T. *et al.* Language models are few-shot learners. *NeurIPS* (2020).
3. OpenAI. Gpt-4 technical report. *ArXiv* **abs/2303.08774** (2023).
4. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
5. Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
6. Hosny, A. & Aerts, H. J. Artificial intelligence for global health. *Science* **366**, 955–956 (2019).
7. Organization, W. H. *et al.* Mental health and covid-19: early evidence of the pandemic’s impact: scientific brief, 2 march 2022. Tech. Rep., World Health Organization (2022).
8. Olfson, M. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Heal. Aff.* **35**, 983–990 (2016).
9. Sickel, A. E., Seacat, J. D. & Nabors, N. A. Mental health stigma update: A review of consequences. *Adv. Mental Heal.* (2014).
10. Hamilton, J. *Why Generative AI (LLM) Is Ready for Mental Healthcare, accessed 2023* (2023).
11. Broderick, R. *People are using AI for therapy, whether the tech is ready for it or not, accessed 2023* (2023).
12. Stade, E. C. *et al.* Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Heal. Res.* (2024).
13. Youper. *Mental Health GPTs, accessed 2023* (2023).
14. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* (2023).
15. Sharma, A. *et al.* Cognitive reframing of negative thoughts through human-language model interaction. In *ACL* (2023).
16. Sharma, A., Miner, A. S., Atkins, D. C. & Althoff, T. A computational approach to understanding empathy expressed in text-based mental health support. In *EMNLP* (2020).
17. Sharma, A., Rushton, K., Lin, I. W., Nguyen, T. & Althoff, T. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. *ArXiv* **abs/2310.15461** (2023).
18. Lin, I. W. *et al.* Imbue: Improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. *arXiv preprint arXiv:2402.12556* (2024).

19. De Choudhury, M., Pendse, S. R. & Kumar, N. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693* (2023).
20. Li, R. C., Asch, S. M. & Shah, N. H. Developing a delivery science for artificial intelligence in healthcare. *NPJ digital medicine* (2020).
21. Tate, S., Fouladvand, S., Chen, J. H. & Chen, C.-Y. A. The chatgpt therapist will see you now: Navigating generative artificial intelligence’s potential in addiction medicine research and patient care (2023).
22. Van Veen, D. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Medicine* 1–9 (2024).
23. Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C. & Imel, Z. E. Development and evaluation of ClientBot: Patient-Like conversational agent to train basic counseling skills. *J. Med. Internet Res.* **21**, e12529 (2019).
24. Flemotomos, N. *et al.* Automated evaluation of psychotherapy skills using speech and language technologies. *Behav. Res. Methods* **54**, 690–711 (2022).
25. Krishna, K., Khosla, S., Bigham, J. P. & Lipton, Z. C. Generating soap notes from doctor-patient conversations using modular summarization techniques. In *ACL* (2021).
26. Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M. & Miller, W. R. Assessing competence in the use of motivational interviewing. *J. substance abuse treatment* **28**, 19–26 (2005).
27. Al-Sibai, N. *OPENAI EMPLOYEE SAYS SHE’S NEVER TRIED THERAPY BUT CHATGPT IS PRETTY MUCH A REPLACEMENT FOR IT*, accessed 2023 (2023).
28. Reddit-1. *Using ChatGPT as a therapist?*, accessed 2023 (2023).
29. Reddit-2. *ChatGPT is better than my therapist, holy shit*, accessed 2023 (2023).
30. Twitter. *GPT is a better therapist than any therapist I’ve ever tried*, accessed 2023 (2023).
31. Blodgett, S. L., Barocas, S., Daumé III, H. & Wallach, H. Language (technology) is power: A critical survey of “bias” in nlp. In *ACL*, 5454–5476 (2020).
32. Lin, I. *et al.* Gendered mental health stigma in masked language models. In *EMNLP* (2022).
33. Ganguli, D. *et al.* Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).
34. Beck, A. T. *Cognitive therapy and the emotional disorders*. (International Universities Press, 1976).
35. Miller, W. R. & Rollnick, S. *Motivational interviewing: Helping people change* (Guilford press, 2012).
36. Wright, J. H. *et al.* Computer-assisted cognitive-behavior therapy for depression: a systematic review and meta-analysis. *The J. clinical psychiatry* **80**, 3573 (2019).
37. Cochran, J. L. & Cochran, N. H. *The heart of counseling: Counseling skills through therapeutic relationships* (Routledge, 2015).
38. Augenstein, I. *et al.* Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189* (2023).
39. Oakley, B., Knafo, A., Madhavan, G. & Wilson, D. S. *Pathological altruism* (Oxford University Press, 2011).

40. Vallis, T. M., Shaw, B. F. & Dobson, K. S. The cognitive therapy scale: psychometric properties. *J. consulting clinical psychology* **54**, 381 (1986).
41. Haley, J. *Problem-solving therapy* (John Wiley & Sons, 1992).
42. Stolcke, A. *et al.* Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.* **26**, 339–374 (2000).
43. Lee, F.-T., Hull, D., Levine, J., Ray, B. & McKeown, K. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In Niederhoffer, K., Hollingshead, K., Resnik, P., Resnik, R. & Loveys, K. (eds.) *CLPsych Workshop, ACL* (Minneapolis, Minnesota, 2019).
44. Cao, J. *et al.* Observing dialogue in therapy: Categorizing and forecasting behavioral codes. *arXiv preprint arXiv:1907.00326* (2019).
45. Pérez-Rosas, V., Wu, X., Resnicow, K. & Mihalcea, R. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In Korhonen, A., Traum, D. & Màrquez, L. (eds.) *ACL* (2019).
46. Bordin, E. S. The generalizability of the psychoanalytic concept of the working alliance. *Psychother. Theory, research & practice* (1979).
47. Horvath, A. O. & Greenberg, L. S. Development and validation of the working alliance inventory. *J. counseling psychology* (1989).
48. Boyd, R. L., Ashokkumar, A., Seraj, S. & Pennebaker, J. W. The development and psychometric properties of liwc-22. *Austin, TX: Univ. Tex. at Austin* (2022).
49. Kanter, J. W. *et al.* What is behavioral activation?: A review of the empirical literature. *Clin. psychology review* **30**, 608–620 (2010).
50. Bai, Y. *et al.* Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).
51. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
52. Sharma, M. *et al.* Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548* (2023).
53. Hill, C. E. *Helping skills: Facilitating, exploration, insight, and action* (American Psychological Association, 2009).
54. Bhattacharjee, D. *et al.* Psychoeducation: A measure to strengthen psychiatric treatment. *Delhi Psychiatry J.* **14**, 33–39 (2011).
55. Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B. & Yang, Q. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *CHI* (2023).
56. Horvath, A. O. The alliance. *Psychother. Theory, research, practice, training* (2001).
57. Fairburn, C. G. & Cooper, Z. Therapist competence, therapy quality, and therapist training. *Behav. research therapy* (2011).
58. Lambert, M. J. *Bergin and Garfield's handbook of psychotherapy and behavior change* (John Wiley & Sons, 2013).
59. Sorensen, T. *et al.* A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070* (2024).

60. Chakraborty, S. *et al.* Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *ArXiv abs/2402.08925* (2024).
61. Hsu, S.-L. *et al.* Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982* (2023).
62. Xu, X. *et al.* Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385* (2023).
63. Pendse, S. R. *et al.* From treatment to healing: envisioning a decolonial digital mental health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–23 (2022).
64. Huang, L. *et al.* A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232* (2023).
65. Duffourc, M. & Gerke, S. Generative ai in health care and liability risks for physicians and safety concerns for patients. *Jama* (2023).
66. De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z. & Puntoni, S. Chatbots and mental health: Insights into the safety of generative ai. *J. Consumer Psychol.* (2022).
67. Weidinger, L. *et al.* Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).
68. Chung, N. C., Dyer, G. & Brocki, L. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857* (2023).
69. Barlow, D. H., Levitt, J. T. & Bufka, L. F. The dissemination of empirically supported treatments: a view to the future. *Behav. Res. Ther.* (1999).
70. Waltz, J., Addis, M. E., Koerner, K. & Jacobson, N. S. Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. *J. consulting clinical psychology* (1993).
71. Malhotra, G., Waheed, A., Srivastava, A., Akhtar, M. S. & Chakraborty, T. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *WSDM* (2022).
72. Ziems, C. *et al.* Can large language models transform computational social science? *Comput. Linguist.* (2023).
73. Arnold, K. Behind the mirror: Reflective listening and its tain in the work of carl rogers. *The Humanist. Psychol.* **42**, 354–369 (2014).
74. Braillon, A. & Taiebi, F. Practicing “reflective listening” is a mandatory prerequisite for empathy. *Patient education counseling* **103**, 1866–1867 (2020).
75. Rautalinko, E., Lisper, H.-O. & Ekehammar, B. Reflective listening in counseling: effects of training time and evaluator social skills. *Am. journal psychotherapy* **61**, 191–209 (2007).
76. Lundh, L.-G. Three modes of psychotherapy and their requisite core skills. *Couns. Psychother. Res.* **19**, 399–408 (2019).
77. Miller, W. R., Moyers, T. B., Ernst, D. & Amrhein, P. Manual for the motivational interviewing skill code (misc). *Unpubl. manuscript. Albuquerque: Cent. on Alcohol. Subst. Abus. Addict. Univ. New Mexico* (2003).
78. Robert, E., Bohart, A. C., Watson, J. & Greenberg, L. Empathy. *Psychotherapy* (2011).
79. Nelson-Jones, R. Practical counselling and helping skills: text and activities for the lifeskills counselling model. *Pract. Couns. Help. Ski.* 1–528 (2013).

80. Solomonov, N., McCarthy, K. S., Gorman, B. S. & Barber, J. P. The multitheoretical list of therapeutic interventions—30 items (multi-30). *Psychother. Res.* **29**, 565–580 (2019).
81. Fawcett, S. B. & Borck-Jameson, L. *Learning counseling and problem-solving skills* (Routledge, 2014).
82. Carey, T. A. *et al.* Psychological change from the inside looking out: A qualitative investigation. *Couns. Psychother. Res.* **7**, 178–187 (2007).
83. Tulver, K., Kaup, K. K., Laukkonen, R. & Aru, J. Restructuring insight: An integrative review of insight in problem-solving, meditation, psychotherapy, delusions and psychedelics. *Conscious. cognition* **110**, 103494 (2023).

Supplementary Materials

List of supplementary materials

Table S1-S29

Figures S1 to S2

Table S1. Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset⁴⁵. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS by 40.0% while human therapists responds by 23.4% on average ($P = 8.0 \times 10^{-73}$, Cohen’s $d = 0.4$, two-sided Student’s t -test).

Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ($n = 4893$)			GPT-3.5 ($n = 4746$)			Llama2-70b ($n = 4970$)			Llama2-13b ($n = 4820$)			Average ($n = 5446$)	High-quality ($n = 3907$)	Low-quality ($n = 1539$)
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	freq.	freq.
Reflections on Needs	40.0 _{49.0}	8.0×10^{-73}	0.4	32.3 _{46.8}	2.3×10^{-22}	0.2	36.0 _{48.0}	6.7×10^{-44}	0.3	34.5 _{47.5}	4.6×10^{-34}	0.2	23.4 _{43.1}	26.2 _{44.0}	20.6 _{40.5}
Reflections on Emotions	13.8 _{34.5}	4.5×10^{-6}	0.1	19.4 _{39.6}	2.3×10^{-34}	0.2	14.9 _{35.6}	6.3×10^{-10}	0.1	14.4 _{35.1}	4.8×10^{-8}	0.1	10.5 _{32.8}	14.6 _{35.3}	6.3 _{24.3}
Reflections on Values	5.5 _{22.7}	1.0×10^0	0.0	6.2 _{24.1}	5.0×10^{-2}	0.1	4.8 _{21.4}	1.0×10^0	0.0	5.7 _{23.1}	8.6×10^{-1}	0.0	4.8 _{22.8}	6.3 _{24.4}	3.3 _{17.9}
Reflections on Consequences	23.4 _{42.4}	5.4×10^{-1}	0.0	14.5 _{35.2}	5.1×10^{-21}	-0.2	22.4 _{41.7}	1.0×10^0	0.0	24.0 _{42.7}	8.2×10^{-2}	0.1	21.8 _{40.1}	17.9 _{38.4}	25.7 _{43.7}
Reflections on Conflicts	5.9 _{23.5}	1.0×10^0	0.0	4.8 _{21.4}	1.5×10^{-3}	-0.1	8.1 _{27.3}	6.9×10^{-2}	0.1	7.8 _{26.8}	3.6×10^{-1}	0.0	6.6 _{26.1}	8.3 _{27.6}	5.0 _{21.8}
Reflections on Strengths	39.1 _{48.8}	1.9×10^{-281}	0.7	32.1 _{46.7}	3.8×10^{-176}	0.6	24.0 _{42.7}	8.0×10^{-85}	0.4	21.9 _{41.3}	7.0×10^{-64}	0.3	9.5 _{31.3}	12.8 _{33.5}	6.2 _{24.2}
Questions on Experiences	29.8 _{45.7}	2.1×10^{-49}	-0.3	29.9 _{45.8}	7.4×10^{-48}	-0.3	48.2 _{50.0}	2.4×10^{-4}	0.1	40.7 _{49.1}	1.2×10^{-2}	-0.1	44.0 _{49.9}	48.9 _{50.0}	39.1 _{48.8}
Questions on Perspectives	5.9 _{23.5}	2.9×10^{-7}	0.1	2.4 _{15.4}	4.9×10^{-2}	-0.1	8.4 _{27.7}	3.2×10^{-24}	0.2	7.3 _{26.0}	1.0×10^{-15}	0.2	3.5 _{19.9}	4.9 _{21.7}	2.0 _{14.1}
Questions on Emotions	2.2 _{14.7}	7.9×10^{-75}	-0.4	1.7 _{13.0}	8.4×10^{-83}	-0.4	8.9 _{28.5}	4.8×10^{-5}	-0.1	11.7 _{32.2}	1.0×10^0	0.0	11.7 _{33.0}	13.3 _{34.0}	10.1 _{30.2}
Problem-Solving	57.8 _{49.4}	3.5×10^{-256}	0.7	49.2 _{50.0}	7.2×10^{-137}	0.5	45.8 _{49.8}	6.1×10^{-104}	0.4	34.4 _{47.5}	1.3×10^{-20}	0.2	25.9 _{42.6}	21.2 _{40.9}	30.5 _{46.0}
Planning	57.1 _{49.5}	0.0×10^0	0.9	42.6 _{49.5}	1.2×10^{-182}	0.6	44.6 _{49.7}	1.6×10^{-213}	0.6	36.2 _{48.1}	7.1×10^{-109}	0.4	17.1 _{37.6}	16.9 _{37.5}	17.2 _{37.8}
Normalizing	45.5 _{49.8}	2.9×10^{-169}	0.6	36.9 _{48.3}	1.3×10^{-76}	0.4	19.3 _{39.5}	1.0×10^0	0.0	24.9 _{43.3}	8.8×10^{-7}	0.1	20.5 _{39.5}	17.8 _{38.3}	23.2 _{42.2}
Psychoeducation	16.4 _{37.0}	2.5×10^{-45}	0.3	24.5 _{43.0}	1.0×10^{-130}	0.5	12.6 _{33.2}	2.2×10^{-17}	0.2	8.9 _{28.4}	1.7×10^{-1}	0.0	7.6 _{24.8}	5.2 _{22.3}	9.9 _{29.8}

Table S2. Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset⁴⁵ for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS by 31.1% and GPT-4 with full conversation simulation responds by 42.5% when compared with the human therapists responds by 23.4% on average from Table S1. (Single: $P = 2.3 \times 10^{-6}$, Cohen’s $d = 0.2$; two-sided student t-test) (Full: $P = 6.4 \times 10^{-84}$, Cohen’s $d = 0.4$; two-sided student t-test)

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)														
	GPT-4 ($n = 3837$)			GPT-3.5 ($n = 3907$)			Llama2-70b ($n = 3906$)			Llama2-13b ($n = 3895$)			GPT-4 ($n = 1056$)			GPT-3.5 ($n = 839$)			Llama2-70b ($n = 1064$)			Llama2-13b ($n = 925$)					
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d
Reflections on Needs	31.1 _{46.3}	2.3×10^{-6}	0.2	35.4 _{47.8}	2.0×10^{-12}	0.3	20.9 _{40.7}	1.0×10^0	-0.1	19.9 _{39.9}	2.7×10^{-1}	-0.1	42.5 _{49.4}	6.4×10^{-84}	0.4	31.6 _{46.5}	1.6×10^{-17}	0.2	40.1 _{49.0}	2.2×10^{-66}	0.4	38.0 _{48.5}	4.7×10^{-51}	0.3	23.4 _{43.1}	8.0×10^{-73}	0.4
Reflections on Emotions	8.4 _{27.8}	7.7×10^{-1}	-0.1	8.5 _{27.8}	1.0×10^0	-0.1	9.4 _{29.2}	1.0×10^0	0.0	9.8 _{29.8}	1.0×10^0	0.0	15.3 _{36.0}	2.0×10^{-10}	0.1	21.8 _{41.3}	1.9×10^{-47}	0.3	16.4 _{37.0}	6.3×10^{-15}	0.2	15.5 _{36.2}	3.1×10^{-11}	0.1	10.5 _{32.8}	4.5×10^{-6}	0.1
Reflections on Values	4.5 _{20.6}	1.0×10^0	0.0	5.1 _{22.1}	1.0×10^0	0.0	3.1 _{17.3}	2.5×10^{-1}	-0.1	3.5 _{18.3}	1.0×10^0	-0.1	5.7 _{23.3}	8.1×10^{-1}	0.0	6.4 _{24.5}	1.9×10^{-2}	0.1	5.3 _{22.4}	1.0×10^0	0.0	6.2 _{24.1}	7.3×10^{-2}	0.1	4.8 _{22.8}	4.8×10^{-1}	0.0
Reflections on Values	4.5 _{20.6}	6.1×10^{-1}	-0.0	5.1 _{22.1}	7.3×10^{-1}	0.0	3.1 _{17.3}	1.9×10^{-2}	-0.1	3.5 _{18.3}	8.2×10^{-2}	-0.1	5.7 _{23.3}	6.2×10^{-2}	0.0	6.4 _{24.5}	1.5×10^{-3}	0.1	5.3 _{22.4}	3.5×10^{-1}	0.0	6.2 _{24.1}	5.6×10^{-3}	0.1	4.8 _{22.8}	4.8×10^{-1}	0.0
Reflections on Conflicts	3.0 _{17.2}	2.0×10^{-4}	-0.1	2.0 _{14.1}	7.1×10^{-6}	-0.2	3.2 _{17.6}	4.7×10^{-4}	-0.1	3.6 _{18.6}	7.5×10^{-3}	-0.1	6.7 _{25.0}	1.0×10^0	0.0	5.4 _{22.6}	2.1×10^{-1}	-0.1	9.4 _{29.3}	1.6×10^{-5}	0.1	8.8 _{28.3}	1.9×10^{-3}	0.1	6.6 _{26.1}	8.3×10^{-2}	0.1
Reflections on Strengths	68.3 _{46.6}	0.0×10^0	1.7	54.6 _{49.8}	4.0×10^{-249}	1.3	35.4 _{47.9}	9.5×10^{-106}	0.7	37.7 _{48.5}	2.4×10^{-112}	0.8	31.1 _{46.3}	3.0×10^{-151}	0.6	27.3 _{44.5}	4.8×10^{-109}	0.5	20.9 _{47.7}	9.2×10^{-51}	0.3	18.1 _{38.5}	6.1×10^{-31}	0.2	9.5 _{31.3}	$12.8 \times 10^{-33.5}$	0.6
Questions on Experiences	29.4 _{45.6}	1.4×10^{-17}	-0.3	20.9 _{40.7}	6.0×10^{-36}	-0.5	32.7 _{46.9}	1.3×10^{-10}	-0.2	35.5 _{47.9}	1.7×10^{-5}	-0.2	29.9 _{45.8}	4.5×10^{-42}	-0.3	31.8 _{46.6}	1.5×10^{-31}	-0.3	52.4 _{49.9}	1.3×10^{-14}	0.2	42.0 _{49.1}	7.3×10^{-11}	0.0	44.0 _{49.9}	$48.9 \times 10^{-50.0}$	0.0
Questions on Perspectives	3.2 _{17.7}	1.0×10^0	0.0	2.0 _{14.1}	5.4×10^{-1}	-0.1	6.9 _{25.3}	1.7×10^{-5}	0.2	5.3 _{22.4}	1.5×10^{-1}	0.1	6.6 _{24.8}	2.7×10^{-10}	0.1	2.5 _{15.7}	1.8×10^{-1}	-0.1	8.8 _{28.3}	3.2×10^{-25}	0.2	7.7 _{26.7}	1.7×10^{-17}	0.2	11.7 _{32.2}	1.0×10^0	0.0
Questions on Emotions	1.3 _{11.4}	8.1×10^{-23}	-0.3	2.0 _{14.1}	7.1×10^{-16}	-0.3	3.7 _{18.8}	1.6×10^{-13}	-0.3	5.4 _{22.6}	2.8×10^{-7}	-0.2	2.4 _{15.5}	3.6×10^{-57}	-0.3	1.7 _{12.8}	5.4×10^{-71}	-0.4	10.3 _{30.5}	5.1×10^{-1}	0.0	13.2 _{33.9}	3.8×10^{-1}	0.0	11.7 _{32.2}	1.0×10^0	0.0
Problem-Solving	44.2 _{49.7}	4.2×10^{-34}	0.4	55.2 _{49.8}	1.9×10^{-70}	0.7	54.1 _{49.9}	4.9×10^{-79}	0.6	34.7 _{47.6}	1.3×10^{-7}	0.2	61.5 _{49.7}	3.2×10^{-284}	0.8	47.9 _{50.0}	7.5×10^{-112}	0.5	43.5 _{49.6}	8.9×10^{-74}	0.4	34.3 _{47.5}	2.8×10^{-18}	0.2	25.9 _{42.6}	$21.2 \times 10^{-40.9}$	0.7
Planning	66.8 _{47.1}	4.7×10^{-279}	1.3	70.2 _{45.8}	2.4×10^{-269}	1.4	64.1 _{48.0}	7.9×10^{-252}	1.2	44.8 _{49.8}	1.8×10^{-82}	0.7	54.4 _{49.8}	0.0×10^0	0.9	36.7 _{48.2}	6.2×10^{-104}	0.5	39.3 _{48.8}	1.4×10^{-130}	0.5	34.1 _{47.4}	6.5×10^{-81}	0.4	17.1 _{37.6}	$16.9 \times 10^{-37.5}$	0.4
Normalizing	65.7 _{47.5}	2.5×10^{-218}	1.1	70.8 _{45.5}	9.2×10^{-227}	1.2	22.5 _{41.8}	1.0×10^0	0.0	28.6 _{45.2}	1.9×10^{-7}	0.2	39.9 _{49.0}	5.4×10^{-95}	0.4	29.6 _{45.6}	1.7×10^{-23}	0.2	18.4 _{38.8}	1.5×10^{-1}	-0.1	20.4 _{42.7}	5.2×10^{-4}	0.1	20.5 _{39.5}	$17.8 \times 10^{-38.3}$	0.1
Psychoeducation	20.5 _{40.4}	7.3×10^{-42}	0.5	25.3 _{43.5}	1.0×10^{-62}	0.6	26.9 _{44.4}	6.0×10^{-85}	0.7	18.1 _{38.5}	4.2×10^{-26}	0.4	15.2 _{36.0}	7.7×10^{-33}	0.3	24.4 _{43.0}	1.5×10^{-121}	0.5	8.7 _{28.2}	5.5×10^{-1}	0.0	6.7 _{25.0}	1.0×10^0	0.0	7.6 _{24.8}	$5.2 \times 10^{-22.3}$	0.0

Table S3. Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset⁴⁵. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS at 3.1 turns earlier than average human therapists responds. (GPT-4: 3.1 turn, Average human therapists: 6.2 turn; $P = 4.3 \times 10^{-15}$, Cohen’s $d = 0.7$, two-sided Student’s t -test).

Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ($n = 4893$)			GPT-3.5 ($n = 4746$)			Llama2-70b ($n = 4970$)			Llama2-13b ($n = 4820$)			Average ($n = 5446$)	High-quality ($n = 3907$)	Low-quality ($n = 1539$)
	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	turn	turn
Reflections on Needs	3.1 _{2,9}	4.3×10^{-15}	0.7	6.4 _{5,4}	1.0×10^0	0.0	5.1 _{5,0}	2.5×10^{-1}	0.2	5.3 _{5,5}	9.1×10^{-1}	0.2	6.2 _{5,5}	4.4 _{4,5}	8.1 _{6,2}
Reflections on Emotions	6.1 _{5,0}	1.0×10^0	0.0	5.7 _{4,8}	1.0×10^0	0.1	7.2 _{5,6}	1.0×10^0	-0.2	6.1 _{5,4}	1.0×10^0	0.0	6.3 _{4,9}	5.2 _{4,8}	7.4 _{5,0}
Reflections on Values	8.3 _{5,1}	3.6×10^{-1}	0.3	9.1 _{5,5}	1.0×10^0	0.1	10.3 _{5,7}	1.0×10^0	-0.1	10.1 _{5,6}	1.0×10^0	0.0	9.9 _{5,5}	8.8 _{5,3}	11.0 _{5,8}
Reflections on Consequences	6.5 _{4,2}	1.0×10^0	0.1	7.6 _{4,9}	1.0×10^0	-0.1	6.5 _{4,5}	1.0×10^0	0.1	7.0 _{5,4}	1.0×10^0	0.0	7.0 _{5,0}	6.2 _{4,6}	7.8 _{5,4}
Reflections on Conflicts	7.2 _{4,6}	3.9×10^{-4}	0.6	8.1 _{4,6}	1.3×10^{-1}	0.4	7.8 _{4,9}	7.9×10^{-3}	0.4	7.8 _{5,3}	9.1×10^{-3}	0.4	10.0 _{5,2}	8.2 _{4,7}	11.8 _{5,6}
Reflections on Strengths	7.9 _{5,4}	5.9×10^{-4}	0.4	9.1 _{6,3}	1.0×10^0	0.2	8.7 _{6,3}	2.1×10^{-1}	0.2	9.4 _{6,5}	1.0×10^0	0.1	10.2 _{6,1}	8.4 _{6,0}	11.9 _{5,7}
Questions on Experiences	2.7 _{2,5}	1.0×10^0	0.1	3.7 _{4,3}	1.5×10^{-1}	-0.2	2.9 _{3,0}	1.0×10^0	0.0	3.8 _{4,0}	2.8×10^{-2}	-0.3	2.9 _{2,5}	2.2 _{2,1}	3.6 _{2,9}
Questions on Perspectives	8.5 _{4,7}	3.1×10^{-7}	0.8	9.7 _{5,3}	1.6×10^{-2}	0.6	8.7 _{5,7}	4.2×10^{-6}	0.7	8.7 _{5,4}	2.8×10^{-6}	0.7	12.6 _{5,1}	10.0 _{4,7}	15.3 _{4,2}
Questions on Emotions	6.2 _{4,9}	1.0×10^0	0.1	6.0 _{5,9}	1.0×10^0	0.2	5.3 _{5,6}	1.6×10^{-1}	0.2	5.4 _{5,5}	1.6×10^{-1}	0.3	7.0 _{6,4}	6.7 _{6,3}	7.3 _{6,5}
Problem-Solving	7.8 _{5,2}	1.0×10^0	0.1	7.1 _{5,1}	1.5×10^{-1}	0.2	7.5 _{4,8}	7.1×10^{-1}	0.2	9.0 _{5,8}	1.0×10^0	-0.1	8.3 _{5,4}	8.7 _{5,6}	8.0 _{5,1}
Planning	9.3 _{5,4}	4.2×10^{-5}	0.4	10.2 _{6,1}	3.8×10^{-2}	0.3	8.7 _{5,3}	7.2×10^{-8}	0.5	10.7 _{6,1}	6.0×10^{-1}	0.2	11.9 _{6,5}	11.5 _{6,8}	12.4 _{6,0}
Normalizing	7.9 _{6,0}	9.0×10^{-1}	-0.2	7.0 _{6,0}	1.0×10^0	0.0	9.3 _{6,8}	2.8×10^{-3}	-0.4	9.0 _{6,3}	5.3×10^{-3}	-0.3	6.9 _{6,2}	6.2 _{6,3}	7.7 _{6,0}
Psychoeducation	11.1 _{5,5}	2.5×10^{-2}	-0.4	8.2 _{5,7}	1.0×10^0	0.1	11.8 _{5,7}	2.3×10^{-4}	-0.5	11.5 _{5,8}	5.6×10^{-3}	-0.5	8.9 _{5,5}	8.5 _{5,6}	9.2 _{5,5}

Table S4. Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset⁷¹ for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS at 4.3 turn which is 1.9 turns earlier than average human therapists. Similarly, GPT-4 with full conversation simulation responds at 2.0 turn which is 4.2 turns earlier than average human therapists. (Single: 4.3 turn, Average human therapists: 6.2 turn from Table S3; $P = 1.1 \times 10^{-3}$, Cohen’s $d = 0.4$, two-sided Student’s t -test) (Full: 2.0 turn, Average human therapists: 6.2 turn from Table S3; $P = 9.2 \times 10^{-16}$, Cohen’s $d = 0.7$).

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)														
	GPT-4 ($n = 3837$)			GPT-3.5 ($n = 3907$)			Llama2-70b ($n = 3906$)			Llama2-13b ($n = 3895$)			GPT-4 ($n = 1056$)			GPT-3.5 ($n = 839$)			Llama2-70b ($n = 1064$)			Llama2-13b ($n = 925$)					
	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d
Reflections on Needs	4.3 _{2,6}	1.1×10^{-3}	0.4	8.2 _{6,2}	1.7×10^{-2}	-0.3	8.3 _{5,1}	1.4×10^{-2}	-0.4	9.6 _{5,9}	6.5×10^{-6}	-0.6	2.0 _{2,8}	9.2×10^{-16}	0.9	4.7 _{3,9}	4.5×10^{-2}	0.3	2.9 _{3,4}	1.0×10^{-9}	0.7	2.4 _{2,5}	6.7×10^{-14}	0.8	4.8 _{5,0}	1.4×10^{-1}	0.3
Reflections on Emotions	8.8 _{5,1}	9.3×10^{-3}	-0.5	7.8 _{5,3}	8.1×10^{-1}	-0.3	9.3 _{5,2}	3.1×10^{-4}	-0.6	8.8 _{5,3}	8.7×10^{-3}	-0.5	4.6 _{4,2}	3.3×10^{-2}	0.4	4.7 _{4,2}	7.1×10^{-2}	0.3	6.0 _{5,4}	1.0×10^0	0.1	4.8 _{5,0}	1.4×10^{-1}	0.3	9.3 _{5,5}	1.0×10^0	0.1
Reflections on Values	8.2 _{4,1}	1.0×10^0	0.3	9.5 _{6,7}	1.0×10^0	0.1	13.6 _{6,1}	2.7×10^{-2}	-0.7	13.1 _{3,3}	1.6×10^{-1}	-0.6	8.4 _{5,6}	9.3×10^{-1}	0.3	8.9 _{5,0}	1.0×10^0	0.2	9.2 _{5,5}	1.0×10^0	0.1	9.3 _{5,5}	1.0×10^0	0.1	9.3 _{5,5}	1.0×10^0	0.1
Reflections on Consequences	7.5 _{3,4}	1.0×10^0	-0.1	9.1 _{4,7}	3.7×10^{-2}	-0.4	9.0 _{4,6}	1.1×10^{-1}	-0.4	11.3 _{5,3}	6.5×10^{-7}	-0.8	6.0 _{4,6}	8.3×10^{-1}	0.2	6.7 _{4,8}	1.0×10^0	0.1	5.7 _{4,2}	1.0×10^{-1}	0.3	5.5 _{4,6}	4.5×10^{-2}	0.3	7.1 _{5,0}	2.1×10^{-4}	0.6
Reflections on Conflicts	7.1 _{3,5}	1.8×10^{-1}	0.6	6.9 _{2,3}	3.4×10^{-1}	0.6	8.0 _{3,2}	1.0×10^0	0.4	11.0 _{3,5}	1.0×10^0	-0.2	7.2 _{4,9}	2.7×10^{-3}	0.6	8.4 _{5,1}	6.7×10^{-1}	0.3	7.8 _{5,3}	1.7×10^{-2}	0.4	7.1 _{5,0}	1.1×10^{-4}	0.5	9.0 _{5,2}	1.0×10^0	0.2
Reflections on Strengths	9.0 _{5,2}	1.0×10^0	0.2	11.4 _{6,1}	1.0×10^0	-0.2	11.2 _{6,1}	1.0×10^0	-0.2	12.2 _{6,6}	1.1×10^{-1}	-0.3	6.8 _{5,4}	5.4×10^{-6}	0.6	7.0 _{5,7}	7.2×10^{-5}	0.5	6.1 _{5,6}	1.1×10^{-7}	0.7	7.1 _{5,6}	1.1×10^{-4}	0.5	9.0 _{5,2}	1.4×10^{-4}	0.5
Questions on Experiences	3.8 _{2,1}	5.7×10^{-3}	-0.4	5.9 _{5,2}	8.3×10^{-12}	-0.9	4.5 _{3,5}	5.4×10^{-6}	-0.5	6.1 _{4,5}	6.0×10^{-17}	-0.9	1.7 _{2,3}	3.1×10^{-5}	0.5	2.0 _{2,5}	1.6×10^{-2}	0.3	1.4 _{1,2}	4.5×10^{-10}	0.7	1.6 _{1,3}	3.5×10^{-8}	0.6	10.1 _{5,4}	1.0×10^{-1}	0.5
Questions on Perspectives	8.0 _{5,1}	3.6×10^{-3}	0.9	6.3 _{2,0}	4.1×10^{-2}	1.3	11.2 _{4,5}	1.0×10^0	0.3	10.4 _{4,7}	4.4×10^{-1}	0.5	8.6 _{4,6}	3.9×10^{-6}	0.8	10.1 _{3,4}	1.0×10^{-1}	0.5	7.6 _{6,8}	3.4×10^{-8}	0.9	8.2 _{5,5}	7.4×10^{-7}	0.8	9.0 _{5,2}	1.0×10^0	0.2
Questions on Emotions	4.9 _{3,1}	1.0×10^0	0.3	7.0 _{5,9}	1.0×10^0	0.0	6.9 _{6,0}	1.0×10^0	0.0	5.7 _{4,2}	1.0×10^0	0.2	6.4 _{5,1}	1.0×10^0	0.1	5.8 _{6,0}	1.0×10^0	0.2	4.9 _{5,5}	4.6×10^{-2}	0.4	5.3 _{5,9}	2.3×10^{-1}	0.3	10.3 _{4,9}	3.9×10^{-3}	-0.4
Problem-Solving	11.4 _{5,1}	1.0×10^0	0.1	12.2 _{5,4}	1.0×10^0	0.0	9.8 _{5,1}	1.6×10^{-2}	0.4	12.7 _{6,5}	1.0×10^0	-0.1	7.5 _{5,0}	5.0×10^{-10}	0.7	8.4 _{6,2}	1.2×10^{-5}	0.5	7.7 _{5,3}	5.0×10^{-9}	0.7	9.0 _{5,2}	1.4×10^{-4}	0.5	11.2 _{5,4}	1.0×10^0	0.0
Normalizing	10.3 _{6,2}	1.6×10^{-3}	-0.5	9.6 _{6,1}	2.4×10^{-3}	-0.4	12.4 _{6,7}	2.0×10^{-10}	-0.9	11.2 _{6,3}	1.3×10^{-6}	-0.7	5.7 _{4,9}	5.5×10^{-1}	0.2	4.7 _{5,0}	6.1×10^{-3}	0.4	6.7 _{5,7}	1.0×10^0	0.0	7.6 _{6,0}	1.0×10^0	-0.1	10.3 _{4,9}	3.9×10^{-3}	-0.4
Psychoeducation	12.0 _{5,1}	7.3×10^{-3}	-0.6	12.2 _{5,4}	3.8×10^{-3}	-0.6	13.2 _{5,3}	3.3×10^{-7}	-0.8	13.5 _{5,9}	1.1×10^{-5}	-0.8	10.4 _{5,7}	7.0×10^{-1}	-0.3	6.2 _{4,7}	1.7×10^{-3}	0.5	10.1 _{5,8}	1.0×10^0	-0.2	9.6 _{5,1}	1.0×10^0	-0.1	12.0 _{5,1}	7.3×10^{-3}	-0.6

Table S5. Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset⁴⁵. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13 \times 6 = 78$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON CONSEQUENCES by 19.9% and human therapists responds by 27.3% on average when client expresses CHANGING UNHEALTHY BEHAVIOR. ($P = 4.1 \times 10^{-6}$, Cohen’s $d = -0.2$, two-sided Student’s t -test).

Therapist Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			Average	High-quality	Low-quality
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	freq.	freq.
Client behavior: Changing unhealthy behavior															
n	2284			2148			2211			2082			1553	1198	355
Reflections on Needs	29.3 _{45.5}	1.0×10^0	0.0	29.1 _{45.4}	1.0×10^0	0.0	26.9 _{44.4}	1.0×10^0	0.0	27.8 _{44.8}	1.0×10^0	0.0	27.7 _{45.5}	30.6 _{46.1}	24.8 _{43.2}
Reflections on Emotions	8.2 _{27.4}	1.0×10^0	-0.1	11.8 _{32.2}	1.0×10^0	0.0	8.7 _{28.2}	1.0×10^0	-0.1	8.6 _{28.0}	1.0×10^0	-0.1	10.4 _{33.3}	14.6 _{35.3}	6.2 _{24.1}
Reflections on Values	4.1 _{19.9}	1.0×10^0	0.0	4.9 _{21.7}	1.0×10^0	0.0	3.6 _{18.6}	1.0×10^0	-0.1	4.5 _{20.8}	1.0×10^0	0.0	4.8 _{23.7}	7.0 _{25.5}	2.5 _{15.7}
Reflections on Consequences	19.9 _{39.9}	4.1×10^{-6}	-0.2	12.3 _{32.9}	4.9×10^{-31}	-0.4	16.6 _{37.2}	3.1×10^{-14}	-0.3	19.6 _{39.7}	1.8×10^{-6}	-0.2	27.3 _{42.3}	19.9 _{39.9}	34.6 _{47.7}
Reflections on Conflicts	4.1 _{19.9}	1.3×10^{-2}	-0.1	3.3 _{17.9}	5.2×10^{-5}	-0.2	4.4 _{20.5}	6.7×10^{-2}	-0.1	6.1 _{24.0}	1.0×10^0	0.0	6.9 _{26.6}	8.3 _{27.5}	5.6 _{23.1}
Reflections on Strengths	44.4 _{49.7}	1.0×10^{-121}	0.8	35.6 _{47.9}	5.5×10^{-73}	0.6	25.9 _{43.8}	2.7×10^{-34}	0.4	22.6 _{41.8}	4.7×10^{-23}	0.3	9.3 _{32.4}	14.1 _{34.8}	4.5 _{20.8}
Questions on Experiences	15.2 _{35.9}	2.6×10^{-47}	-0.5	16.3 _{36.9}	5.1×10^{-40}	-0.5	27.7 _{44.8}	2.0×10^{-5}	-0.2	24.3 _{42.9}	8.5×10^{-12}	-0.3	35.6 _{48.6}	40.2 _{49.1}	31.0 _{46.3}
Questions on Perspectives	2.8 _{16.4}	1.0×10^0	0.0	0.8 _{9.1}	3.2×10^{-6}	-0.2	3.9 _{19.3}	1.0×10^0	0.0	4.8 _{21.4}	1.0×10^0	0.1	3.5 _{20.2}	4.8 _{21.5}	2.3 _{14.9}
Questions on Emotions	0.7 _{8.1}	2.6×10^{-37}	-0.4	0.3 _{5.3}	2.9×10^{-40}	-0.5	3.2 _{17.5}	4.6×10^{-14}	-0.3	4.2 _{20.0}	1.3×10^{-8}	-0.2	9.6 _{30.7}	11.3 _{31.6}	7.9 _{27.0}
Problem-Solving	40.7 _{49.1}	8.0×10^{-14}	0.3	38.1 _{48.6}	1.9×10^{-8}	0.2	37.3 _{48.4}	4.5×10^{-7}	0.2	29.4 _{45.6}	1.0×10^0	0.0	28.4 _{42.7}	20.4 _{40.3}	36.3 _{48.2}
Planning	32.4 _{46.8}	1.2×10^{-39}	0.4	32.4 _{46.8}	2.4×10^{-39}	0.5	32.4 _{46.8}	1.4×10^{-39}	0.5	25.9 _{43.8}	2.6×10^{-18}	0.3	13.7 _{32.6}	10.7 _{30.9}	16.6 _{37.3}
Normalizing	28.0 _{44.9}	1.1×10^{-23}	0.3	23.4 _{42.3}	2.7×10^{-11}	0.2	9.9 _{29.9}	5.6×10^{-3}	-0.1	13.8 _{34.5}	1.0×10^0	0.0	14.0 _{32.6}	10.5 _{30.7}	17.5 _{38.0}
Psychoeducation	6.1 _{24.0}	1.0×10^0	0.0	12.0 _{32.5}	1.7×10^{-10}	0.2	8.5 _{27.8}	1.9×10^{-2}	0.1	5.8 _{23.4}	1.0×10^0	0.0	5.4 _{20.5}	3.5 _{18.4}	7.3 _{26.1}
Client behavior: Sustaining unhealthy behavior															
n	436			419			553			445			389	275	114
Reflections on Needs	36.5 _{48.2}	5.1×10^{-1}	0.2	35.8 _{48.0}	9.7×10^{-1}	0.2	31.5 _{46.5}	1.0×10^0	0.1	30.1 _{45.9}	1.0×10^0	0.1	27.6 _{45.3}	30.5 _{46.1}	24.6 _{43.2}
Reflections on Emotions	9.2 _{28.9}	1.0×10^0	-0.1	14.8 _{35.5}	1.0×10^0	0.1	11.0 _{31.4}	1.0×10^0	0.0	14.2 _{34.9}	1.0×10^0	0.1	11.7 _{33.2}	13.8 _{34.6}	9.6 _{29.7}
Reflections on Values	4.1 _{19.9}	1.0×10^0	-0.1	3.6 _{18.6}	1.0×10^0	-0.1	4.7 _{21.2}	1.0×10^0	-0.1	3.8 _{19.2}	1.0×10^0	-0.1	6.8 _{25.0}	6.5 _{24.8}	7.0 _{25.7}
Reflections on Consequences	35.1 _{47.8}	1.0×10^0	0.0	27.4 _{44.7}	5.8×10^{-1}	-0.2	26.6 _{44.2}	1.2×10^{-1}	-0.2	31.5 _{46.5}	1.0×10^0	-0.1	36.1 _{46.8}	26.5 _{44.2}	45.6 _{50.0}
Reflections on Conflicts	11.2 _{31.6}	1.0×10^0	0.1	6.9 _{25.4}	1.0×10^0	0.0	6.7 _{25.0}	1.0×10^0	-0.1	8.1 _{27.3}	1.0×10^0	0.0	8.2 _{27.1}	7.0 _{26.6}	8.8 _{28.4}
Reflections on Strengths	11.7 _{32.2}	8.8×10^{-2}	0.2	8.1 _{27.3}	1.0×10^0	0.1	7.6 _{26.5}	1.0×10^0	0.1	4.7 _{21.2}	1.0×10^0	0.0	5.2 _{23.6}	6.9 _{25.4}	3.5 _{18.5}
Questions on Experiences	19.5 _{39.7}	2.3×10^{-8}	-0.4	20.5 _{40.4}	3.5×10^{-7}	-0.4	29.7 _{45.7}	1.6×10^{-1}	-0.2	30.6 _{46.1}	6.5×10^{-1}	-0.2	39.3 _{49.4}	46.2 _{49.9}	32.5 _{47.0}
Questions on Perspectives	7.6 _{26.5}	9.9×10^{-1}	0.2	2.6 _{16.0}	1.0×10^0	0.0	6.9 _{25.3}	1.0×10^0	0.1	6.1 _{23.9}	1.0×10^0	0.1	3.5 _{19.3}	4.4 _{20.5}	2.6 _{16.1}
Questions on Emotions	0.5 _{6.8}	2.5×10^{-7}	-0.4	0.5 _{6.9}	5.5×10^{-7}	-0.4	2.5 _{15.7}	6.7×10^{-4}	-0.3	5.8 _{23.5}	1.0×10^0	-0.1	9.2 _{29.7}	10.5 _{30.8}	7.9 _{27.1}
Problem-Solving	31.4 _{46.5}	8.2×10^{-1}	0.2	27.9 _{44.9}	1.0×10^0	0.1	34.0 _{47.4}	3.5×10^{-2}	0.2	21.8 _{41.3}	1.0×10^0	0.0	23.6 _{40.3}	15.6 _{36.4}	31.6 _{46.7}
Planning	12.8 _{33.5}	4.7×10^{-2}	0.2	9.3 _{29.1}	1.0×10^0	0.1	19.9 _{40.0}	5.1×10^{-8}	0.4	10.1 _{30.2}	1.0×10^0	0.2	5.9 _{22.1}	4.0 _{19.6}	7.9 _{27.1}
Normalizing	14.7 _{35.4}	1.0×10^0	0.0	15.5 _{36.2}	1.0×10^0	0.0	6.9 _{25.3}	1.4×10^{-2}	-0.2	7.4 _{26.2}	1.1×10^{-1}	-0.2	14.0 _{33.2}	10.5 _{30.8}	17.5 _{38.2}
Psychoeducation	2.1 _{14.2}	1.3×10^{-5}	-0.4	11.9 _{32.5}	1.0×10^0	0.1	5.2 _{22.3}	3.4×10^{-1}	-0.2	2.9 _{16.9}	5.6×10^{-4}	-0.3	9.8 _{26.7}	4.7 _{21.3}	14.9 _{35.8}
Client behavior: Sharing positive emotions															
n	658			554			408			534			294	226	68
Reflections on Needs	12.9 _{33.6}	1.0×10^0	-0.1	17.9 _{38.3}	1.0×10^0	0.0	20.6 _{40.5}	1.0×10^0	0.1	16.7 _{37.3}	1.0×10^0	0.0	17.3 _{40.9}	24.3 _{43.0}	10.3 _{30.6}
Reflections on Emotions	5.3 _{22.5}	1.0×10^0	-0.1	8.1 _{27.3}	1.0×10^0	0.0	5.6 _{23.1}	1.0×10^0	-0.1	5.8 _{23.4}	1.0×10^0	-0.1	7.5 _{32.0}	15.0 _{35.8}	0.0 _{0.0}
Reflections on Values	2.9 _{16.8}	1.0×10^0	0.0	4.0 _{19.5}	1.0×10^0	0.0	5.1 _{22.1}	1.0×10^0	0.1	6.4 _{24.4}	1.0×10^0	0.1	3.4 _{20.6}	5.3 _{22.5}	1.5 _{12.1}
Reflections on Consequences	6.2 _{24.2}	6.6×10^{-5}	-0.3	5.2 _{22.3}	5.1×10^{-6}	-0.4	14.2 _{35.0}	1.0×10^0	0.0	7.9 _{26.9}	1.8×10^{-2}	-0.3	15.8 _{34.0}	11.1 _{31.4}	20.6 _{40.7}
Reflections on Conflicts	1.8 _{13.4}	1.0×10^0	-0.1	2.0 _{14.0}	1.0×10^0	-0.1	4.9 _{21.6}	1.0×10^0	0.0	1.9 _{13.6}	1.0×10^0	-0.1	3.4 _{20.6}	5.3 _{22.5}	1.5 _{12.1}
Reflections on Strengths	66.1 _{47.4}	3.7×10^{-62}	1.3	52.3 _{50.0}	2.7×10^{-33}	0.9	29.2 _{45.5}	1.9×10^{-7}	0.5	39.3 _{48.9}	5.8×10^{-17}	0.7	10.2 _{34.0}	15.9 _{36.7}	4.4 _{20.7}
Questions on Experiences	7.9 _{27.0}	1.5×10^{-19}	-0.7	10.6 _{30.9}	2.2×10^{-12}	-0.6	24.3 _{42.9}	1.0×10^0	-0.2	15.7 _{36.4}	8.3×10^{-6}	-0.4	31.6 _{47.6}	36.7 _{48.3}	26.5 _{44.4}
Questions on Perspectives	2.1 _{14.4}	1.0×10^0	0.1	0.7 _{8.5}	1.0×10^0	-0.1	3.9 _{19.4}	1.0×10^0	0.1	3.4 _{18.1}	1.0×10^0	0.1	1.3 _{14.2}	2.7 _{16.1}	0.0 _{0.0}
Questions on Emotions	0.5 _{6.7}	3.6×10^{-7}	-0.4	0.5 _{7.3}	8.4×10^{-6}	-0.4	4.9 _{21.6}	1.0×10^0	-0.1	3.6 _{18.5}	1.0×10^0	-0.2	7.5 _{28.9}	10.6 _{30.9}	4.4 _{20.7}
Problem-Solving	19.8 _{39.8}	9.3×10^{-1}	0.2	23.1 _{42.2}	3.3×10^{-2}	0.3	26.0 _{43.9}	2.0×10^{-3}	0.3	14.2 _{35.0}	1.0×10^0	0.0	13.0 _{33.6}	12.8 _{33.5}	13.2 _{34.1}
Planning	47.9 _{50.0}	7.0×10^{-33}	0.9	39.4 _{48.9}	5.1×10^{-21}	0.7	32.6 _{46.9}	3.4×10^{-13}	0.6	25.1 _{43.4}	1.5×10^{-7}	0.4	8.2 _{26.9}	7.5 _{26.4}	8.8 _{28.6}
Normalizing	59.0 _{49.2}	7.1×10^{-25}	0.8	42.6 _{49.5}	4.0×10^{-7}	0.4	12.5 _{33.1}	2.0×10^{-2}	-0.3	29.0 _{45.4}	1.0×10^0	0.1	22.7 _{40.6}	19.0 _{39.3}	26.5 _{44.4}
Psychoeducation	4.3 _{20.2}	1.0×10^0	0.1	7.4 _{26.2}	7.5×10^{-1}	0.2	5.6 _{23.1}	1.0×10^0	0.1	4.7 _{21.1}	1.0×10^0	0.1	3.1 _{15.3}	1.8 _{13.2}	4.4 _{20.7}

Table S6. (Continue Table S5) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset⁴⁵. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13 \times 6 = 78$). d is the Cohen’s d when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON NEEDS by 38.9% and human therapists responds by 24.5% on average when client expresses SHARING NEGATIVE EMOTIONS. ($P = 4.0 \times 10^{-9}$, Cohen’s $d = 0.3$, two-sided Student’s t -test).

Therapist Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			Average	High-quality	Low-quality
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	freq.	freq.
Client behavior: Sharing negative emotions															
<i>n</i>	1063			1016			1236			1114			820	646	174
Reflections on Needs	38.9 _{48.8}	4.0×10^{-9}	0.3	32.3 _{46.8}	2.5×10^{-2}	0.2	28.9 _{45.3}	1.0×10^0	0.1	31.9 _{46.6}	3.9×10^{-2}	0.2	24.5 _{44.5}	28.9 _{45.4}	20.1 _{40.2}
Reflections on Emotions	27.4 _{44.6}	8.0×10^{-3}	0.2	33.9 _{47.3}	1.4×10^{-9}	0.3	25.1 _{43.4}	2.9×10^{-1}	0.1	24.0 _{42.7}	1.0×10^0	0.1	19.4 _{42.7}	27.4 _{44.6}	11.5 _{32.0}
Reflections on Values	4.5 _{20.8}	1.0×10^0	-0.1	4.9 _{21.6}	1.0×10^0	-0.1	4.0 _{19.5}	1.0×10^0	-0.1	5.6 _{22.9}	1.0×10^0	0.0	6.3 _{25.4}	7.4 _{26.2}	5.2 _{22.2}
Reflections on Consequences	21.3 _{40.9}	1.0×10^0	0.0	14.0 _{34.7}	6.1×10^{-4}	-0.2	16.7 _{37.4}	2.8×10^{-1}	-0.1	22.2 _{41.6}	1.0×10^0	0.0	21.8 _{39.8}	18.3 _{38.7}	25.3 _{43.6}
Reflections on Conflicts	7.9 _{27.0}	1.0×10^0	0.0	5.3 _{22.4}	8.2×10^{-1}	-0.1	8.4 _{27.8}	1.0×10^0	0.0	8.6 _{28.1}	1.0×10^0	0.0	8.4 _{29.5}	10.5 _{30.7}	6.3 _{24.4}
Reflections on Strengths	23.1 _{42.2}	2.0×10^{-20}	0.5	17.2 _{37.8}	8.9×10^{-10}	0.3	17.8 _{38.3}	2.8×10^{-11}	0.3	11.8 _{32.2}	1.3×10^{-2}	0.2	6.6 _{26.6}	8.5 _{27.9}	4.6 _{21.0}
Questions on Experiences	22.3 _{41.6}	2.8×10^{-7}	-0.3	18.4 _{38.8}	2.2×10^{-13}	-0.4	32.9 _{47.0}	1.0×10^0	0.0	34.9 _{47.7}	1.0×10^0	0.0	34.6 _{48.6}	40.6 _{49.1}	28.7 _{45.4}
Questions on Perspectives	3.3 _{17.9}	1.0×10^0	0.0	0.8 _{8.8}	3.1×10^{-2}	-0.2	5.6 _{23.0}	1.0×10^0	0.1	4.9 _{21.7}	1.0×10^0	0.1	3.3 _{19.9}	4.8 _{21.4}	1.7 _{13.1}
Questions on Emotions	1.6 _{12.6}	2.8×10^{-12}	-0.4	1.1 _{10.4}	3.6×10^{-14}	-0.4	3.2 _{17.7}	3.1×10^{-7}	-0.3	4.5 _{20.7}	1.1×10^{-3}	-0.2	9.6 _{31.1}	11.8 _{32.2}	7.5 _{26.4}
Problem-Solving	37.1 _{48.3}	4.4×10^{-17}	0.4	35.5 _{47.9}	2.8×10^{-14}	0.4	33.2 _{47.1}	2.3×10^{-11}	0.3	24.4 _{43.0}	2.5×10^{-1}	0.1	19.0 _{35.7}	12.1 _{32.6}	25.9 _{43.9}
Planning	13.9 _{34.6}	8.8×10^{-4}	0.2	12.8 _{33.4}	1.9×10^{-2}	0.2	19.3 _{39.4}	7.0×10^{-12}	0.3	13.5 _{34.2}	2.7×10^{-3}	0.2	7.6 _{25.0}	6.0 _{23.8}	9.2 _{29.0}
Normalizing	23.0 _{42.1}	3.4×10^{-2}	0.2	30.3 _{46.0}	2.8×10^{-10}	0.3	12.9 _{33.6}	1.0×10^0	-0.1	14.3 _{35.0}	1.0×10^0	-0.1	16.5 _{36.2}	14.7 _{35.4}	18.4 _{38.9}
Psychoeducation	4.5 _{20.8}	1.0×10^0	0.0	10.3 _{30.5}	2.5×10^{-4}	0.2	6.3 _{24.3}	1.0×10^0	0.1	3.0 _{17.0}	1.0×10^0	-0.1	4.7 _{17.5}	2.0 _{14.1}	7.5 _{26.4}
Client behavior: Sharing experiences															
<i>n</i>	865			842			940			879			759	533	226
Reflections on Needs	33.3 _{47.2}	1.4×10^{-4}	0.2	28.0 _{44.9}	9.6×10^{-1}	0.1	26.0 _{43.9}	1.0×10^0	0.1	27.0 _{44.4}	1.0×10^0	0.1	22.5 _{42.8}	26.5 _{44.2}	18.6 _{39.0}
Reflections on Emotions	17.9 _{38.4}	1.0×10^0	0.1	23.9 _{42.7}	1.8×10^{-5}	0.3	20.0 _{40.0}	4.3×10^{-2}	0.2	19.3 _{39.5}	1.6×10^{-1}	0.2	13.5 _{36.8}	19.9 _{40.0}	7.1 _{25.7}
Reflections on Values	5.7 _{23.1}	1.0×10^0	0.0	5.5 _{22.7}	1.0×10^0	0.0	4.5 _{20.7}	1.0×10^0	0.0	5.1 _{22.1}	1.0×10^0	0.0	4.7 _{23.1}	7.1 _{25.8}	2.2 _{14.7}
Reflections on Consequences	19.4 _{39.6}	1.0×10^0	0.0	10.8 _{31.1}	1.8×10^{-3}	-0.2	14.1 _{34.9}	1.0×10^0	-0.1	16.3 _{36.9}	1.0×10^0	-0.1	18.2 _{38.5}	17.8 _{38.3}	18.6 _{39.0}
Reflections on Conflicts	6.4 _{24.4}	1.0×10^0	0.0	3.4 _{18.2}	1.0×10^0	-0.1	7.2 _{25.9}	1.0×10^0	0.1	5.5 _{22.7}	1.0×10^0	0.0	5.8 _{25.1}	8.1 _{27.3}	3.5 _{18.5}
Reflections on Strengths	21.5 _{41.1}	5.4×10^{-11}	0.4	18.3 _{38.7}	1.4×10^{-6}	0.3	13.5 _{34.2}	1.0×10^{-1}	0.2	12.5 _{33.1}	7.3×10^{-1}	0.1	8.4 _{29.9}	12.0 _{32.5}	4.9 _{21.6}
Questions on Experiences	23.5 _{42.4}	3.6×10^{-18}	-0.5	25.3 _{43.5}	1.3×10^{-14}	-0.4	35.1 _{47.8}	4.1×10^{-3}	-0.2	36.4 _{48.1}	4.7×10^{-2}	-0.2	44.7 _{49.8}	45.2 _{49.8}	44.2 _{49.8}
Questions on Perspectives	4.4 _{20.5}	7.9×10^{-1}	0.1	1.4 _{11.9}	1.0×10^0	0.0	5.0 _{21.8}	1.3×10^{-1}	0.2	4.6 _{20.9}	5.1×10^{-1}	0.1	2.0 _{15.6}	3.2 _{17.6}	0.9 _{9.4}
Questions on Emotions	2.2 _{14.7}	2.4×10^{-7}	-0.3	1.8 _{13.2}	1.5×10^{-8}	-0.3	3.9 _{19.5}	2.6×10^{-3}	-0.2	5.8 _{23.4}	1.0×10^0	-0.1	8.8 _{28.6}	9.2 _{28.9}	8.4 _{27.8}
Problem-Solving	29.6 _{45.7}	7.1×10^{-13}	0.4	23.9 _{42.7}	2.1×10^{-5}	0.3	21.2 _{40.9}	6.7×10^{-3}	0.2	15.1 _{35.9}	1.0×10^0	0.0	14.0 _{32.7}	9.4 _{29.2}	18.6 _{39.0}
Planning	12.9 _{33.6}	5.9×10^{-3}	0.2	10.5 _{30.6}	1.0×10^0	0.1	13.1 _{33.7}	3.3×10^{-3}	0.2	8.3 _{27.6}	1.0×10^0	0.0	7.0 _{25.5}	6.9 _{25.4}	7.1 _{25.7}
Normalizing	16.2 _{36.9}	1.0×10^0	0.1	18.2 _{38.6}	1.0×10^0	0.1	8.8 _{28.4}	5.5×10^{-2}	-0.2	11.0 _{31.4}	1.0×10^0	-0.1	14.0 _{34.7}	13.9 _{34.6}	14.2 _{34.9}
Psychoeducation	3.1 _{17.4}	4.5×10^{-1}	-0.1	5.7 _{23.2}	1.0×10^0	0.0	2.8 _{16.4}	7.7×10^{-2}	-0.2	1.1 _{10.6}	7.4×10^{-7}	-0.3	5.7 _{20.7}	2.6 _{16.0}	8.8 _{28.5}
Client behavior: Gained insights															
<i>n</i>	450			365			430			437			258	214	44
Reflections on Needs	28.0 _{44.9}	1.0×10^0	0.0	29.3 _{45.6}	1.0×10^0	0.1	27.9 _{44.9}	1.0×10^0	0.0	22.9 _{42.1}	1.0×10^0	-0.1	27.0 _{45.1}	29.0 _{45.5}	25.0 _{43.8}
Reflections on Emotions	9.1 _{28.8}	1.0×10^0	-0.1	11.8 _{32.3}	1.0×10^0	0.0	10.5 _{30.6}	1.0×10^0	-0.1	6.9 _{25.3}	1.0×10^0	-0.2	12.5 _{35.5}	15.9 _{36.6}	9.1 _{29.1}
Reflections on Values	5.6 _{22.9}	1.0×10^0	-0.1	8.8 _{28.3}	1.0×10^0	0.0	6.7 _{25.1}	1.0×10^0	-0.1	7.8 _{26.8}	1.0×10^0	0.0	8.1 _{28.6}	9.3 _{29.2}	6.8 _{25.5}
Reflections on Consequences	18.2 _{38.6}	1.0×10^0	0.0	12.6 _{33.2}	1.0×10^0	-0.2	15.8 _{36.5}	1.0×10^0	-0.1	14.9 _{35.6}	1.0×10^0	-0.1	18.8 _{37.0}	15.0 _{35.7}	22.7 _{42.4}
Reflections on Conflicts	4.2 _{20.1}	1.0×10^0	0.0	3.8 _{19.2}	1.0×10^0	0.0	6.3 _{24.3}	1.0×10^0	0.1	4.8 _{21.4}	1.0×10^0	0.0	3.9 _{21.9}	5.6 _{23.1}	2.3 _{15.1}
Reflections on Strengths	51.3 _{50.0}	4.0×10^{-24}	0.9	36.4 _{48.2}	2.1×10^{-9}	0.6	31.4 _{46.5}	2.1×10^{-6}	0.4	32.7 _{47.0}	2.8×10^{-7}	0.5	12.5 _{35.5}	15.9 _{36.6}	9.1 _{29.1}
Questions on Experiences	13.1 _{33.8}	2.3×10^{-5}	-0.4	16.2 _{36.9}	1.2×10^{-2}	-0.3	24.7 _{43.1}	1.0×10^0	-0.1	17.4 _{37.9}	3.0×10^{-2}	-0.3	29.2 _{48.6}	42.5 _{49.6}	15.9 _{37.0}
Questions on Perspectives	3.1 _{17.4}	1.0×10^0	-0.1	2.2 _{14.7}	1.0×10^0	-0.1	4.4 _{20.6}	1.0×10^0	0.0	4.3 _{20.4}	1.0×10^0	0.0	4.2 _{22.7}	6.1 _{23.9}	2.3 _{15.1}
Questions on Emotions	0.7 _{8.1}	5.2×10^{-9}	-0.5	1.9 _{13.7}	9.7×10^{-5}	-0.4	4.2 _{20.1}	5.0×10^{-2}	-0.3	4.1 _{19.9}	4.0×10^{-2}	-0.3	10.8 _{30.7}	10.3 _{30.4}	11.4 _{32.1}
Problem-Solving	41.1 _{49.3}	2.0×10^{-2}	0.3	37.5 _{48.5}	7.0×10^{-1}	0.2	37.0 _{48.3}	8.9×10^{-1}	0.2	25.9 _{43.8}	1.0×10^0	0.0	27.9 _{39.6}	15.0 _{35.7}	40.9 _{49.7}
Planning	34.9 _{47.7}	4.4×10^{-10}	0.5	28.2 _{45.1}	3.5×10^{-5}	0.4	38.4 _{48.7}	9.5×10^{-13}	0.6	25.9 _{43.8}	4.4×10^{-4}	0.4	11.5 _{32.1}	11.7 _{32.2}	11.4 _{32.1}
Normalizing	32.4 _{46.9}	3.1×10^{-3}	0.3	30.1 _{45.9}	5.6×10^{-2}	0.3	13.7 _{34.4}	1.0×10^0	-0.1	21.7 _{41.3}	1.0×10^0	0.1	18.6 _{35.1}	12.1 _{32.7}	25.0 _{43.8}
Psychoeducation	8.7 _{28.2}	1.0×10^0	0.0	12.6 _{33.2}	1.0×10^0	0.1	9.5 _{29.4}	1.0×10^0	0.0	9.2 _{28.9}	1.0×10^0	0.0	8.4 _{18.4}	0.9 _{9.6}	15.9 _{37.0}

Table S7. Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in High-low quality dataset⁴⁵. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student's t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13 \times 6 = 78$). d is the cohen's d when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON CONSEQUENCES by 38.9% which is 15.8% less than average human therapists. (Single: 11.5%, Average human therapists: 27.3% from Table S5; $P = 5.3 \times 10^{-17}$, Cohen's $d = -0.4$, two-sided Student's t -test)

Therapist Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			GPT-4			GPT-3.5			Llama2-70b			Llama2-13b		
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d
Client behavior: Changing unhealthy behavior																								
n	1553			731			1553			595			1553			658			1553			529		
Reflections on Needs	25.243.4	1.0×10^0	-0.1	30.846.2	1.0×10^0	0.1	17.037.6	9.7×10^{-6}	-0.2	21.040.8	2.0×10^{-1}	-0.2	31.246.4	1.0×10^0	0.1	28.445.1	1.0×10^0	0.0	31.146.3	1.0×10^0	0.1	30.145.9	1.0×10^0	0.1
Reflections on Emotions	7.526.4	1.0×10^0	-0.1	7.927.0	1.0×10^0	-0.1	5.023.1	6.2×10^{-2}	-0.2	8.127.4	1.0×10^0	-0.1	8.527.9	1.0×10^0	-0.1	13.333.9	1.0×10^0	0.1	10.030.0	1.0×10^0	0.0	8.828.3	1.0×10^0	-0.1
Reflections on Values	3.819.2	1.0×10^0	0.0	1.813.5	3.6×10^{-1}	-0.1	1.813.4	2.2×10^{-1}	-0.1	4.320.4	1.0×10^0	0.0	4.202.2	1.0×10^0	0.0	6.124.0	1.0×10^0	0.1	4.320.3	1.0×10^0	0.0	4.620.9	1.0×10^0	0.0
Reflections on Consequences	11.531.9	5.3×10^{-17}	-0.4	9.229.0	2.2×10^{-5}	-0.5	5.222.2	8.8×10^{-34}	-0.6	8.127.4	4.7×10^{-20}	-0.5	23.942.7	1.0×10^0	-0.1	13.534.2	4.1×10^{-21}	-0.4	21.441.1	8.0×10^{-3}	-0.1	23.524.4	1.0×10^0	-0.1
Reflections on Conflicts	2.315.1	9.8×10^{-4}	-0.2	1.311.5	5.9×10^{-5}	-0.2	1.512.2	4.5×10^{-5}	-0.2	3.217.7	2.1×10^{-1}	-0.2	5.021.7	1.0×10^0	-0.1	4.119.7	4.6×10^{-2}	-0.1	5.023.0	1.0×10^0	-0.1	7.125.8	1.0×10^0	0.0
Reflections on Strengths	76.242.6	2.0×10^{-278}	1.9	62.748.4	5.0×10^{-159}	1.4	44.407.7	2.0×10^{-77}	0.9	42.349.5	2.9×10^{-62}	0.9	29.515.6	1.4×10^{-42}	0.5	25.243.4	2.5×10^{-28}	0.4	18.138.5	5.6×10^{-10}	0.2	15.836.5	1.1×10^{-5}	0.2
Questions on Experiences	16.336.9	2.8×10^{-19}	-0.4	8.427.8	1.0×10^{-34}	-0.6	21.310.0	3.2×10^{-9}	-0.3	27.444.6	5.0×10^{-2}	-0.2	14.735.4	9.7×10^{-40}	-0.5	19.339.4	2.1×10^{-22}	-0.4	30.446.0	1.7×10^{-1}	-0.1	23.242.2	2.8×10^{-12}	-0.3
Questions on Perspectives	2.214.6	1.0×10^0	-0.1	0.00.0	1.5×10^{-3}	-0.2	4.320.2	1.0×10^0	0.0	4.320.4	1.0×10^0	0.0	3.017.1	1.0×10^0	0.0	1.210.7	3.0×10^{-3}	-0.1	3.719.0	1.0×10^0	0.0	5.021.7	1.0×10^0	0.1
Questions on Emotions	0.35.2	4.7×10^{-14}	-0.4	0.24.1	9.8×10^{-12}	-0.4	1.110.3	3.5×10^{-10}	-0.3	0.46.1	7.4×10^{-10}	-0.3	0.89.1	1.1×10^{-24}	-0.4	0.35.7	4.5×10^{-29}	-0.4	4.119.7	2.1×10^{-7}	-0.2	5.522.8	1.8×10^{-3}	-0.2
Problem-Solving	38.348.6	5.8×10^{-5}	0.2	48.290.0	6.9×10^{-18}	0.4	56.249.6	3.0×10^{-37}	0.6	41.449.3	5.2×10^{-7}	0.3	41.949.3	4.0×10^{-14}	0.3	34.347.5	2.1×10^{-2}	0.1	29.245.5	1.0×10^0	0.0	25.443.5	1.0×10^0	-0.1
Planning	47.600.0	5.8×10^{-76}	0.9	55.049.8	9.3×10^{-99}	1.1	64.048.0	1.9×10^{-151}	1.3	48.090.0	4.5×10^{-66}	0.9	25.243.4	6.7×10^{-15}	0.3	23.842.6	7.8×10^{-12}	0.3	19.139.3	2.4×10^{-3}	0.1	18.438.7	2.0×10^{-2}	0.1
Normalizing	52.740.0	2.6×10^{-96}	1.0	52.800.0	1.2×10^{-87}	1.0	18.518.9	3.7×10^{-1}	0.1	21.741.3	9.1×10^{-4}	0.2	16.437.0	1.0×10^0	0.1	12.132.6	1.0×10^0	-0.1	6.224.2	5.9×10^{-12}	-0.3	11.311.5	1.0×10^0	-0.1
Psychoeducation	9.228.9	2.9×10^{-2}	0.2	8.285.5	1.3×10^{-1}	0.2	21.441.1	2.7×10^{-31}	0.6	14.034.7	7.7×10^{-10}	0.3	4.721.2	1.0×10^0	0.0	13.233.9	9.3×10^{-13}	0.3	3.017.0	2.2×10^{-2}	-0.1	3.017.1	3.3×10^{-2}	-0.1
Client behavior: Sustaining unhealthy behavior																								
n	389			47			389			30			389			164			389			56		
Reflections on Needs	59.649.6	6.0×10^{-4}	0.7	43.340.4	1.0×10^0	0.3	34.147.6	1.0×10^0	0.1	41.149.6	1.0×10^0	0.3	33.747.3	1.0×10^0	0.1	35.247.8	1.0×10^0	0.2	30.346.0	1.0×10^0	0.1	28.545.2	1.0×10^0	0.0
Reflections on Emotions	17.038.0	1.0×10^0	0.2	13.314.6	1.0×10^0	0.0	9.829.8	1.0×10^0	-0.1	25.031.7	6.0×10^{-1}	0.4	8.227.5	1.0×10^0	-0.1	14.935.7	1.0×10^0	0.1	11.612.0	1.0×10^0	0.0	12.032.2	1.0×10^0	0.0
Reflections on Values	4.320.4	1.0×10^0	-0.1	3.318.3	1.0×10^0	-0.1	5.522.8	1.0×10^0	-0.1	0.00.0	1.0×10^0	-0.3	4.119.9	1.0×10^0	-0.1	3.618.7	1.0×10^0	-0.1	4.420.5	1.0×10^0	-0.1	4.420.5	1.0×10^0	-0.1
Reflections on Consequences	34.047.9	1.0×10^0	0.0	43.340.4	1.0×10^0	0.2	23.242.3	1.9×10^{-1}	-0.3	32.147.1	1.0×10^0	-0.1	35.247.8	1.0×10^0	0.0	26.244.0	2.0×10^{-1}	-0.2	28.045.0	1.0×10^0	-0.2	31.446.5	1.0×10^0	-0.1
Reflections on Conflicts	21.314.4	2.8×10^{-1}	0.5	10.030.5	1.0×10^0	0.1	7.326.1	1.0×10^0	0.0	14.335.3	1.0×10^0	0.2	10.030.1	1.0×10^0	0.1	6.725.0	1.0×10^0	0.1	6.424.6	1.0×10^0	-0.1	7.225.9	1.0×10^0	0.0
Reflections on Strengths	14.936.0	1.0×10^0	0.4	13.314.6	1.0×10^0	0.3	13.434.2	9.8×10^{-2}	0.3	5.422.7	1.0×10^0	0.0	11.331.7	1.9×10^{-1}	0.2	7.726.7	1.0×10^0	0.1	5.122.1	1.0×10^0	0.0	4.621.0	1.0×10^0	0.0
Questions on Experiences	34.047.9	1.0×10^0	-0.1	23.343.0	1.0×10^0	-0.3	29.945.9	1.0×10^0	-0.2	50.040.5	1.0×10^0	0.2	17.738.2	1.5×10^{-9}	-0.5	20.340.3	4.8×10^{-7}	-0.4	29.645.7	3.4×10^{-1}	-0.2	27.844.8	5.2×10^{-2}	-0.2
Questions on Perspectives	14.936.0	5.6×10^{-2}	0.5	6.725.4	1.0×10^0	0.2	10.440.6	1.2×10^{-1}	0.3	12.533.4	2.8×10^{-1}	0.4	6.725.0	1.0×10^0	0.1	2.315.1	1.0×10^0	-0.1	5.422.6	1.0×10^0	0.1	5.122.1	1.0×10^0	0.1
Questions on Emotions	0.00.0	1.0×10^0	-0.3	0.00.0	1.0×10^0	-0.3	1.211.0	6.7×10^{-2}	-0.3	0.00.0	1.0×10^0	-0.3	0.51.2	2.1×10^{-6}	-0.4	0.51.2	2.1×10^{-6}	-0.4	3.117.3	3.6×10^{-2}	-0.3	6.725.0	1.0×10^0	-0.1
Problem-Solving	40.449.6	6.8×10^{-1}	0.4	50.050.9	6.0×10^{-2}	0.6	58.549.4	3.3×10^{-15}	0.8	46.460.3	1.1×10^{-2}	0.5	30.346.0	1.0×10^0	0.2	26.244.0	1.0×10^0	0.1	23.742.5	1.0×10^0	0.0	18.338.7	1.0×10^0	-0.1
Planning	14.936.0	1.0×10^0	0.4	26.745.0	7.5×10^{-4}	0.8	40.949.3	1.1×10^{-25}	1.1	21.441.4	1.8×10^{-3}	0.6	12.613.2	8.3×10^{-2}	0.2	8.027.1	1.0×10^0	0.1	11.131.4	6.9×10^{-1}	0.2	8.527.9	1.0×10^0	0.1
Normalizing	31.947.1	7.9×10^{-2}	0.5	46.750.7	8.1×10^{-5}	0.9	11.031.4	1.0×10^0	-0.1	5.422.7	1.0×10^0	-0.3	12.613.2	1.0×10^0	0.0	13.133.8	1.0×10^0	0.0	5.122.1	9.6×10^{-4}	-0.3	7.726.7	2.7×10^{-1}	-0.2
Psychoeducation	4.320.4	1.0×10^0	-0.2	3.318.3	1.0×10^0	-0.2	11.632.1	1.0×10^0	0.1	10.731.2	1.0×10^0	0.0	1.813.3	1.2×10^{-5}	-0.4	12.613.2	1.0×10^0	0.1	2.615.8	3.8×10^{-4}	-0.3	1.813.3	1.2×10^{-5}	-0.4
Client behavior: Sharing positive emotions																								
n	294			364			294			260			294			114			294			240		
Reflections on Needs	4.420.5	1.5×10^{-5}	-0.4	17.337.9	1.0×10^0	0.0	11.431.9	1.0×10^0	-0.2	5.823.5	9.8×10^{-3}	-0.3	23.542.5	1.0×10^0	0.1	18.438.8	1.0×10^0	0.0	24.142.9	1.0×10^0	0.2	25.543.7	1.0×10^0	0.2
Reflections on Emotions	3.618.6	1.0×10^0	-0.2	3.117.3	1.0×10^0	-0.2	1.813.2	1.0×10^0	-0.2	2.515.6	1.0×10^0	-0.2	7.526.4	1.0×10^0	0.0	12.613.2	1.0×10^0	0.2	7.125.8	1.0×10^0	0.0	8.527.9	1.0×10^0	0.0
Reflections on Values	0.89.1	1.0×10^0	-0.2	1.913.8	1.0×10^0	-0.1	2.616.1	1.0×10^0	0.0	3.819.0	1.0×10^0	0.0	5.422.7	1.0×10^0	0.1	5.823.4	1.0×10^0	0.1	6.124.0	1.0×10^0	0.1	8.527.9	9.2×10^{-1}	0.2
Reflections on Consequences	0.89.1	2.5×10^{-13}	-0.6	1.512.3	2.4×10^{-8}	-0.5	7.025.7	9.9×10^{-1}	-0.3	0.89.1	6.0×10^{-9}	-0.6	12.933.6	1.0×10^0	-0.1	8.527.9	3.5×10^{-1}	-0.2	17.037.6	1.0×10^0	-0.1	10.634.3	1.0×10^0	-0.1
Reflections on Conflicts	0.00.0	1.4×10^{-1}	-0.2	0.00.0	6.4×10^{-1}	-0.2	1.813.2	1.0×10^0	-0.1	0.00.0	8.6×10^{-1}	-0.2	4.119.8	1.0×10^0	0.0	3.719.0	1.0×10^0	0.0	6.124.0	1.0×10^0	0.1	3.418.2	1.0×10^0	0.0
Reflections on Strengths	97.814.7	1.0×10^{-197}																						

Table S8. (Continue Table S7) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in High-low quality dataset⁴⁵. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13 \times 6 = 78$). d is the cohen’s d when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON NEEDS by 38.9% which is 23.6% more than average human therapists. (Single: 48.1%, Average human therapists: 24.5% from Table S6; $P = 2.3 \times 10^{-10}$, Cohen’s $d = 0.5$, two-sided Student’s t -test)

Therapist Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			GPT-4			GPT-3.5			Llama2-70b			Llama2-13b		
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d
Client behavior: Sharing negative emotions																								
n	820			243			820			196			820			416			820			294		
Reflections on Needs	48.1501	2.3×10^{-10}	0.5	41.8495	1.5×10^{-4}	0.4	23.1222	1.0×10^0	0.0	30.3460	1.0×10^0	0.1	36.2481	2.8×10^{-5}	0.3	30.0259	1.0×10^0	0.1	31.8468	9.4×10^{-2}	0.2	32.4468	3.6×10^{-2}	0.2
Reflections on Emotions	23.9427	1.0×10^0	0.1	24.5431	1.0×10^0	0.1	19.5396	1.0×10^0	0.0	22.1416	1.0×10^0	0.1	28.4451	3.0×10^{-3}	0.2	36.1481	1.5×10^{-11}	0.4	27.9449	7.3×10^{-3}	0.2	24.6311	1.0×10^0	0.1
Reflections on Values	7.4262	1.0×10^0	0.0	2.6158	1.0×10^0	-0.2	3.8193	1.0×10^0	-0.1	4.4206	1.0×10^0	-0.1	3.7188	1.0×10^0	-0.1	5.5228	1.0×10^0	0.0	4.0197	1.0×10^0	-0.1	6.0237	1.0×10^0	0.0
Reflections on Consequences	21.4411	1.0×10^0	0.0	18.4388	1.0×10^0	-0.1	8.9285	4.2×10^{-7}	-0.4	14.6354	5.2×10^{-1}	-0.2	21.2409	1.0×10^0	0.0	12.9336	9.8×10^{-5}	-0.2	20.7406	1.0×10^0	0.0	24.9433	1.0×10^0	0.1
Reflections on Conflicts	6.2341	1.0×10^0	-0.1	5.1221	1.0×10^0	-0.1	4.8214	1.0×10^0	-0.1	8.5279	1.0×10^0	0.0	8.4278	1.0×10^0	0.0	5.4225	1.0×10^0	-0.1	10.2303	1.0×10^0	0.1	8.7281	1.0×10^0	0.0
Reflections on Strengths	41.6494	2.7×10^{-41}	1.1	27.6448	4.3×10^{-15}	0.7	29.3456	2.8×10^{-25}	0.7	21.1409	7.3×10^{-10}	0.5	17.7382	8.5×10^{-10}	0.3	14.8355	1.1×10^{-5}	0.3	12.0325	1.9×10^{-2}	0.2	8.4278	1.0×10^0	0.1
Questions on Experiences	35.4479	1.0×10^0	0.0	23.5225	2.4×10^{-1}	-0.2	31.0463	1.0×10^0	-0.1	52.0600	1.6×10^{-5}	0.4	18.4388	9.6×10^{-12}	-0.4	17.2378	7.0×10^{-14}	-0.4	33.9474	1.0×10^0	0.0	28.8453	9.0×10^{-1}	-0.1
Questions on Perspectives	3.7189	1.0×10^0	0.0	0.571	1.0×10^0	-0.2	7.5263	1.4×10^{-1}	0.2	6.1240	1.0×10^0	0.1	3.2175	1.0×10^0	0.0	0.92	1.3×10^{-1}	-0.2	4.6210	1.0×10^0	0.1	4.5208	1.0×10^0	0.1
Questions on Emotions	0.891	1.2×10^{-3}	-0.3	1.5123	2.9×10^{-2}	-0.3	2.2146	3.1×10^{-4}	-0.3	1.4116	7.8×10^{-4}	-0.3	1.8134	4.9×10^{-9}	-0.3	1.038	4.4×10^{-12}	-0.4	3.8191	3.9×10^{-4}	-0.2	5.0230	2.4×10^{-1}	-0.1
Problem-Solving	42.8496	2.5×10^{-14}	0.6	60.7490	6.9×10^{-37}	1.1	55.5498	5.8×10^{-44}	0.9	42.9496	3.4×10^{-16}	0.6	35.4478	5.2×10^{-13}	0.4	29.5456	1.7×10^{-5}	0.3	21.8413	1.0×10^0	0.1	17.8383	1.0×10^0	0.0
Planning	17.7382	1.2×10^{-4}	0.4	25.5437	4.7×10^{-12}	0.6	41.8494	1.4×10^{-51}	1.0	29.3456	8.5×10^{-21}	0.7	12.8334	3.0×10^{-2}	0.2	9.8297	1.0×10^0	0.1	7.8268	1.0×10^0	0.0	7.8268	1.0×10^0	0.0
Normalizing	33.7474	1.7×10^{-7}	0.4	42.9496	7.5×10^{-15}	0.7	19.5396	1.0×10^0	0.1	17.3379	1.0×10^0	0.0	19.9399	1.0×10^0	0.0	27.3446	7.0×10^{-6}	0.3	9.6295	1.9×10^{-3}	-0.2	13.2338	1.0×10^0	-0.1
Psychoeducation	10.3304	2.7×10^{-2}	0.3	10.7310	2.5×10^{-2}	0.3	16.1368	2.8×10^{-11}	0.4	8.2274	1.0×10^0	0.2	2.8165	1.0×10^0	-0.1	10.2303	5.8×10^{-4}	0.2	1.3115	2.9×10^{-4}	-0.2	1.1104	2.7×10^{-5}	-0.3
Client behavior: Sharing experiences																								
n	759			106			759			83			759			181			759			120		
Reflections on Needs	57.5497	2.3×10^{-12}	0.8	38.6190	1.1×10^{-1}	0.4	26.0410	1.0×10^0	0.1	35.8482	1.5×10^{-1}	0.3	29.9458	9.3×10^{-2}	0.2	26.9414	1.0×10^0	0.1	26.0438	1.0×10^0	0.1	25.6316	1.0×10^0	0.1
Reflections on Emotions	23.6427	7.5×10^{-1}	0.3	24.1430	1.0×10^0	0.3	23.8247	8.6×10^{-2}	0.3	29.2456	2.4×10^{-3}	0.4	17.1377	1.0×10^0	0.1	23.8426	3.5×10^{-5}	0.3	19.1393	3.2×10^{-1}	0.1	17.8383	1.0×10^0	0.1
Reflections on Values	8.5280	1.0×10^0	0.2	6.0239	1.0×10^0	0.1	3.9193	1.0×10^0	0.0	1.7129	1.0×10^0	-0.1	5.3224	1.0×10^0	0.0	5.4226	1.0×10^0	0.0	4.6210	1.0×10^0	0.0	5.7231	1.0×10^0	0.0
Reflections on Consequences	20.8407	1.0×10^0	0.1	14.5154	1.0×10^0	-0.1	7.7268	4.3×10^{-2}	-0.3	16.7374	1.0×10^0	0.0	19.2394	1.0×10^0	0.0	10.4306	1.0×10^{-3}	-0.2	15.7364	1.0×10^0	-0.1	16.2369	1.0×10^0	-0.1
Reflections on Conflicts	8.5280	1.0×10^0	0.1	6.0239	1.0×10^0	0.0	6.0249	1.0×10^0	0.0	20.8408	5.4×10^{-3}	0.4	8.3378	1.0×10^0	0.1	6.1239	1.0×10^0	0.0	3.2175	1.0×10^0	-0.1	7.4262	1.0×10^0	0.0
Reflections on Strengths	34.9479	1.1×10^{-12}	0.8	27.7450	1.4×10^{-5}	0.6	21.0408	2.3×10^{-4}	0.4	20.8408	5.4×10^{-3}	0.4	19.6197	5.6×10^{-8}	0.3	17.3378	4.0×10^{-5}	0.3	11.7322	1.0×10^0	0.1	11.2316	1.0×10^0	0.1
Questions on Experiences	40.6493	1.0×10^0	-0.1	44.6300	1.0×10^0	0.0	40.3492	1.0×10^0	-0.1	56.7498	1.0×10^0	0.2	21.1408	1.9×10^{-21}	-0.5	23.2422	2.3×10^{-17}	-0.5	33.9474	1.1×10^{-3}	-0.2	33.2471	3.0×10^{-4}	-0.2
Questions on Perspectives	4.7213	1.0×10^0	0.2	1.2110	1.0×10^0	-0.1	7.2259	4.7×10^{-2}	0.3	5.0219	1.0×10^0	0.2	4.3204	1.0×10^0	0.1	1.4120	1.0×10^0	0.0	4.5207	7.5×10^{-1}	0.1	4.5207	7.5×10^{-1}	0.1
Questions on Emotions	2.8167	1.0×10^0	-0.2	7.2261	1.0×10^0	-0.1	3.3180	1.0×10^0	-0.2	4.2201	1.0×10^0	-0.2	2.1144	7.8×10^{-7}	-0.3	1.2108	7.6×10^{-10}	-0.4	4.1198	1.5×10^{-2}	-0.2	6.1239	1.0×10^0	-0.1
Problem-Solving	41.5495	8.4×10^{-12}	0.8	39.8492	1.6×10^{-8}	0.7	46.4500	2.1×10^{-23}	0.9	35.8482	3.1×10^{-8}	0.6	27.9449	5.1×10^{-10}	0.4	22.1415	1.8×10^{-3}	0.2	15.2359	1.0×10^0	0.0	11.9324	1.0×10^0	-0.1
Planning	19.8400	7.3×10^{-4}	0.5	16.9377	1.3×10^{-1}	0.4	34.3476	4.2×10^{-23}	0.9	23.3425	5.4×10^{-7}	0.6	12.0325	7.2×10^{-2}	0.2	9.7297	1.0×10^0	0.1	8.0272	1.0×10^0	0.0	5.9236	1.0×10^0	0.0
Normalizing	24.5432	3.7×10^{-1}	0.3	43.4499	5.1×10^{-10}	0.8	12.7334	1.0×10^0	0.0	15.0359	1.0×10^0	0.0	15.0358	1.0×10^0	0.0	15.4361	1.0×10^0	0.0	7.9270	1.0×10^{-2}	-0.2	10.4306	1.0×10^0	-0.1
Psychoeducation	3.8191	1.0×10^0	-0.1	4.8215	1.0×10^0	0.0	11.0314	4.4×10^{-1}	0.2	5.8235	1.0×10^0	0.0	3.0172	4.4×10^{-1}	-0.1	5.8234	1.0×10^0	0.0	0.889	1.4×10^{-7}	-0.3	0.463	1.1×10^{-9}	-0.3
Client behavior: Gained insights																								
n	258			192			258			107			258			172			258			179		
Reflections on Needs	19.3395	1.0×10^0	-0.2	34.6178	1.0×10^0	0.2	17.4381	1.0×10^0	-0.2	13.4342	5.7×10^{-2}	-0.3	34.5176	1.0×10^0	0.2	27.1446	1.0×10^0	0.0	34.9478	1.0×10^0	0.2	25.5457	1.0×10^0	0.1
Reflections on Emotions	3.1174	6.5×10^{-2}	-0.3	2.8166	5.7×10^{-1}	-0.3	2.9186	8.2×10^{-2}	-0.3	2.2148	2.3×10^{-2}	-0.4	13.6343	1.0×10^0	0.0	15.5363	1.0×10^0	0.1	15.5363	1.0×10^0	0.1	10.1302	1.0×10^0	-0.1
Reflections on Values	5.7233	1.0×10^0	-0.1	7.5264	1.0×10^0	0.0	6.4245	1.0×10^0	-0.1	7.3260	1.0×10^0	0.0	5.4272	1.0×10^0	-0.1	9.3291	1.0×10^0	0.0	7.0255	1.0×10^0	0.0	8.1274	1.0×10^0	0.0
Reflections on Consequences	9.4292	2.8×10^{-1}	-0.3	12.1328	1.0×10^0	-0.2	1.7131	9.0×10^{-7}	-0.6	2.2148	1.8×10^{-6}	-0.6	24.8433	1.0×10^0	0.1	12.8335	1.0×10^0	-0.2	25.2435	1.0×10^0	0.2	23.6226	1.0×10^0	0.1
Reflections on Conflicts	1.0102	1.0×10^0	-0.2	1.9136	1.0×10^0	-0.1	1.7131	1.0×10^0	-0.1	0.675	1.0×10^0	-0.2	6.6249	1.0×10^0	0.1	4.7211	1.0×10^0	0.0	9.3291	1.0×10^0	0.2	7.8268	1.0×10^0	0.2
Reflections on Strengths	82.8378	4.0×10^{-63}	1.9	58.9494	2.7×10^{-19}	1.2	51.7801	1.1×10^{-17}	0.9	57.0496	8.1×10^{-23}	1.1	27.9449	1.4×10^{-3}	0.4	27.1446	3.3×10^{-3}	0.4	17.8384	$$				

Table S9. Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset⁴⁵. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student's t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 14$). d is the cohen's d when compared with human therapist average. For instance, GPT-4 exhibits POSITIVE EMOTIONS by 1.5% while human therapists responds by 1.0% on average ($P = 8.4 \times 10^{-13}$, Cohen's $d = 0.1$, two-sided Student's t -test).

Linguistic Attributes	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ($n = 4893$)			GPT-3.5 ($n = 4746$)			Llama2-70b ($n = 4970$)			Llama2-13b ($n = 4820$)			Average ($n = 5446$)	High-quality ($n = 3907$)	Low-quality ($n = 1539$)
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	freq.	freq.
Positive Emotions	1.5 _{2,1}	8.4×10^{-13}	0.1	1.2 _{1,8}	9.2×10^{-5}	0.1	1.1 _{3,1}	1.0×10^0	0.0	1.1 _{3,3}	2.9×10^{-1}	0.0	1.0 _{4,2}	1.2 _{4,3}	0.7 _{3,6}
Negative Emotions	0.8 _{1,5}	1.0×10^0	0.0	0.9 _{1,6}	9.3×10^{-6}	0.1	0.8 _{1,9}	6.4×10^{-2}	0.1	0.6 _{1,8}	1.0×10^0	0.0	0.7 _{3,6}	0.7 _{3,4}	0.6 _{4,3}
BigWords	24.5 _{7,1}	0.0×10^0	1.6	23.1 _{7,7}	0.0×10^0	1.4	23.7 _{11,0}	0.0×10^0	1.2	19.8 _{12,6}	0.0×10^0	0.8	10.7 _{9,9}	11.0 _{9,9}	10.4 _{10,1}
Self	1.5 _{2,3}	1.1×10^{-84}	-0.4	1.8 _{2,6}	4.3×10^{-68}	-0.3	1.3 _{2,4}	7.9×10^{-134}	-0.5	1.9 _{3,0}	9.3×10^{-61}	-0.3	3.1 _{5,1}	2.5 _{4,5}	3.8 _{6,1}
Other	8.7 _{3,3}	8.2×10^{-1}	0.0	8.2 _{3,3}	8.7×10^{-12}	-0.1	7.5 _{4,8}	2.4×10^{-38}	-0.2	6.7 _{4,9}	2.0×10^{-85}	-0.4	8.9 _{7,1}	8.7 _{6,9}	9.2 _{7,4}
Health	2.7 _{2,8}	1.1×10^{-30}	0.2	3.0 _{3,3}	1.6×10^{-60}	0.3	2.0 _{2,9}	4.8×10^{-2}	0.1	1.7 _{3,9}	6.1×10^{-1}	0.0	1.8 _{4,3}	1.2 _{3,2}	2.5 _{6,1}
Wellness	0.9 _{1,6}	8.8×10^{-62}	0.3	0.8 _{1,6}	2.9×10^{-51}	0.3	0.6 _{1,3}	1.1×10^{-14}	0.1	0.5 _{1,9}	1.2×10^{-3}	0.1	0.3 _{2,1}	0.2 _{1,3}	0.4 _{3,5}
Behavioral Activation	6.8 _{4,0}	7.8×10^{-260}	0.7	5.9 _{4,3}	4.5×10^{-175}	0.5	5.7 _{5,2}	7.6×10^{-129}	0.5	5.1 _{5,7}	4.6×10^{-69}	0.3	3.2 _{5,9}	3.5 _{5,9}	2.9 _{5,9}
Meaning	3.5 _{2,3}	1.4×10^{-253}	0.7	3.1 _{2,2}	1.1×10^{-209}	0.6	2.1 _{2,4}	4.7×10^{-38}	0.2	1.9 _{2,7}	6.5×10^{-15}	0.2	1.4 _{3,5}	1.4 _{3,4}	1.4 _{3,7}
Purpose	3.8 _{2,6}	1.3×10^{-285}	0.7	3.5 _{2,5}	1.8×10^{-257}	0.7	2.5 _{2,6}	3.5×10^{-63}	0.3	2.2 _{3,0}	2.4×10^{-27}	0.2	1.5 _{3,6}	1.5 _{3,5}	1.5 _{3,8}
Motivation	3.4 _{2,4}	6.8×10^{-264}	0.7	3.1 _{2,3}	9.2×10^{-248}	0.6	2.1 _{2,3}	7.1×10^{-56}	0.3	1.9 _{2,8}	4.8×10^{-26}	0.2	1.3 _{3,3}	1.3 _{3,3}	1.3 _{3,4}
Sadness	0.7 _{1,3}	1.4×10^{-37}	0.3	0.8 _{1,4}	3.5×10^{-67}	0.3	0.6 _{1,5}	1.5×10^{-33}	0.2	0.5 _{1,6}	9.3×10^{-13}	0.1	0.3 _{1,5}	0.3 _{1,5}	0.2 _{1,6}
Sympathy	0.3 _{0,7}	4.1×10^{-11}	0.1	0.4 _{0,9}	8.5×10^{-41}	0.3	0.3 _{1,1}	6.7×10^{-18}	0.2	0.3 _{1,1}	1.4×10^{-9}	0.1	0.1 _{1,0}	0.2 _{1,1}	0.1 _{0,8}
Stress	1.1 _{1,8}	1.0×10^0	0.0	1.4 _{1,9}	2.7×10^{-7}	0.1	1.4 _{2,2}	3.6×10^{-8}	0.1	1.1 _{2,2}	1.0×10^0	0.0	1.1 _{3,5}	1.1 _{3,2}	1.1 _{4,2}

Table S10. Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset⁴⁵ for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student's t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen's d when compared with human therapist average. For instance, GPT-4 with single response simulation responds with POSITIVE EMOTIONS by 2.0% and GPT-4 with full conversation simulation responds by 1.3% when compared with the human therapists responds by 1.0% on average from Table S9. (Single: $P = 2.0 \times 10^{-14}$, Cohen's $d = 0.3$; two-sided student t-test) (Full: $P = 3.7 \times 10^{-5}$, Cohen's $d = 0.1$; two-sided student t-test).

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4 ($n = 3837$)			GPT-3.5 ($n = 3907$)			Llama2-70b ($n = 3906$)			Llama2-13b ($n = 3895$)			GPT-4 ($n = 1056$)			GPT-3.5 ($n = 839$)			Llama2-70b ($n = 1064$)			Llama2-13b ($n = 925$)		
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d
Positive Emotions	2.0 _{1,7}	2.0×10^{-14}	0.3	1.7 _{1,7}	1.5×10^{-5}	0.2	1.3 _{1,4}	5.2×10^{-2}	0.1	1.7 _{2,1}	3.2×10^{-6}	0.2	1.3 _{2,2}	3.7×10^{-5}	0.1	1.2 _{1,8}	2.5×10^{-2}	0.1	1.0 _{3,3}	1.0×10^0	0.0	1.0 _{3,5}	1.0×10^0	0.0
Negative Emotions	0.5 _{1,0}	1.0×10^0	-0.1	0.6 _{1,1}	1.0×10^0	0.0	0.7 _{1,1}	1.0×10^0	0.0	0.6 _{1,1}	1.0×10^0	0.0	0.8 _{1,6}	2.7×10^{-1}	0.0	1.0 _{1,7}	1.7×10^{-7}	0.1	0.9 _{2,0}	2.5×10^{-2}	0.1	0.6 _{1,9}	1.0×10^0	0.0
BigWords	28.2 _{4,8}	0.0×10^0	1.9	25.1 _{4,8}	0.0×10^0	1.5	32.5 _{11,1}	0.0×10^0	2.2	32.0 _{6,0}	0.0×10^0	1.9	23.5 _{7,3}	0.0×10^0	1.4	22.8 _{7,7}	0.0×10^0	1.4	22.0 _{0,2}	0.0×10^0	1.1	17.8 _{10,6}	3.1×10^{-268}	0.7
Self	2.2 _{1,9}	3.1×10^{-8}	-0.2	2.4 _{2,3}	1.1×10^{-3}	-0.1	1.0 _{1,6}	7.0×10^{-39}	-0.4	1.3 _{1,8}	8.9×10^{-25}	-0.4	1.4 _{2,4}	7.5×10^{-84}	-0.4	1.7 _{2,6}	2.1×10^{-69}	-0.3	1.4 _{2,5}	2.1×10^{-106}	-0.4	1.9 _{3,2}	4.8×10^{-46}	-0.3
Other	9.6 _{2,3}	2.9×10^{-2}	0.1	8.4 _{3,0}	3.4×10^{-1}	-0.1	5.6 _{4,8}	2.1×10^{-43}	-0.5	5.6 _{4,8}	1.3×10^{-41}	-0.5	8.5 _{3,5}	3.2×10^{-3}	-0.1	8.2 _{3,4}	3.3×10^{-11}	-0.1	7.8 _{5,0}	3.5×10^{-20}	-0.2	6.9 _{4,9}	1.0×10^{-64}	-0.3
Health	2.8 _{2,3}	3.8×10^{-12}	0.2	3.2 _{3,1}	4.0×10^{-18}	0.3	1.9 _{2,0}	1.0×10^0	0.0	1.6 _{2,3}	1.0×10^0	-0.1	2.7 _{2,9}	6.7×10^{-23}	0.2	3.0 _{3,4}	3.3×10^{-52}	0.3	2.1 _{3,0}	2.3×10^{-2}	0.1	1.7 _{4,1}	1.0×10^0	0.0
Wellness	1.1 _{1,5}	1.7×10^{-31}	0.4	1.2 _{1,8}	1.4×10^{-31}	0.4	0.7 _{1,2}	5.6×10^{-8}	0.2	0.8 _{1,5}	1.6×10^{-38}	0.2	0.9 _{1,6}	2.0×10^{-43}	0.3	0.8 _{1,6}	1.3×10^{-37}	0.3	0.5 _{1,4}	2.2×10^{-10}	0.1	0.4 _{2,0}	2.0×10^{-9}	0.0
Behavioral Activation	7.8 _{3,0}	4.0×10^{-128}	0.8	8.9 _{4,5}	4.0×10^{-147}	1.0	6.9 _{3,8}	1.1×10^{-84}	0.7	6.4 _{4,2}	1.4×10^{-54}	0.6	6.5 _{4,2}	2.6×10^{-182}	0.6	5.4 _{4,1}	3.5×10^{-116}	0.4	5.4 _{4,5}	6.3×10^{-93}	0.4	4.9 _{5,9}	6.4×10^{-49}	0.3
Meaning	3.6 _{1,8}	5.0×10^{-84}	0.7	3.4 _{2,1}	7.6×10^{-60}	0.6	2.3 _{1,7}	1.6×10^{-16}	0.3	2.3 _{1,8}	1.9×10^{-12}	0.3	3.4 _{2,4}	1.3×10^{-201}	0.7	3.0 _{2,1}	6.8×10^{-177}	0.6	2.1 _{2,5}	5.9×10^{-29}	0.2	1.8 _{2,8}	1.0×10^{-9}	0.1
Purpose	4.3 _{1,9}	1.1×10^{-124}	0.8	4.7 _{2,7}	9.1×10^{-126}	0.9	2.8 _{2,0}	1.3×10^{-29}	0.4	2.7 _{2,0}	1.3×10^{-20}	0.3	3.7 _{2,7}	2.3×10^{-211}	0.7	3.3 _{2,4}	2.4×10^{-199}	0.6	2.4 _{2,7}	8.3×10^{-48}	0.3	2.1 _{3,1}	9.6×10^{-19}	0.2
Motivation	3.9 _{1,8}	6.7×10^{-124}	0.8	4.3 _{2,6}	5.4×10^{-128}	0.9	2.4 _{1,8}	1.7×10^{-23}	0.3	2.3 _{1,9}	9.7×10^{-17}	0.3	3.2 _{2,5}	1.8×10^{-190}	0.6	2.9 _{2,2}	1.8×10^{-190}	0.6	2.1 _{2,4}	2.9×10^{-43}	0.3	1.8 _{2,9}	9.6×10^{-19}	0.2
Sadness	0.5 _{0,9}	6.2×10^{-5}	0.2	0.5 _{0,9}	2.6×10^{-2}	0.1	0.6 _{1,0}	7.5×10^{-10}	0.2	0.7 _{1,4}	5.8×10^{-12}	0.3	0.7 _{1,4}	8.1×10^{-38}	0.3	0.8 _{1,5}	3.0×10^{-72}	0.3	0.6 _{1,6}	6.7×10^{-30}	0.2	0.5 _{1,7}	7.1×10^{-9}	0.1
Sympathy	0.2 _{0,5}	2.9×10^{-1}	0.1	0.2 _{0,6}	5.9×10^{-1}	0.1	0.3 _{0,7}	1.5×10^{-6}	0.2	0.4 _{1,3}	7.2×10^{-12}	0.3	0.3 _{0,8}	4.4×10^{-11}	0.1	0.4 _{1,0}	9.8×10^{-45}	0.3	0.3 _{1,2}	8.8×10^{-16}	0.2	0.2 _{1,0}	4.2×10^{-6}	0.1
Stress	0.7 _{1,2}	1.2×10^{-2}	-0.1	0.9 _{1,4}	6.5×10^{-1}	-0.1	1.2 _{1,5}	1.0×10^0	0.0	1.0 _{1,5}	1.0×10^0	0.0	1.2 _{1,9}	2.3×10^{-1}	0.1	1.5 _{1,9}	1.5×10^{-10}	0.1	1.4 _{2,3}	1.9×10^{-8}	0.1	1.1 _{2,3}	1.0×10^0	0.0

Table S11. Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset⁷¹. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS by 40.2% while human therapists responds by 23.4% on average ($P = 6.9 \times 10^{-90}$, Cohen’s $d = 0.4$, two-sided Student’s t -test).

Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ($n = 7705$)			GPT-3.5 ($n = 7135$)			Llama2-70b ($n = 7461$)			Llama2-13b ($n = 7464$)			Average ($n = 5446$)	High-quality ($n = 3907$)	Low-quality ($n = 1539$)
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	freq.	freq.
Reflections on Needs	40.2 _{49.0}	6.9×10^{-90}	0.4	35.0 _{47.7}	7.2×10^{-44}	0.3	29.0 _{45.4}	2.6×10^{-11}	0.1	28.8 _{45.3}	9.1×10^{-11}	0.1	23.4 _{43.1}	26.2 _{44.0}	20.6 _{40.5}
Reflections on Emotions	23.1 _{42.2}	1.2×10^{-74}	0.3	25.2 _{43.4}	8.4×10^{-95}	0.4	17.1 _{37.6}	3.1×10^{-24}	0.2	21.3 _{41.0}	4.5×10^{-57}	0.3	10.5 _{32.8}	14.6 _{35.3}	6.3 _{24.3}
Reflections on Values	4.4 _{20.6}	1.0×10^0	0.0	3.5 _{18.4}	3.9×10^{-3}	-0.1	3.7 _{19.0}	4.0×10^{-2}	-0.1	4.1 _{19.8}	6.2×10^{-1}	0.0	4.8 _{22.8}	6.3 _{24.4}	3.3 _{17.9}
Reflections on Consequences	22.3 _{41.7}	1.0×10^0	0.0	21.4 _{41.0}	1.0×10^0	0.0	19.6 _{39.7}	3.2×10^{-2}	-0.1	21.2 _{40.9}	1.0×10^0	0.0	21.8 _{40.1}	17.9 _{38.4}	25.7 _{43.7}
Reflections on Conflicts	4.1 _{19.8}	2.9×10^{-9}	-0.1	3.5 _{18.5}	9.1×10^{-14}	-0.1	6.3 _{24.2}	1.0×10^0	0.0	6.6 _{24.8}	1.0×10^0	0.0	6.6 _{26.1}	8.3 _{27.6}	5.0 _{21.8}
Reflections on Strengths	33.5 _{47.2}	4.2×10^{-225}	0.6	27.4 _{44.6}	1.4×10^{-135}	0.5	16.8 _{37.4}	1.9×10^{-30}	0.2	15.6 _{36.3}	5.8×10^{-22}	0.2	9.5 _{31.3}	12.8 _{33.5}	6.2 _{24.2}
Questions on Experiences	36.7 _{48.2}	6.0×10^{-16}	-0.1	29.5 _{45.6}	6.2×10^{-63}	-0.3	57.6 _{49.4}	6.7×10^{-52}	0.3	56.6 _{49.6}	1.5×10^{-44}	0.3	44.0 _{49.9}	48.9 _{50.0}	39.1 _{48.8}
Questions on Perspectives	7.6 _{26.6}	2.3×10^{-21}	0.2	3.8 _{19.1}	1.0×10^0	0.0	12.1 _{32.6}	1.4×10^{-65}	0.3	11.3 _{31.7}	2.2×10^{-56}	0.3	3.5 _{19.9}	4.9 _{21.7}	2.0 _{14.1}
Questions on Emotions	3.0 _{17.1}	4.7×10^{-84}	-0.3	6.1 _{24.0}	3.7×10^{-27}	-0.2	12.5 _{33.0}	1.0×10^0	0.0	18.1 _{38.5}	7.7×10^{-22}	0.2	11.7 _{33.0}	13.3 _{34.0}	10.1 _{30.2}
Problem-Solving	53.4 _{49.9}	1.7×10^{-229}	0.6	60.1 _{49.0}	0.0×10^0	0.7	33.7 _{47.3}	6.9×10^{-21}	0.2	24.3 _{42.9}	5.4×10^{-1}	0.0	25.9 _{42.6}	21.2 _{40.9}	30.5 _{46.0}
Planning	47.9 _{50.0}	3.6×10^{-306}	0.7	53.3 _{49.9}	0.0×10^0	0.8	32.5 _{46.8}	7.2×10^{-87}	0.4	26.8 _{44.3}	5.8×10^{-38}	0.2	17.1 _{37.6}	16.9 _{37.5}	17.2 _{37.8}
Normalizing	43.5 _{49.6}	1.9×10^{-171}	0.5	47.6 _{49.9}	8.9×10^{-228}	0.6	16.6 _{37.2}	1.4×10^{-7}	-0.1	19.4 _{39.5}	1.0×10^0	0.0	20.5 _{39.5}	17.8 _{38.3}	23.2 _{42.2}
Psychoeducation	19.1 _{39.3}	1.6×10^{-79}	0.3	18.7 _{39.0}	9.5×10^{-74}	0.3	12.8 _{33.5}	1.2×10^{-21}	0.2	7.4 _{26.3}	1.0×10^0	0.0	7.6 _{24.8}	5.2 _{22.3}	9.9 _{29.8}

Table S12. Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset⁷¹ for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS by 33.5% and GPT-4 with full conversation simulation responds by 41.9% when compared with the human therapists responds by 23.4% on average from Table S11. (Single: $P = 2.6 \times 10^{-14}$, Cohen’s $d = 0.2$; two-sided student t-test) (Full: $P = 8.1 \times 10^{-99}$, Cohen’s $d = 0.4$; two-sided student t-test)

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4 ($n = 6158$)			GPT-3.5 ($n = 6144$)			Llama2-70b ($n = 6174$)			Llama2-13b ($n = 6282$)			GPT-4 ($n = 1547$)			GPT-3.5 ($n = 991$)			Llama2-70b ($n = 1287$)			Llama2-13b ($n = 1182$)		
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d
Reflections on Needs	33.5 _{47.2}	2.6×10^{-14}	0.2	39.6 _{48.9}	4.5×10^{-25}	0.4	19.6 _{39.7}	4.9×10^{-2}	-0.1	19.5 _{39.7}	6.2×10^{-2}	-0.1	41.9 _{49.3}	8.1×10^{-99}	0.4	34.3 _{47.5}	1.3×10^{-36}	0.2	30.9 _{46.2}	2.2×10^{-18}	0.2	30.6 _{46.1}	6.6×10^{-17}	0.2
Reflections on Emotions	10.0 _{29.9}	1.0×10^0	0.0	13.5 _{34.2}	9.4×10^{-2}	0.1	9.9 _{29.9}	1.0×10^0	0.0	16.1 _{36.7}	2.4×10^{-6}	0.2	26.4 _{44.1}	5.5×10^{-103}	0.4	27.1 _{44.5}	8.8×10^{-111}	0.4	18.6 _{38.9}	3.0×10^{-32}	0.2	22.3 _{41.6}	1.5×10^{-62}	0.3
Reflections on Values	5.1 _{22.0}	1.0×10^0	0.0	4.3 _{20.4}	1.0×10^0	0.0	2.7 _{16.3}	2.2×10^{-2}	-0.1	3.1 _{17.4}	2.0×10^{-1}	-0.1	4.3 _{20.2}	1.0×10^0	0.0	3.4 _{18.0}	1.6×10^{-3}	-0.1	4.0 _{19.5}	3.3×10^{-1}	0.0	4.3 _{20.2}	1.0×10^0	0.0
Reflections on Consequences	10.9 _{31.2}	1.3×10^{-21}	-0.3	12.0 _{32.5}	5.6×10^{-12}	-0.3	5.9 _{23.6}	7.3×10^{-41}	-0.4	8.5 _{28.0}	7.1×10^{-26}	-0.3	25.2 _{43.4}	1.6×10^{-4}	0.1	22.9 _{42.0}	1.0×10^0	0.0	22.5 _{41.8}	1.0×10^0	0.0	23.6 _{42.5}	2.6×10^{-1}	0.0
Reflections on Conflicts	2.0 _{14.0}	2.5×10^{-10}	-0.2	1.8 _{13.4}	1.7×10^{-7}	-0.2	2.3 _{15.1}	1.5×10^{-7}	-0.2	3.9 _{19.3}	7.9×10^{-3}	-0.1	4.6 _{21.0}	5.3×10^{-5}	-0.1	3.8 _{19.2}	3.2×10^{-10}	-0.1	7.1 _{25.6}	1.0×10^0	0.0	7.1 _{25.7}	1.0×10^0	0.0
Reflections on Strengths	70.4 _{45.7}	0.0×10^0	1.7	56.5 _{49.6}	5.5×10^{-300}	1.4	35.3 _{47.8}	3.4×10^{-118}	0.7	31.6 _{46.5}	4.3×10^{-85}	0.6	24.3 _{42.9}	4.5×10^{-94}	0.4	22.7 _{41.9}	6.2×10^{-78}	0.4	13.0 _{33.6}	1.4×10^{-7}	0.1	12.6 _{33.2}	5.2×10^{-6}	0.1
Questions on Experiences	31.5 _{46.5}	2.3×10^{-17}	-0.3	21.6 _{41.2}	5.9×10^{-39}	-0.5	38.3 _{48.6}	2.8×10^{-3}	-0.1	41.1 _{49.2}	9.3×10^{-1}	-0.1	38.0 _{48.5}	8.5×10^{-10}	-0.1	30.7 _{46.1}	2.4×10^{-48}	-0.3	61.6 _{48.6}	2.0×10^{-80}	0.4	59.5 _{49.1}	1.3×10^{-62}	0.3
Questions on Perspectives	4.5 _{20.8}	9.1×10^{-1}	0.1	2.7 _{16.3}	1.0×10^0	0.0	5.5 _{22.8}	1.7×10^{-2}	0.1	5.4 _{22.6}	4.0×10^{-2}	0.1	8.4 _{27.8}	2.4×10^{-26}	0.2	4.0 _{19.5}	1.0×10^0	0.0	13.5 _{34.2}	2.9×10^{-78}	0.4	12.4 _{33.0}	5.0×10^{-66}	0.3
Questions on Emotions	1.5 _{12.1}	1.2×10^{-31}	-0.3	3.0 _{17.1}	8.2×10^{-15}	-0.3	3.1 _{17.4}	1.8×10^{-18}	-0.3	8.5 _{28.0}	2.7×10^{-2}	-0.1	3.4 _{18.2}	1.9×10^{-63}	-0.3	6.6 _{24.8}	3.5×10^{-20}	-0.2	14.4 _{35.1}	2.9×10^{-4}	0.1	19.9 _{39.9}	7.3×10^{-32}	0.2
Problem-Solving	40.5 _{49.1}	2.1×10^{-29}	0.3	58.5 _{49.3}	3.6×10^{-99}	0.7	51.4 _{50.0}	1.4×10^{-74}	0.6	34.6 _{47.6}	5.4×10^{-9}	0.2	56.6 _{49.6}	1.1×10^{-262}	0.7	60.4 _{48.9}	0.0×10^0	0.7	30.0 _{45.8}	7.6×10^{-6}	0.1	22.4 _{41.7}	9.8×10^{-5}	-0.1
Planning	66.3 _{47.3}	0.0×10^0	1.2	73.2 _{44.3}	0.0×10^0	1.4	61.6 _{48.7}	2.6×10^{-258}	1.1	42.0 _{49.4}	4.2×10^{-81}	0.6	43.3 _{49.6}	1.8×10^{-211}	0.6	50.1 _{50.0}	0.0×10^0	0.7	26.4 _{44.1}	5.3×10^{-33}	0.2	23.9 _{42.7}	7.0×10^{-19}	0.2
Normalizing	61.9 _{48.6}	7.1×10^{-240}	1.0	71.2 _{45.3}	6.9×10^{-262}	1.3	23.9 _{42.7}	7.7×10^{-2}	0.1	23.8 _{42.6}	1.4×10^{-1}	0.1	38.9 _{48.7}	6.0×10^{-105}	0.4	43.8 _{49.6}	3.8×10^{-163}	0.5	15.1 _{35.8}	1.1×10^{-13}	-0.1	18.5 _{38.9}	8.9×10^{-2}	-0.1
Psychoeducation	23.4 _{42.4}	2.8×10^{-74}	0.5	33.4 _{47.2}	1.4×10^{-135}	0.9	29.4 _{45.6}	7.5×10^{-117}	0.7	17.8 _{38.2}	3.1×10^{-29}	0.4	18.0 _{38.4}	3.0×10^{-64}	0.3	16.3 _{37.0}	5.3×10^{-48}	0.3	9.4 _{29.2}	3.8×10^{-3}	0.1	5.5 _{22.8}	3.9×10^{-5}	-0.1

Table S13. Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset⁷¹. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS at 3.3 turns earlier than average human therapists responds. (GPT-4: 2.9 turn, Average human therapists: 6.2 turn; $P = 1.0 \times 10^{-19}$, Cohen’s $d = 0.8$, two-sided Student’s t -test).

Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ($n = 7705$)			GPT-3.5 ($n = 7135$)			Llama2-70b ($n = 7461$)			Llama2-13b ($n = 7464$)			Average ($n = 5446$)	High-quality ($n = 3907$)	Low-quality ($n = 1539$)
	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	turn	turn
Reflections on Needs	2.9 _{3,3}	1.0×10^{-19}	0.8	6.1 _{6,7}	1.0×10^0	0.0	5.3 _{5,5}	4.8×10^{-1}	0.2	6.1 _{6,4}	1.0×10^0	0.0	6.2 _{5,5}	4.4 _{4,5}	8.1 _{6,2}
Reflections on Emotions	5.8 _{4,9}	1.0×10^0	0.1	6.0 _{5,3}	1.0×10^0	0.1	6.4 _{5,1}	1.0×10^0	0.0	7.0 _{5,8}	1.0×10^0	-0.1	6.3 _{4,9}	5.2 _{4,8}	7.4 _{5,0}
Reflections on Values	9.0 _{5,9}	1.0×10^0	0.2	9.3 _{6,4}	1.0×10^0	0.1	10.3 _{6,0}	1.0×10^0	-0.1	10.3 _{5,8}	1.0×10^0	-0.1	9.9 _{5,5}	8.8 _{5,3}	11.0 _{5,8}
Reflections on Consequences	7.0 _{4,8}	1.0×10^0	0.0	7.1 _{5,7}	1.0×10^0	0.0	6.2 _{4,9}	1.0×10^0	0.2	7.3 _{5,7}	1.0×10^0	-0.1	7.0 _{5,0}	6.2 _{4,6}	7.8 _{5,4}
Reflections on Conflicts	8.2 _{4,6}	5.8×10^{-2}	0.4	8.8 _{5,3}	1.0×10^0	0.2	8.7 _{5,3}	4.2×10^{-1}	0.2	8.1 _{5,3}	2.3×10^{-2}	0.4	10.0 _{5,2}	8.2 _{4,7}	11.8 _{5,6}
Reflections on Strengths	7.8 _{5,1}	4.6×10^{-5}	0.4	8.9 _{6,2}	3.9×10^{-1}	0.2	9.1 _{6,3}	1.0×10^0	0.2	10.1 _{6,8}	1.0×10^0	0.0	10.2 _{6,1}	8.4 _{6,0}	11.9 _{5,7}
Questions on Experiences	2.7 _{2,8}	1.0×10^0	0.1	4.1 _{5,7}	2.9×10^{-2}	-0.3	3.1 _{3,1}	1.0×10^0	-0.1	4.0 _{4,1}	1.8×10^{-3}	-0.3	2.9 _{2,5}	2.2 _{2,1}	3.6 _{2,9}
Questions on Perspectives	7.7 _{4,1}	1.4×10^{-13}	1.1	7.8 _{4,5}	2.9×10^{-9}	1.0	8.9 _{5,4}	8.1×10^{-7}	0.7	8.4 _{5,2}	6.1×10^{-9}	0.8	12.6 _{5,1}	10.0 _{4,7}	15.3 _{4,2}
Questions on Emotions	6.4 _{5,6}	1.0×10^0	0.1	7.7 _{6,0}	1.0×10^0	-0.1	5.0 _{5,2}	1.0×10^{-2}	0.3	4.4 _{4,4}	2.9×10^{-6}	0.5	7.0 _{6,4}	6.7 _{6,3}	7.3 _{6,5}
Problem-Solving	7.9 _{5,2}	1.0×10^0	0.1	7.0 _{5,6}	4.7×10^{-2}	0.2	8.7 _{5,4}	1.0×10^0	-0.1	9.5 _{6,2}	1.9×10^{-1}	-0.2	8.3 _{5,4}	8.7 _{5,6}	8.0 _{5,1}
Planning	9.7 _{5,7}	2.7×10^{-4}	0.4	9.3 _{5,6}	8.6×10^{-6}	0.4	9.7 _{5,7}	2.7×10^{-4}	0.4	10.5 _{6,1}	1.3×10^{-1}	0.2	11.9 _{6,5}	11.5 _{6,8}	12.4 _{6,0}
Normalizing	8.0 _{5,8}	4.5×10^{-1}	-0.2	7.3 _{6,1}	1.0×10^0	-0.1	9.5 _{6,6}	7.5×10^{-5}	-0.4	8.8 _{6,6}	1.4×10^{-2}	-0.3	6.9 _{6,2}	6.2 _{6,3}	7.7 _{6,0}
Psychoeducation	11.0 _{5,9}	2.1×10^{-2}	-0.4	10.1 _{5,9}	9.8×10^{-1}	-0.2	12.1 _{5,7}	9.3×10^{-6}	-0.6	12.8 _{6,2}	2.2×10^{-6}	-0.7	8.9 _{5,5}	8.5 _{5,6}	9.2 _{5,5}

Table S14. Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset⁷¹ for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS at 4.0 turn which is 2.2 turns earlier than average human therapists. Similarly, GPT-4 with full conversation simulation responds at 1.8 turn which is 4.4 turns earlier than average human therapists. (Single: 4.0 turn, Average human therapists: 6.2 turn from Table S13; $P = 6.4 \times 10^{-6}$, Cohen’s $d = 0.5$, two-sided Student’s t -test) (Full: 1.8 turn, Average human therapists: 6.2 turn from Table S13; $P = 4.7 \times 10^{-21}$, Cohen’s $d = 1.0$).

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)														
	GPT-4 ($n = 6158$)			GPT-3.5 ($n = 6144$)			Llama2-70b ($n = 6174$)			Llama2-13b ($n = 6282$)			GPT-4 ($n = 1547$)			GPT-3.5 ($n = 991$)			Llama2-70b ($n = 1287$)			Llama2-13b ($n = 1182$)					
	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d	turn	p	d
Reflections on Needs	4.0 _{3,0}	6.4×10^{-6}	0.5	9.3 _{7,4}	1.3×10^{-5}	-0.5	9.3 _{3,3}	1.3×10^{-6}	-0.6	11.5 _{6,3}	2.1×10^{-15}	-0.9	1.8 _{3,2}	4.7×10^{-21}	1.0	2.8 _{3,8}	7.5×10^{-12}	0.7	2.4 _{3,4}	5.3×10^{-16}	0.8	2.6 _{3,1}	1.5×10^{-15}	0.8	4.2 _{4,0}	9.6×10^{-5}	0.5
Reflections on Emotions	8.2 _{5,0}	3.5×10^{-2}	-0.4	9.5 _{8,8}	6.9×10^{-5}	-0.6	9.5 _{4,7}	2.4×10^{-5}	-0.7	11.7 _{5,2}	3.6×10^{-16}	-1.1	4.5 _{4,3}	2.9×10^{-3}	0.4	4.3 _{4,0}	3.8×10^{-4}	0.4	5.0 _{4,7}	1.7×10^{-1}	0.3	4.2 _{4,0}	9.6×10^{-5}	0.5	9.2 _{5,9}	1.0×10^0	0.1
Reflections on Values	8.4 _{5,3}	1.0×10^0	0.3	8.9 _{9,2}	1.0×10^0	0.2	14.2 _{4,9}	4.4×10^{-3}	-0.8	14.4 _{4,5}	1.1×10^{-3}	-0.8	9.4 _{6,2}	1.0×10^0	0.1	9.6 _{8,8}	1.0×10^0	0.1	9.2 _{5,9}	1.0×10^0	0.1	9.2 _{5,5}	1.0×10^0	0.1	9.2 _{5,5}	1.0×10^0	0.1
Reflections on Consequences	8.6 _{4,4}	4.4×10^{-2}	-0.3	9.9 _{6,4}	3.7×10^{-4}	-0.5	10.1 _{4,7}	6.0×10^{-4}	-0.6	13.0 _{8,3}	2.4×10^{-14}	-1.2	6.0 _{4,8}	6.0×10^{-1}	0.2	5.8 _{4,9}	2.0×10^{-1}	0.2	5.1 _{4,3}	5.1×10^{-4}	0.4	5.5 _{4,4}	1.3×10^{-2}	0.3	9.9 _{5,8}	1.0×10^0	0.1
Reflections on Conflicts	9.2 _{3,3}	1.0×10^0	0.2	8.8 _{5,2}	1.0×10^0	0.2	10.3 _{5,0}	1.0×10^0	-0.1	11.8 _{4,9}	1.0×10^0	-0.3	8.0 _{4,8}	4.1×10^{-2}	0.4	8.8 _{5,3}	1.0×10^0	0.2	8.4 _{3,3}	1.6×10^{-1}	0.3	7.2 _{5,1}	1.7×10^{-4}	0.5	9.2 _{5,9}	1.0×10^0	0.1
Reflections on Strengths	8.4 _{4,8}	2.1×10^{-2}	0.3	11.5 _{6,2}	5.3×10^{-1}	-0.2	11.2 _{5,8}	1.0×10^0	-0.2	12.4 _{6,3}	2.0×10^{-2}	-0.4	7.3 _{5,3}	3.0×10^{-5}	0.5	6.6 _{6,2}	5.5×10^{-8}	0.6	7.0 _{6,1}	2.2×10^{-5}	0.5	8.3 _{6,6}	9.8×10^{-2}	0.3	9.2 _{5,9}	1.0×10^0	0.1
Questions on Experiences	3.9 _{2,6}	7.9×10^{-4}	-0.4	8.1 _{7,1}	3.2×10^{-21}	-1.1	4.9 _{3,1}	3.9×10^{-12}	-0.7	6.5 _{4,2}	1.5×10^{-25}	-1.1	1.5 _{2,5}	3.9×10^{-8}	0.6	1.6 _{2,4}	1.7×10^{-7}	0.5	1.4 _{1,9}	1.9×10^{-11}	0.7	1.5 _{1,9}	1.8×10^{-9}	0.6	9.2 _{5,9}	1.0×10^0	0.1
Questions on Perspectives	8.3 _{3,9}	3.5×10^{-4}	0.9	8.0 _{7,9}	5.2×10^{-1}	0.9	13.0 _{4,3}	1.0×10^0	-0.1	13.5 _{4,8}	1.0×10^0	-0.2	7.5 _{4,1}	1.3×10^{-12}	1.1	7.8 _{4,2}	2.3×10^{-9}	1.0	7.7 _{5,1}	4.1×10^{-11}	1.0	7.4 _{7,1}	3.5×10^{-14}	1.1	9.2 _{5,9}	1.0×10^0	0.1
Questions on Emotions	6.0 _{5,9}	1.0×10^0	0.2	8.1 _{6,5}	1.0×10^0	-0.2	7.6 _{6,1}	1.0×10^0	-0.1	5.5 _{3,3}	5.3×10^{-1}	0.3	6.5 _{5,5}	1.0×10^0	0.1	7.6 _{6,0}	1.0×10^0	-0.1	4.5 _{4,9}	5.5×10^{-4}	0.4	3.9 _{4,7}	1.1×10^{-6}	0.6	9.2 _{5,9}	1.0×10^0	0.1
Problem-Solving	10.6 _{4,8}	8.3×10^{-5}	-0.4	10.2 _{5,7}	8.3×10^{-3}	-0.3	11.2 _{4,8}	1.0×10^{-7}	-0.6	13.1 _{5,5}	2.0×10^{-15}	-0.9	5.3 _{4,0}	2.6×10^{-9}	0.6	4.2 _{3,6}	3.4×10^{-17}	0.9	6.1 _{4,8}	1.9×10^{-4}	0.4	6.6 _{5,1}	8.9×10^{-3}	0.3	9.2 _{5,9}	1.0×10^0	0.1
Planning	11.4 _{5,5}	1.0×10^0	0.1	12.3 _{5,3}	1.0×10^0	-0.1	10.9 _{5,4}	1.0×10^0	0.2	12.6 _{6,0}	1.0×10^0	-0.1	8.0 _{5,4}	3.4×10^{-9}	0.7	6.9 _{4,6}	5.9×10^{-16}	0.9	8.4 _{5,8}	7.4×10^{-7}	0.6	8.6 _{5,7}	3.0×10^{-6}	0.5	9.2 _{5,9}	1.0×10^0	0.1
Normalizing	10.6 _{5,4}	8.6×10^{-9}	-0.6	10.9 _{5,8}	1.4×10^{-8}	-0.7	12.9 _{6,2}	2.1×10^{-15}	-1.0	12.3 _{6,5}	1.1×10^{-10}	-0.9	5.5 _{5,1}	1.5×10^{-1}	0.2	4.6 _{4,8}	2.8×10^{-4}	0.4	7.1 _{5,8}	1.0×10^0	0.0	7.0 _{5,8}	1.0×10^0	0.0	9.2 _{5,9}	1.0×10^0	0.1
Psychoeducation	13.1 _{5,2}	2.3×10^{-6}	-0.8	11.7 _{5,8}	1.2×10^{-2}	-0.5	13.3 _{5,3}	1.0×10^{-8}	-0.8	15.6 _{5,4}	8.1×10^{-14}	-1.2	9.6 _{5,9}	1.0×10^0	-0.1	9.2 _{5,8}	1.0×10^0	-0.1	10.7 _{5,7}	2.0×10^{-1}	-0.3	10.0 _{5,6}	1.0×10^0	-0.2	9.2 _{5,9}	1.0×10^0	0.1

Table S15. Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset⁷¹. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13 \times 6 = 78$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON NEEDS by 10.2% and human therapists responds by 27.7% on average when client expresses CHANGING UNHEALTHY BEHAVIOR. ($P = 1.9 \times 10^{-46}$, Cohen’s $d = -0.5$, two-sided Student’s t -test).

Therapist Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			Average	High-quality	Low-quality
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	freq.	freq.
Client behavior: Changing unhealthy behavior															
n	2552			2193			2349			2109			1553	1198	355
Reflections on Needs	10.2 _{30.3}	1.9×10^{-46}	-0.5	8.2 _{27.4}	2.9×10^{-56}	-0.5	6.0 _{23.8}	3.6×10^{-78}	-0.6	5.3 _{22.4}	1.7×10^{-79}	-0.7	27.7 _{45.5}	30.6 _{46.1}	24.8 _{43.2}
Reflections on Emotions	2.7 _{16.2}	5.4×10^{-21}	-0.3	2.8 _{16.6}	5.9×10^{-18}	-0.3	2.1 _{14.4}	3.9×10^{-24}	-0.3	3.7 _{18.8}	7.0×10^{-13}	-0.3	10.4 _{33.3}	14.6 _{35.3}	6.2 _{24.1}
Reflections on Values	2.0 _{14.1}	3.0×10^{-4}	-0.1	0.1 _{3.0}	8.2×10^{-18}	-0.3	1.0 _{9.8}	4.7×10^{-10}	-0.2	0.9 _{9.2}	4.4×10^{-10}	-0.2	4.8 _{23.7}	7.0 _{25.5}	2.5 _{15.7}
Reflections on Consequences	4.5 _{20.8}	6.6×10^{-108}	-0.7	2.1 _{14.5}	3.9×10^{-133}	-0.9	1.8 _{13.3}	5.2×10^{-148}	-0.9	2.8 _{16.4}	5.6×10^{-119}	-0.8	27.3 _{42.3}	19.9 _{39.9}	34.6 _{47.7}
Reflections on Conflicts	0.5 _{7.4}	5.3×10^{-28}	-0.4	0.3 _{5.6}	2.6×10^{-27}	-0.4	0.6 _{7.7}	2.2×10^{-25}	-0.4	0.9 _{9.5}	7.3×10^{-20}	-0.3	6.9 _{26.6}	8.3 _{27.5}	5.6 _{23.1}
Reflections on Strengths	31.6 _{46.5}	5.7×10^{-58}	0.5	20.1 _{40.1}	1.9×10^{-16}	0.3	14.8 _{35.5}	7.5×10^{-5}	0.2	12.0 _{32.6}	9.2×10^{-1}	0.1	9.3 _{32.4}	14.1 _{34.8}	4.5 _{20.8}
Questions on Experiences	6.5 _{24.7}	1.4×10^{-129}	-0.8	1.2 _{11.0}	6.8×10^{-197}	-1.1	6.3 _{24.2}	8.0×10^{-126}	-0.8	6.2 _{24.1}	7.0×10^{-117}	-0.8	35.6 _{48.6}	40.2 _{49.1}	31.0 _{46.3}
Questions on Perspectives	1.1 _{10.4}	2.6×10^{-5}	-0.2	0.0 _{2.1}	7.7×10^{-14}	-0.3	1.7 _{12.8}	2.8×10^{-2}	-0.1	1.0 _{9.9}	3.9×10^{-5}	-0.2	3.5 _{20.2}	4.8 _{21.5}	2.3 _{14.9}
Questions on Emotions	0.2 _{4.4}	6.3×10^{-49}	-0.5	0.1 _{3.7}	3.6×10^{-43}	-0.5	0.4 _{6.5}	1.4×10^{-41}	-0.5	0.2 _{4.4}	8.3×10^{-41}	-0.5	9.6 _{30.7}	11.3 _{31.6}	7.9 _{27.0}
Problem-Solving	14.0 _{34.7}	1.5×10^{-29}	-0.4	14.4 _{35.1}	1.0×10^{-25}	-0.4	20.3 _{40.2}	2.1×10^{-7}	-0.2	10.3 _{30.4}	3.5×10^{-47}	-0.5	28.4 _{42.7}	20.4 _{40.3}	36.3 _{48.2}
Planning	19.5 _{39.6}	8.7×10^{-5}	0.2	15.3 _{36.0}	1.0×10^0	0.0	22.9 _{42.0}	2.5×10^{-11}	0.2	11.9 _{32.4}	1.0×10^0	-0.1	13.7 _{32.6}	10.7 _{30.9}	16.6 _{37.3}
Normalizing	19.3 _{39.5}	6.3×10^{-4}	0.1	15.8 _{36.5}	1.0×10^0	0.1	6.5 _{24.7}	4.4×10^{-14}	-0.3	5.1 _{22.0}	7.8×10^{-21}	-0.3	14.0 _{32.6}	10.5 _{30.7}	17.5 _{38.0}
Psychoeducation	3.4 _{18.0}	6.5×10^{-2}	-0.1	3.0 _{17.0}	5.1×10^{-3}	-0.1	8.3 _{27.5}	3.9×10^{-2}	0.1	4.1 _{19.9}	1.0×10^0	-0.1	5.4 _{20.5}	3.5 _{18.4}	7.3 _{26.1}
Client behavior: Sustaining unhealthy behavior															
n	428			402			485			462			389	275	114
Reflections on Needs	5.4 _{22.6}	1.5×10^{-16}	-0.6	2.2 _{14.8}	7.6×10^{-23}	-0.8	7.0 _{25.6}	9.8×10^{-15}	-0.6	6.1 _{23.9}	4.4×10^{-16}	-0.6	27.6 _{45.3}	30.5 _{46.1}	24.6 _{43.2}
Reflections on Emotions	0.5 _{6.8}	1.1×10^{-9}	-0.5	0.2 _{5.0}	1.1×10^{-9}	-0.5	2.5 _{15.5}	5.4×10^{-6}	-0.4	3.7 _{18.8}	8.2×10^{-4}	-0.3	11.7 _{33.2}	13.8 _{34.6}	9.6 _{29.7}
Reflections on Values	1.2 _{10.8}	2.0×10^{-3}	-0.3	0.0 _{0.0}	5.6×10^{-6}	-0.4	0.6 _{7.8}	2.9×10^{-5}	-0.3	0.9 _{9.3}	2.2×10^{-4}	-0.3	6.8 _{25.0}	6.5 _{24.8}	7.0 _{25.7}
Reflections on Consequences	4.7 _{21.1}	2.7×10^{-31}	-0.9	1.0 _{9.9}	1.5×10^{-41}	-1.0	4.5 _{20.8}	5.3×10^{-35}	-0.9	5.4 _{22.6}	5.4×10^{-31}	-0.9	36.1 _{46.8}	26.5 _{44.2}	45.6 _{50.0}
Reflections on Conflicts	0.9 _{9.6}	2.0×10^{-5}	-0.4	1.2 _{11.1}	1.9×10^{-4}	-0.3	1.9 _{13.5}	5.9×10^{-4}	-0.3	0.9 _{9.3}	5.1×10^{-6}	-0.4	8.2 _{27.1}	7.6 _{26.6}	8.8 _{28.4}
Reflections on Strengths	3.5 _{18.4}	1.0×10^0	-0.1	1.2 _{11.1}	1.9×10^{-1}	-0.2	4.3 _{20.4}	1.0×10^0	0.0	3.7 _{18.8}	1.0×10^0	-0.1	5.2 _{23.6}	6.9 _{25.4}	3.5 _{18.5}
Questions on Experiences	3.0 _{17.2}	1.8×10^{-39}	-1.0	0.5 _{7.0}	5.0×10^{-46}	-1.1	6.0 _{23.7}	5.6×10^{-34}	-0.9	5.4 _{22.6}	2.2×10^{-34}	-0.9	39.3 _{49.4}	46.2 _{49.9}	32.5 _{47.0}
Questions on Perspectives	0.7 _{8.4}	5.0×10^{-1}	-0.2	0.0 _{0.0}	2.3×10^{-2}	-0.3	1.4 _{11.9}	1.0×10^0	-0.1	1.7 _{13.1}	1.0×10^0	-0.1	3.5 _{19.3}	4.4 _{20.5}	2.6 _{16.1}
Questions on Emotions	0.0 _{0.0}	1.8×10^{-8}	-0.4	0.0 _{0.0}	6.3×10^{-8}	-0.4	0.0 _{0.0}	1.2×10^{-9}	-0.5	0.2 _{4.7}	1.8×10^{-8}	-0.4	9.2 _{29.7}	10.5 _{30.8}	7.9 _{27.1}
Problem-Solving	2.8 _{16.5}	1.1×10^{-19}	-0.7	1.5 _{12.1}	1.9×10^{-22}	-0.7	12.2 _{32.7}	3.2×10^{-4}	-0.3	6.9 _{25.4}	4.0×10^{-11}	-0.5	23.6 _{40.3}	15.6 _{36.4}	31.6 _{46.7}
Planning	1.9 _{13.6}	1.1×10^{-1}	-0.2	0.5 _{7.0}	2.4×10^{-4}	-0.3	8.9 _{28.5}	1.0×10^0	0.1	4.1 _{19.9}	1.0×10^0	-0.1	5.9 _{22.1}	4.0 _{19.6}	7.9 _{27.1}
Normalizing	2.1 _{14.4}	1.9×10^{-9}	-0.5	0.7 _{8.6}	2.0×10^{-12}	-0.6	2.3 _{14.9}	4.5×10^{-10}	-0.5	1.7 _{13.1}	4.4×10^{-11}	-0.5	14.0 _{33.2}	10.5 _{30.8}	17.5 _{38.2}
Psychoeducation	0.2 _{4.8}	5.6×10^{-11}	-0.5	0.5 _{7.0}	2.1×10^{-9}	-0.5	3.3 _{17.9}	1.4×10^{-3}	-0.3	1.5 _{12.2}	2.6×10^{-7}	-0.4	9.8 _{26.7}	4.7 _{21.3}	14.9 _{35.8}
Client behavior: Sharing positive emotions															
n	815			596			389			437			294	226	68
Reflections on Needs	4.0 _{19.7}	7.2×10^{-11}	-0.5	10.7 _{31.0}	6.0×10^{-1}	-0.2	2.8 _{16.6}	3.3×10^{-8}	-0.5	3.0 _{17.0}	8.8×10^{-9}	-0.5	17.3 _{40.9}	24.3 _{43.0}	10.3 _{30.6}
Reflections on Emotions	2.3 _{15.1}	2.2×10^{-2}	-0.2	3.4 _{18.0}	1.0×10^0	-0.2	0.3 _{5.1}	9.7×10^{-4}	-0.3	0.9 _{9.5}	4.4×10^{-3}	-0.3	7.5 _{32.0}	15.0 _{35.8}	0.0 _{0.0}
Reflections on Values	0.9 _{9.2}	3.9×10^{-1}	-0.2	0.2 _{4.1}	1.9×10^{-2}	-0.3	0.8 _{8.8}	1.0×10^0	-0.2	0.2 _{4.8}	1.6×10^{-1}	-0.2	3.4 _{20.6}	5.3 _{22.5}	1.5 _{12.1}
Reflections on Consequences	0.6 _{7.8}	4.0×10^{-29}	-0.8	0.7 _{8.2}	1.0×10^{-21}	-0.7	0.5 _{7.2}	3.1×10^{-15}	-0.7	0.9 _{9.5}	2.0×10^{-15}	-0.7	15.8 _{34.0}	11.1 _{31.4}	20.6 _{40.7}
Reflections on Conflicts	0.0 _{0.0}	2.2×10^{-4}	-0.3	0.0 _{0.0}	4.9×10^{-3}	-0.3	0.0 _{0.0}	9.5×10^{-2}	-0.3	0.7 _{8.3}	1.0×10^0	-0.2	3.4 _{20.6}	5.3 _{22.5}	1.5 _{12.1}
Reflections on Strengths	62.2 _{48.5}	1.4×10^{-55}	1.2	42.8 _{49.5}	3.6×10^{-21}	0.7	14.4 _{35.2}	1.0×10^0	0.1	17.2 _{37.7}	8.4×10^{-1}	0.2	10.2 _{34.0}	15.9 _{36.7}	4.4 _{20.7}
Questions on Experiences	0.9 _{9.2}	3.1×10^{-59}	-1.2	0.2 _{4.1}	4.9×10^{-49}	-1.1	4.1 _{19.9}	3.2×10^{-21}	-0.8	6.6 _{24.9}	2.0×10^{-17}	-0.7	31.6 _{47.6}	36.7 _{48.3}	26.5 _{44.4}
Questions on Perspectives	0.1 _{3.5}	1.0×10^0	-0.2	0.2 _{4.1}	1.0×10^0	-0.1	0.3 _{5.1}	1.0×10^0	-0.1	0.5 _{6.8}	1.0×10^0	-0.1	1.3 _{14.2}	2.7 _{16.1}	0.0 _{0.0}
Questions on Emotions	0.0 _{0.0}	1.8×10^{-11}	-0.5	0.3 _{5.8}	6.0×10^{-7}	-0.4	0.0 _{0.0}	3.0×10^{-5}	-0.4	0.2 _{4.8}	2.4×10^{-5}	-0.4	10.6 _{33.6}	10.6 _{30.9}	4.4 _{20.7}
Problem-Solving	7.7 _{26.7}	5.2×10^{-1}	-0.2	10.6 _{30.8}	1.0×10^0	-0.1	7.2 _{25.9}	8.2×10^{-1}	-0.2	3.4 _{18.2}	6.1×10^{-5}	-0.4	13.0 _{33.6}	12.8 _{33.5}	13.2 _{34.1}
Planning	43.6 _{49.6}	1.1×10^{-27}	0.8	31.0 _{46.3}	1.0×10^{-12}	0.6	18.5 _{38.9}	8.1×10^{-3}	0.3	12.4 _{32.9}	1.0×10^0	0.1	8.2 _{26.9}	7.5 _{26.4}	8.8 _{28.6}
Normalizing	58.5 _{49.3}	1.5×10^{-25}	0.8	38.9 _{48.8}	8.6×10^{-5}	0.3	8.2 _{27.5}	3.0×10^{-6}	-0.4	11.4 _{31.9}	2.3×10^{-3}	-0.3	22.7 _{40.6}	19.0 _{39.3}	26.5 _{44.4}
Psychoeducation	3.4 _{18.2}	1.0×10^0	0.0	1.8 _{13.5}	1.0×10^0	-0.1	5.1 _{22.1}	1.0×10^0	0.1	3.0 _{17.0}	1.0×10^0	0.0	3.1 _{15.3}	1.8 _{13.2}	4.4 _{20.7}

Table S16. (Continue Table S15) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset⁷¹. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13 \times 6 = 78$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON NEEDS by 14.0% and human therapists responds by 24.5% on average when client expresses SHARING NEGATIVE EMOTIONS. ($P = 3.2 \times 10^{-7}$, Cohen’s $d = -0.3$, two-sided Student’s t -test).

Therapist Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			Average	High-quality	Low-quality
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	freq.	freq.
Client behavior: Sharing negative emotions															
n	1175			1052			1354			1348			820	646	174
Reflections on Needs	14.0 _{34.8}	3.2×10^{-7}	-0.3	7.2 _{25.9}	2.3×10^{-23}	-0.5	9.5 _{29.4}	5.6×10^{-19}	-0.4	9.2 _{28.9}	5.1×10^{-20}	-0.4	24.5 _{44.5}	28.9 _{45.4}	20.1 _{40.2}
Reflections on Emotions	9.0 _{28.7}	6.9×10^{-9}	-0.3	7.2 _{25.9}	2.6×10^{-12}	-0.4	8.1 _{27.3}	5.7×10^{-12}	-0.3	12.2 _{32.8}	8.3×10^{-4}	-0.2	19.4 _{42.7}	27.4 _{44.6}	11.5 _{32.0}
Reflections on Values	2.5 _{15.5}	2.5×10^{-3}	-0.2	0.2 _{4.4}	2.5×10^{-12}	-0.4	0.8 _{9.0}	5.2×10^{-11}	-0.3	1.3 _{11.2}	2.2×10^{-8}	-0.3	6.3 _{25.4}	7.4 _{26.2}	5.2 _{22.2}
Reflections on Consequences	7.1 _{25.8}	6.6×10^{-21}	-0.5	3.6 _{18.7}	1.7×10^{-35}	-0.6	3.5 _{18.3}	5.4×10^{-44}	-0.6	4.4 _{20.5}	1.8×10^{-37}	-0.6	21.8 _{39.8}	18.3 _{38.7}	25.3 _{43.6}
Reflections on Conflicts	1.4 _{11.6}	1.3×10^{-11}	-0.3	0.6 _{7.5}	1.7×10^{-14}	-0.4	1.1 _{10.5}	1.4×10^{-14}	-0.4	2.3 _{15.0}	1.7×10^{-8}	-0.3	8.4 _{29.5}	10.5 _{30.7}	6.3 _{24.4}
Reflections on Strengths	13.5 _{34.2}	8.4×10^{-5}	0.2	7.0 _{25.6}	1.0×10^0	0.0	12.6 _{33.1}	9.1×10^{-4}	0.2	10.4 _{30.5}	2.3×10^{-1}	0.1	6.6 _{26.6}	8.5 _{27.9}	4.6 _{21.0}
Questions on Experiences	11.6 _{32.0}	4.0×10^{-34}	-0.6	3.5 _{18.4}	1.6×10^{-72}	-0.9	13.4 _{34.0}	3.6×10^{-30}	-0.5	15.7 _{36.4}	1.8×10^{-22}	-0.5	34.6 _{48.6}	40.6 _{49.1}	28.7 _{45.4}
Questions on Perspectives	1.1 _{10.5}	1.4×10^{-1}	-0.1	0.2 _{4.4}	1.1×10^{-4}	-0.2	2.7 _{16.1}	1.0×10^0	0.0	2.2 _{14.5}	1.0×10^0	-0.1	3.3 _{19.9}	4.8 _{21.4}	1.7 _{13.1}
Questions on Emotions	0.2 _{4.1}	2.7×10^{-22}	-0.5	0.4 _{6.2}	1.3×10^{-18}	-0.4	0.5 _{7.2}	1.7×10^{-22}	-0.5	0.4 _{6.1}	1.2×10^{-23}	-0.5	9.6 _{31.1}	11.8 _{32.2}	7.5 _{26.4}
Problem-Solving	11.1 _{31.5}	2.1×10^{-5}	-0.2	14.2 _{34.9}	2.7×10^{-1}	-0.1	22.6 _{41.8}	1.0×10^0	0.1	15.4 _{36.1}	1.0×10^0	-0.1	19.0 _{35.7}	12.1 _{32.6}	25.9 _{43.9}
Planning	5.3 _{22.4}	1.0×10^0	-0.1	6.0 _{23.7}	1.0×10^0	-0.1	17.6 _{38.1}	2.5×10^{-9}	0.3	10.4 _{30.5}	1.0×10^0	0.1	7.6 _{25.0}	6.0 _{23.8}	9.2 _{29.0}
Normalizing	9.8 _{29.7}	4.1×10^{-4}	-0.2	10.6 _{30.7}	8.7×10^{-3}	-0.2	6.7 _{25.0}	9.3×10^{-12}	-0.3	6.7 _{25.0}	7.2×10^{-12}	-0.3	16.5 _{36.2}	14.7 _{35.4}	18.4 _{38.9}
Psychoeducation	2.6 _{15.8}	2.8×10^{-1}	-0.1	2.9 _{16.9}	1.0×10^0	-0.1	7.6 _{26.5}	4.7×10^{-1}	0.1	4.3 _{20.3}	1.0×10^0	0.0	4.7 _{17.5}	2.0 _{14.1}	7.5 _{26.4}
Client behavior: Sharing experiences															
n	874			812			908			926			759	533	226
Reflections on Needs	6.5 _{24.7}	1.5×10^{-18}	-0.5	1.4 _{11.6}	7.5×10^{-38}	-0.7	3.9 _{19.3}	5.6×10^{-29}	-0.6	5.0 _{21.7}	8.2×10^{-25}	-0.5	22.5 _{42.8}	26.5 _{44.2}	18.6 _{39.0}
Reflections on Emotions	5.8 _{23.5}	3.3×10^{-5}	-0.3	3.0 _{16.9}	2.2×10^{-11}	-0.4	4.4 _{20.5}	2.1×10^{-8}	-0.3	6.0 _{23.8}	4.8×10^{-5}	-0.2	13.5 _{36.8}	19.9 _{40.0}	7.1 _{25.7}
Reflections on Values	0.6 _{7.5}	6.6×10^{-5}	-0.2	0.0 _{0.0}	8.2×10^{-7}	-0.3	0.8 _{8.8}	2.3×10^{-4}	-0.2	1.3 _{11.3}	7.6×10^{-3}	-0.2	4.7 _{23.1}	7.1 _{25.8}	2.2 _{14.7}
Reflections on Consequences	2.4 _{15.3}	4.9×10^{-26}	-0.6	1.0 _{9.9}	1.8×10^{-31}	-0.6	1.9 _{13.6}	1.2×10^{-29}	-0.6	1.7 _{13.0}	5.5×10^{-31}	-0.6	18.2 _{38.5}	17.8 _{38.3}	18.6 _{39.0}
Reflections on Conflicts	1.0 _{10.1}	2.0×10^{-5}	-0.3	0.1 _{3.5}	1.7×10^{-8}	-0.3	0.9 _{9.4}	3.8×10^{-6}	-0.3	1.1 _{10.3}	1.6×10^{-5}	-0.3	5.8 _{25.1}	8.1 _{27.3}	3.5 _{18.5}
Reflections on Strengths	4.5 _{20.7}	1.3×10^{-1}	-0.2	1.8 _{13.5}	1.1×10^{-6}	-0.3	4.1 _{19.8}	2.9×10^{-2}	-0.2	3.7 _{18.8}	5.5×10^{-3}	-0.2	8.4 _{29.9}	12.0 _{32.5}	4.9 _{21.6}
Questions on Experiences	5.8 _{23.5}	3.2×10^{-82}	-1.0	2.3 _{15.1}	2.3×10^{-100}	-1.2	8.6 _{28.0}	1.3×10^{-68}	-0.9	9.1 _{28.7}	5.0×10^{-67}	-0.9	44.7 _{49.8}	45.2 _{49.8}	44.2 _{49.8}
Questions on Perspectives	0.1 _{3.4}	3.2×10^{-2}	-0.2	0.0 _{0.0}	1.6×10^{-2}	-0.2	1.1 _{10.4}	1.0×10^0	-0.1	1.0 _{9.8}	1.0×10^0	-0.1	2.0 _{15.6}	3.2 _{17.6}	0.9 _{9.4}
Questions on Emotions	0.1 _{3.4}	1.0×10^{-16}	-0.4	0.4 _{6.1}	3.5×10^{-14}	-0.4	0.2 _{4.7}	1.1×10^{-16}	-0.4	0.9 _{9.3}	2.5×10^{-13}	-0.4	8.8 _{28.6}	9.2 _{28.9}	8.4 _{27.8}
Problem-Solving	4.7 _{21.2}	5.5×10^{-10}	-0.3	3.2 _{17.6}	3.4×10^{-14}	-0.4	5.9 _{23.7}	5.8×10^{-7}	-0.3	6.2 _{24.0}	1.4×10^{-6}	-0.3	14.0 _{32.7}	9.4 _{29.2}	18.6 _{39.0}
Planning	2.3 _{15.0}	3.0×10^{-4}	-0.2	1.7 _{13.0}	1.6×10^{-5}	-0.3	5.3 _{22.4}	1.0×10^0	-0.1	4.1 _{19.8}	6.7×10^{-1}	-0.1	7.0 _{25.5}	6.9 _{25.4}	7.1 _{25.7}
Normalizing	4.0 _{19.6}	3.5×10^{-11}	-0.4	3.2 _{17.6}	5.2×10^{-13}	-0.4	2.5 _{15.7}	7.7×10^{-17}	-0.4	2.6 _{15.9}	7.4×10^{-17}	-0.4	14.0 _{34.7}	13.9 _{34.6}	14.2 _{34.9}
Psychoeducation	0.7 _{8.3}	3.5×10^{-9}	-0.3	0.2 _{5.0}	2.7×10^{-11}	-0.4	1.5 _{12.3}	2.7×10^{-5}	-0.3	0.5 _{7.3}	1.3×10^{-10}	-0.3	5.7 _{20.7}	2.6 _{16.0}	8.8 _{28.5}
Client behavior: Gained insights															
n	561			382			490			425			258	214	44
Reflections on Needs	10.0 _{30.0}	2.4×10^{-8}	-0.5	7.3 _{26.1}	6.5×10^{-10}	-0.6	6.1 _{24.0}	5.3×10^{-14}	-0.6	7.8 _{26.8}	5.3×10^{-10}	-0.6	27.0 _{45.1}	29.0 _{45.5}	25.0 _{43.8}
Reflections on Emotions	3.4 _{18.1}	1.1×10^{-4}	-0.4	2.9 _{16.7}	4.2×10^{-4}	-0.4	3.1 _{17.2}	9.8×10^{-5}	-0.4	3.5 _{18.5}	1.4×10^{-3}	-0.3	12.5 _{35.5}	15.9 _{36.6}	9.1 _{29.1}
Reflections on Values	1.8 _{13.2}	1.4×10^{-3}	-0.3	0.3 _{5.1}	1.8×10^{-5}	-0.4	2.2 _{14.8}	1.9×10^{-2}	-0.3	1.4 _{11.8}	1.9×10^{-3}	-0.3	8.1 _{28.6}	9.3 _{29.2}	6.8 _{25.5}
Reflections on Consequences	4.6 _{21.0}	5.1×10^{-10}	-0.5	2.4 _{15.2}	2.1×10^{-12}	-0.6	1.0 _{10.1}	3.0×10^{-20}	-0.8	1.4 _{11.8}	1.9×10^{-16}	-0.7	18.8 _{37.0}	15.0 _{35.7}	22.7 _{42.4}
Reflections on Conflicts	0.7 _{8.4}	1.9×10^{-1}	-0.2	0.3 _{5.1}	1.3×10^{-1}	-0.3	0.8 _{9.0}	4.8×10^{-1}	-0.2	1.4 _{11.8}	1.0×10^0	-0.2	3.9 _{21.9}	5.6 _{23.1}	2.3 _{15.1}
Reflections on Strengths	47.4 _{50.0}	7.8×10^{-21}	-0.8	23.3 _{42.3}	6.1×10^{-2}	0.3	21.4 _{41.1}	2.5×10^{-1}	0.2	17.9 _{38.4}	1.0×10^0	0.1	12.5 _{35.5}	15.9 _{36.6}	9.1 _{29.1}
Questions on Experiences	4.5 _{20.7}	2.8×10^{-21}	-0.8	0.8 _{8.8}	1.2×10^{-24}	-0.9	6.7 _{25.1}	2.7×10^{-14}	-0.6	6.6 _{24.8}	3.5×10^{-13}	-0.6	29.2 _{48.6}	42.5 _{49.6}	15.9 _{37.0}
Questions on Perspectives	2.1 _{14.5}	1.0×10^0	-0.1	0.3 _{5.1}	9.5×10^{-2}	-0.3	2.0 _{14.2}	1.0×10^0	-0.1	2.6 _{15.9}	1.0×10^0	-0.1	4.2 _{22.7}	6.1 _{23.9}	2.3 _{15.1}
Questions on Emotions	0.0 _{0.0}	2.0×10^{-14}	-0.6	0.3 _{5.1}	6.9×10^{-9}	-0.5	0.4 _{6.4}	9.4×10^{-11}	-0.6	0.0 _{0.0}	7.3×10^{-11}	-0.6	10.8 _{30.7}	10.3 _{30.4}	11.4 _{32.1}
Problem-Solving	18.0 _{38.5}	5.5×10^{-2}	-0.3	17.3 _{37.9}	5.1×10^{-2}	-0.3	24.9 _{43.3}	1.0×10^0	-0.1	11.1 _{31.4}	9.9×10^{-8}	-0.5	27.9 _{39.6}	15.0 _{35.7}	40.9 _{49.7}
Planning	25.8 _{43.8}	2.4×10^{-4}	0.4	16.2 _{36.9}	1.0×10^0	0.1	33.7 _{47.3}	2.4×10^{-9}	0.5	16.7 _{37.3}	1.0×10^0	0.1	11.5 _{32.1}	11.7 _{32.2}	11.4 _{32.1}
Normalizing	32.4 _{46.9}	2.0×10^{-3}	0.3	15.7 _{36.4}	1.0×10^0	-0.1	10.8 _{31.1}	1.6×10^{-1}	-0.2	10.6 _{30.8}	1.5×10^{-1}	-0.2	18.6 _{35.1}	12.1 _{32.7}	25.0 _{43.8}
Psychoeducation	5.3 _{22.5}	1.0×10^0	-0.1	5.2 _{22.3}	1.0×10^0	-0.2	13.1 _{33.7}	1.0×10^0	0.2	6.6 _{24.8}	1.0×10^0	-0.1	8.4 _{18.4}	0.9 _{9.6}	15.9 _{37.0}

Table S17. Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in HOPE dataset⁷¹. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13 \times 6 = 78$). d is the cohen’s d when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON CONSEQUENCES by 11.6% which is 15.7% less than average human therapists. (Single: 11.6%, Average human therapists: 27.3% from Table S15; $P = 3.2 \times 10^{-21}$, Cohen’s $d = -0.4$, two-sided Student’s t -test)

Therapist Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)														
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			GPT-4			GPT-3.5			Llama2-70b			Llama2-13b					
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d			
Client behavior: Changing unhealthy behavior																											
n	1553			999			1553			640			1553			796			1553			556					
Reflections on Needs	26.0439	1.0×10^0	0.0	28.0449	1.0×10^0	0.0	17.8383	1.4×10^{-5}	-0.2	20.1401	4.2×10^{-2}	-0.2	0.000	6.8×10^{-115}	-0.9	0.000	6.8×10^{-115}	-0.9	0.000	6.8×10^{-115}	-0.9	0.000	6.8×10^{-115}	-0.9	0.000	6.8×10^{-115}	-0.9
Reflections on Emotions	6.9254	3.6×10^{-1}	-0.1	9.7296	1.0×10^0	0.0	6.3243	1.5×10^{-1}	-0.1	13.8346	1.0×10^0	0.1	0.000	3.7×10^{-32}	-0.4	0.000	3.7×10^{-32}	-0.4	0.000	3.7×10^{-32}	-0.4	0.000	3.7×10^{-32}	-0.4	0.000	3.7×10^{-32}	-0.4
Reflections on Values	5.2222	1.0×10^0	0.0	0.356	2.1×10^{-4}	-0.2	2.9168	1.0×10^0	-0.1	3.2177	1.0×10^0	-0.1	0.000	2.4×10^{-13}	-0.3	0.000	2.4×10^{-13}	-0.3	0.000	2.4×10^{-13}	-0.3	0.000	2.4×10^{-13}	-0.3	0.000	2.4×10^{-13}	-0.3
Reflections on Consequences	11.6321	3.2×10^{-21}	-0.4	7.3261	6.5×10^{-26}	-0.5	5.3224	2.2×10^{-39}	-0.6	10.4306	1.0×10^{-15}	-0.4	0.000	8.9×10^{-128}	-0.9	0.000	8.9×10^{-128}	-0.9	0.000	8.9×10^{-128}	-0.9	0.000	8.9×10^{-128}	-0.9	0.000	8.9×10^{-128}	-0.9
Reflections on Conflicts	1.4118	4.8×10^{-8}	-0.3	1.1104	5.8×10^{-6}	-0.3	1.8132	1.8×10^{-5}	-0.2	3.4182	3.0×10^{-1}	-0.1	0.000	1.5×10^{-22}	-0.4	0.000	1.5×10^{-22}	-0.4	0.000	1.5×10^{-22}	-0.4	0.000	1.5×10^{-22}	-0.4	0.000	1.5×10^{-22}	-0.4
Reflections on Strengths	80.7395	0.0×10^0	2.0	68.9463	1.0×10^{-204}	1.6	43.7496	8.1×10^{-82}	0.9	45.7499	6.3×10^{-76}	1.0	0.000	3.1×10^{-27}	-0.4	0.000	3.1×10^{-27}	-0.4	0.000	3.1×10^{-27}	-0.4	0.000	3.1×10^{-27}	-0.4	0.000	3.1×10^{-27}	-0.4
Questions on Experiences	16.7373	3.2×10^{-23}	-0.4	4.2201	2.0×10^{-51}	-0.7	18.5388	7.9×10^{-16}	-0.4	23.6225	1.9×10^{-5}	-0.3	0.000	8.4×10^{-161}	-1.0	0.000	8.4×10^{-161}	-1.0	0.000	8.4×10^{-161}	-1.0	0.000	8.4×10^{-161}	-1.0	0.000	8.4×10^{-161}	-1.0
Questions on Perspectives	2.8165	1.0×10^0	0.0	0.240	2.0×10^{-3}	-0.2	4.9216	1.0×10^0	0.1	3.8191	1.0×10^0	0.0	0.000	4.0×10^{-10}	-0.2	0.000	4.0×10^{-10}	-0.2	0.000	4.0×10^{-10}	-0.2	0.000	4.0×10^{-10}	-0.2	0.000	4.0×10^{-10}	-0.2
Questions on Emotions	0.571	5.6×10^{-18}	-0.4	0.568	1.1×10^{-11}	-0.3	1.3111	1.4×10^{-11}	-0.3	0.785	1.8×10^{-9}	-0.3	0.000	3.8×10^{-32}	-0.4	0.000	3.8×10^{-32}	-0.4	0.000	3.8×10^{-32}	-0.4	0.000	3.8×10^{-32}	-0.4	0.000	3.8×10^{-32}	-0.4
Problem-Solving	35.7479	4.0×10^{-3}	0.2	49.2500	1.2×10^{-20}	0.5	59.9490	1.5×10^{-53}	0.7	39.0488	9.7×10^{-5}	0.2	0.000	1.8×10^{-134}	-0.9	0.000	1.8×10^{-134}	-0.9	0.000	1.8×10^{-134}	-0.9	0.000	1.8×10^{-134}	-0.9	0.000	1.8×10^{-134}	-0.9
Planning	49.7500	3.9×10^{-97}	0.9	52.5800	2.5×10^{-91}	1.0	67.5469	1.9×10^{-189}	1.4	45.3498	2.0×10^{-58}	0.8	0.000	5.7×10^{-57}	-0.6	0.000	5.7×10^{-57}	-0.6	0.000	5.7×10^{-57}	-0.6	0.000	5.7×10^{-57}	-0.6	0.000	5.7×10^{-57}	-0.6
Normalizing	49.3500	2.3×10^{-93}	0.9	54.1499	1.4×10^{-96}	1.0	19.2394	5.0×10^{-22}	0.1	19.2395	1.7×10^{-14}	0.2	0.000	1.9×10^{-59}	-0.6	0.000	1.9×10^{-59}	-0.6	0.000	1.9×10^{-59}	-0.6	0.000	1.9×10^{-59}	-0.6	0.000	1.9×10^{-59}	-0.6
Psychoeducation	8.6281	7.2×10^{-2}	0.1	10.2302	1.7×10^{-3}	0.2	24.4430	1.1×10^{-43}	0.6	15.6364	8.5×10^{-14}	0.4	0.000	3.8×10^{-23}	-0.4	0.000	3.8×10^{-23}	-0.4	0.000	3.8×10^{-23}	-0.4	0.000	3.8×10^{-23}	-0.4	0.000	3.8×10^{-23}	-0.4
Client behavior: Sustaining unhealthy behavior																											
n	389			39			389			13			389			96			389			73					
Reflections on Needs	59.0498	4.1×10^{-3}	0.7	69.2480	9.6×10^{-2}	0.9	35.4481	1.0×10^0	0.2	38.4490	1.0×10^0	0.2	0.000	1.3×10^{-28}	-0.9	0.000	1.3×10^{-28}	-0.9	0.000	1.3×10^{-28}	-0.9	0.000	1.3×10^{-28}	-0.9	0.000	1.3×10^{-28}	-0.9
Reflections on Emotions	5.1223	1.0×10^0	-0.2	7.7271	1.0×10^0	-0.1	12.5332	1.0×10^0	0.0	23.3426	7.5×10^{-1}	0.3	0.000	5.4×10^{-10}	-0.5	0.000	5.4×10^{-10}	-0.5	0.000	5.4×10^{-10}	-0.5	0.000	5.4×10^{-10}	-0.5	0.000	5.4×10^{-10}	-0.5
Reflections on Values	12.3339	1.0×10^0	0.2	0.000	1.0×10^0	-0.3	3.1175	1.0×10^0	-0.2	5.5229	1.0×10^0	-0.1	0.000	9.1×10^{-6}	-0.4	0.000	9.1×10^{-6}	-0.4	0.000	9.1×10^{-6}	-0.4	0.000	9.1×10^{-6}	-0.4	0.000	9.1×10^{-6}	-0.4
Reflections on Consequences	51.3506	1.0×10^0	0.3	30.8480	1.0×10^0	-0.1	22.9423	9.5×10^{-1}	-0.3	34.2478	1.0×10^0	0.0	0.000	4.5×10^{-44}	-1.1	0.000	4.5×10^{-44}	-1.1	0.000	4.5×10^{-44}	-1.1	0.000	4.5×10^{-44}	-1.1	0.000	4.5×10^{-44}	-1.1
Reflections on Conflicts	10.3307	1.0×10^0	0.1	38.5506	1.2×10^{-2}	1.1	9.4293	1.0×10^0	0.0	5.5229	1.0×10^0	-0.1	0.000	2.9×10^{-7}	-0.4	0.000	2.9×10^{-7}	-0.4	0.000	2.9×10^{-7}	-0.4	0.000	2.9×10^{-7}	-0.4	0.000	2.9×10^{-7}	-0.4
Reflections on Strengths	38.5493	7.7×10^{-11}	1.2	38.5506	2.3×10^{-4}	1.3	21.9416	2.2×10^{-5}	0.6	23.3426	2.8×10^{-5}	0.7	0.000	1.2×10^{-3}	-0.3	0.000	1.2×10^{-3}	-0.3	0.000	1.2×10^{-3}	-0.3	0.000	1.2×10^{-3}	-0.3	0.000	1.2×10^{-3}	-0.3
Questions on Experiences	33.3478	1.0×10^0	-0.1	15.4376	1.0×10^0	-0.5	30.2462	1.0×10^0	-0.2	34.2478	1.0×10^0	-0.1	0.000	1.8×10^{-46}	-1.1	0.000	1.8×10^{-46}	-1.1	0.000	1.8×10^{-46}	-1.1	0.000	1.8×10^{-46}	-1.1	0.000	1.8×10^{-46}	-1.1
Questions on Perspectives	7.7270	1.0×10^0	0.2	0.000	1.0×10^0	-0.2	7.3261	1.0×10^0	0.2	11.0315	5.6×10^{-1}	0.3	0.000	2.9×10^{-2}	-0.3	0.000	2.9×10^{-2}	-0.3	0.000	2.9×10^{-2}	-0.3	0.000	2.9×10^{-2}	-0.3	0.000	2.9×10^{-2}	-0.3
Questions on Emotions	0.000	1.0×10^0	-0.3	0.000	1.0×10^0	-0.3	0.000	2.0×10^{-1}	-0.3	1.4117	1.0×10^0	-0.3	0.000	1.2×10^{-7}	-0.4	0.000	1.2×10^{-7}	-0.4	0.000	1.2×10^{-7}	-0.4	0.000	1.2×10^{-7}	-0.4	0.000	1.2×10^{-7}	-0.4
Problem-Solving	30.8468	1.0×10^0	0.2	46.2519	1.0×10^0	0.6	61.5489	1.6×10^{-12}	0.9	43.8500	1.4×10^{-2}	0.5	0.000	9.8×10^{-27}	-0.8	0.000	9.8×10^{-27}	-0.8	0.000	9.8×10^{-27}	-0.8	0.000	9.8×10^{-27}	-0.8	0.000	9.8×10^{-27}	-0.8
Planning	20.5409	3.3×10^{-2}	0.6	15.4376	1.0×10^0	0.4	44.8500	3.5×10^{-25}	1.3	26.0442	6.3×10^{-7}	0.7	0.000	1.1×10^{-5}	-0.4	0.000	1.1×10^{-5}	-0.4	0.000	1.1×10^{-5}	-0.4	0.000	1.1×10^{-5}	-0.4	0.000	1.1×10^{-5}	-0.4
Normalizing	23.1427	1.0×10^0	0.3	23.1439	1.0×10^0	0.3	11.5320	1.0×10^0	-0.1	11.0315	1.0×10^0	-0.1	0.000	2.7×10^{-14}	-0.6	0.000	2.7×10^{-14}	-0.6	0.000	2.7×10^{-14}	-0.6	0.000	2.7×10^{-14}	-0.6	0.000	2.7×10^{-14}	-0.6
Psychoeducation	2.6160	1.0×10^0	-0.3	15.4376	1.0×10^0	0.2	16.7375	1.0×10^0	0.2	9.6296	1.0×10^0	0.0	0.000	7.9×10^{-11}	-0.5	0.000	7.9×10^{-11}	-0.5	0.000	7.9×10^{-11}	-0.5	0.000	7.9×10^{-11}	-0.5	0.000	7.9×10^{-11}	-0.5
Client behavior: Sharing positive emotions																											
n	294			521			294			302			294			95			294			143					
Reflections on Needs	6.3244	1.4×10^{-4}	-0.4	21.2409	1.0×10^0	0.1	11.6322	1.0×10^0	-0.1	9.1288	1.0×10^0	-0.2	0.000	9.3×10^{-11}	-0.6	0.000	9.3×10^{-11}	-0.6	0.000	9.3×10^{-11}	-0.6	0.000	9.3×10^{-11}	-0.6	0.000	9.3×10^{-11}	-0.6
Reflections on Emotions	3.6188	1.0×10^0	-0.2	6.6249	1.0×10^0	0.0	1.1103	1.0×10^0	-0.2	2.8165	1.0×10^0	-0.2	0.000	5.0×10^{-3}	-0.3	0.000	5.0×10^{-3}	-0.3	0.000	5.0×10^{-3}	-0.3	0.000	5.0×10^{-3}	-0.3	0.000	5.0×10^{-3}	-0.3
Reflections on Values	1.3115	1.0×10^0	-0.1	0.358	1.0×10^0	-0.2	3.2176	1.0×10^0	0.0	0.784	1.0×10^0	-0.2	0.000	3.8×10^{-1}	-0.2	0.000	3.8×10^{-1}	-0.2	0.000								

Table S18. (Continue Table S17) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in HOPE dataset⁷¹. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13 \times 6 = 78$). d is the cohen’s d when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON NEEDS by 46.5% which is 22.0% more than average human therapists. (Single: 46.5%, Average human therapists: 24.5% from Table S16; $P = 1.1 \times 10^{-11}$, Cohen’s $d = 0.5$, two-sided Student’s t -test)

Therapist Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			GPT-4			GPT-3.5			Llama2-70b			Llama2-13b		
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d
Client behavior: Sharing negative emotions																								
n	820			355			820			232			820			534			820			528		
Reflections on Needs	46.5 _{99.2}	1.1 × 10 ⁻¹¹	0.5	32.8 _{47.0}	1.0 × 10 ⁰	0.2	24.2 _{42.8}	1.0 × 10 ⁰	0.0	23.5 _{42.4}	1.0 × 10 ⁰	0.0	0.0 _{0.0}	1.5 × 10 ⁻⁵⁰	-0.8	0.0 _{0.0}	1.5 × 10 ⁻⁵⁰	-0.8	0.0 _{0.0}	1.5 × 10 ⁻⁵⁰	-0.8	0.0 _{0.0}	1.5 × 10 ⁻⁵⁰	-0.8
Reflections on Emotions	29.9 _{45.8}	1.4 × 10 ⁻²	0.2	32.8 _{47.0}	3.6 × 10 ⁻³	0.3	20.6 _{40.5}	1.0 × 10 ⁰	0.0	31.2 _{46.4}	1.5 × 10 ⁻⁴	0.3	0.0 _{0.0}	4.3 × 10 ⁻³⁵	-0.6	0.0 _{0.0}	4.3 × 10 ⁻³⁵	-0.6	0.0 _{0.0}	4.3 × 10 ⁻³⁵	-0.6	0.0 _{0.0}	4.3 × 10 ⁻³⁵	-0.6
Reflections on Values	8.2 _{27.4}	1.0 × 10 ⁰	0.1	0.9 _{9.3}	1.1 × 10 ⁻¹	-0.2	2.1 _{14.2}	3.6 × 10 ⁻²	-0.2	3.2 _{17.7}	1.0 × 10 ⁰	-0.1	0.0 _{0.0}	1.5 × 10 ⁻¹⁰	-0.4	0.0 _{0.0}	1.5 × 10 ⁻¹⁰	-0.4	0.0 _{0.0}	1.5 × 10 ⁻¹⁰	-0.4	0.0 _{0.0}	1.5 × 10 ⁻¹⁰	-0.4
Reflections on Consequences	23.7 _{42.6}	1.0 × 10 ⁰	0.0	16.4 _{37.1}	1.0 × 10 ⁰	-0.1	8.8 _{28.4}	7.4 × 10 ⁻⁹	-0.4	11.2 _{31.5}	2.2 × 10 ⁻⁵	-0.3	0.0 _{0.0}	1.1 × 10 ⁻⁴⁹	-0.8	0.0 _{0.0}	1.1 × 10 ⁻⁴⁹	-0.8	0.0 _{0.0}	1.1 × 10 ⁻⁴⁹	-0.8	0.0 _{0.0}	1.1 × 10 ⁻⁴⁹	-0.8
Reflections on Conflicts	4.5 _{20.8}	1.0 × 10 ⁰	-0.1	2.6 _{15.9}	3.0 × 10 ⁻¹	-0.2	2.8 _{16.5}	5.1 × 10 ⁻³	-0.2	5.9 _{23.5}	1.0 × 10 ⁰	-0.1	0.0 _{0.0}	4.7 × 10 ⁻¹⁴	-0.4	0.0 _{0.0}	4.7 × 10 ⁻¹⁴	-0.4	0.0 _{0.0}	4.7 × 10 ⁻¹⁴	-0.4	0.0 _{0.0}	4.7 × 10 ⁻¹⁴	-0.4
Reflections on Strengths	44.8 _{99.8}	1.4 × 10 ⁻³⁶	1.1	31.9 _{46.7}	4.1 × 10 ⁻²³	0.8	31.8 _{46.6}	5.9 × 10 ⁻³³	0.7	26.5 _{44.2}	2.8 × 10 ⁻²²	0.6	0.0 _{0.0}	2.1 × 10 ⁻¹⁰	-0.3	0.0 _{0.0}	2.1 × 10 ⁻¹⁰	-0.3	0.0 _{0.0}	2.1 × 10 ⁻¹⁰	-0.3	0.0 _{0.0}	2.1 × 10 ⁻¹⁰	-0.3
Questions on Experiences	38.3 _{48.7}	1.0 × 10 ⁰	0.1	15.9 _{36.7}	5.2 × 10 ⁻⁶	-0.4	33.9 _{47.4}	1.0 × 10 ⁰	0.0	40.2 _{49.1}	1.0 × 10 ⁰	0.1	0.0 _{0.0}	7.1 × 10 ⁻⁸¹	-1.0	0.0 _{0.0}	7.1 × 10 ⁻⁸¹	-1.0	0.0 _{0.0}	7.1 × 10 ⁻⁸¹	-1.0	0.0 _{0.0}	7.1 × 10 ⁻⁸¹	-1.0
Questions on Perspectives	3.7 _{18.8}	1.0 × 10 ⁰	0.0	0.9 _{9.3}	1.0 × 10 ⁰	-0.1	6.7 _{25.1}	3.7 × 10 ⁻¹	0.2	5.5 _{22.8}	1.0 × 10 ⁰	0.1	0.0 _{0.0}	2.4 × 10 ⁻⁴	-0.2	0.0 _{0.0}	2.4 × 10 ⁻⁴	-0.2	0.0 _{0.0}	2.4 × 10 ⁻⁴	-0.2	0.0 _{0.0}	2.4 × 10 ⁻⁴	-0.2
Questions on Emotions	0.6 _{7.5}	5.9 × 10 ⁻⁶	-0.3	1.7 _{13.0}	1.3 × 10 ⁻²	-0.3	1.3 _{11.4}	3.3 × 10 ⁻⁷	-0.3	0.9 _{9.7}	5.5 × 10 ⁻⁸	-0.3	0.0 _{0.0}	1.7 × 10 ⁻¹⁶	-0.4	0.0 _{0.0}	1.7 × 10 ⁻¹⁶	-0.4	0.0 _{0.0}	1.7 × 10 ⁻¹⁶	-0.4	0.0 _{0.0}	1.7 × 10 ⁻¹⁶	-0.4
Problem-Solving	36.9 _{48.3}	2.1 × 10 ⁻¹⁰	0.4	64.2 _{48.0}	3.9 × 10 ⁻⁴⁸	1.2	57.3 _{49.5}	3.8 × 10 ⁻⁵⁴	0.9	39.2 _{48.9}	4.3 × 10 ⁻¹⁶	0.5	0.0 _{0.0}	4.9 × 10 ⁻⁴⁷	-0.8	0.0 _{0.0}	4.9 × 10 ⁻⁴⁷	-0.8	0.0 _{0.0}	4.9 × 10 ⁻⁴⁷	-0.8	0.0 _{0.0}	4.9 × 10 ⁻⁴⁷	-0.8
Planning	17.5 _{38.0}	1.4 × 10 ⁻⁵	0.3	27.2 _{44.6}	1.7 × 10 ⁻¹⁵	0.6	44.6 _{49.8}	1.6 × 10 ⁻⁶³	1.0	26.5 _{44.2}	6.8 × 10 ⁻²¹	0.6	0.0 _{0.0}	5.5 × 10 ⁻¹⁶	-0.4	0.0 _{0.0}	5.5 × 10 ⁻¹⁶	-0.4	0.0 _{0.0}	5.5 × 10 ⁻¹⁶	-0.4	0.0 _{0.0}	5.5 × 10 ⁻¹⁶	-0.4
Normalizing	32.4 _{46.9}	3.7 × 10 ⁻⁸	0.4	47.8 _{50.1}	3.7 × 10 ⁻²³	0.8	17.0 _{37.6}	1.0 × 10 ⁰	0.0	17.0 _{37.6}	1.0 × 10 ⁰	0.0	0.0 _{0.0}	2.0 × 10 ⁻³⁵	-0.6	0.0 _{0.0}	2.0 × 10 ⁻³⁵	-0.6	0.0 _{0.0}	2.0 × 10 ⁻³⁵	-0.6	0.0 _{0.0}	2.0 × 10 ⁻³⁵	-0.6
Psychoeducation	8.5 _{27.9}	4.6 × 10 ⁻¹	0.2	13.4 _{34.1}	1.8 × 10 ⁻⁵	0.4	19.3 _{39.5}	7.1 × 10 ⁻¹⁸	0.5	11.0 _{31.3}	2.4 × 10 ⁻⁴	0.3	0.0 _{0.0}	1.3 × 10 ⁻¹²	-0.4	0.0 _{0.0}	1.3 × 10 ⁻¹²	-0.4	0.0 _{0.0}	1.3 × 10 ⁻¹²	-0.4	0.0 _{0.0}	1.3 × 10 ⁻¹²	-0.4
Client behavior: Sharing experiences																								
n	759			115			759			53			759			149			759			167		
Reflections on Needs	49.6 _{50.2}	8.4 × 10 ⁻⁸	0.6	20.8 _{40.9}	1.0 × 10 ⁰	0.0	23.5 _{42.8}	1.0 × 10 ⁰	0.0	27.5 _{44.8}	1.0 × 10 ⁰	0.1	0.0 _{0.0}	8.7 × 10 ⁻⁴³	-0.7	0.0 _{0.0}	8.7 × 10 ⁻⁴³	-0.7	0.0 _{0.0}	8.7 × 10 ⁻⁴³	-0.7	0.0 _{0.0}	8.7 × 10 ⁻⁴³	-0.7
Reflections on Emotions	44.3 _{49.9}	4.0 × 10 ⁻¹³	0.8	45.3 _{50.3}	3.6 × 10 ⁻⁷	0.8	26.8 _{44.5}	7.7 × 10 ⁻³	0.4	33.5 _{47.4}	1.8 × 10 ⁻⁷	0.5	0.0 _{0.0}	2.1 × 10 ⁻²¹	-0.5	0.0 _{0.0}	2.1 × 10 ⁻²¹	-0.5	0.0 _{0.0}	2.1 × 10 ⁻²¹	-0.5	0.0 _{0.0}	2.1 × 10 ⁻²¹	-0.5
Reflections on Values	4.3 _{20.5}	1.0 × 10 ⁰	0.0	0.0 _{0.0}	1.0 × 10 ⁰	-0.2	4.7 _{21.2}	1.0 × 10 ⁰	0.0	7.2 _{25.9}	1.0 × 10 ⁰	0.1	0.0 _{0.0}	2.4 × 10 ⁻⁶	-0.3	0.0 _{0.0}	2.4 × 10 ⁻⁶	-0.3	0.0 _{0.0}	2.4 × 10 ⁻⁶	-0.3	0.0 _{0.0}	2.4 × 10 ⁻⁶	-0.3
Reflections on Consequences	18.3 _{38.8}	1.0 × 10 ⁰	0.0	11.4 _{31.1}	1.0 × 10 ⁰	0.0	11.4 _{31.1}	1.0 × 10 ⁰	0.0	9.6 _{29.5}	5.1 × 10 ⁻¹	-0.2	0.0 _{0.0}	5.6 × 10 ⁻³⁵	-0.7	0.0 _{0.0}	5.6 × 10 ⁻³⁵	-0.7	0.0 _{0.0}	5.6 × 10 ⁻³⁵	-0.7	0.0 _{0.0}	5.6 × 10 ⁻³⁵	-0.7
Reflections on Conflicts	7.8 _{27.0}	1.0 × 10 ⁰	0.1	1.9 _{13.7}	1.0 × 10 ⁰	-0.2	5.4 _{23.6}	1.0 × 10 ⁰	0.0	6.0 _{23.8}	1.0 × 10 ⁰	0.0	0.0 _{0.0}	1.8 × 10 ⁻⁸	-0.3	0.0 _{0.0}	1.8 × 10 ⁻⁸	-0.3	0.0 _{0.0}	1.8 × 10 ⁻⁸	-0.3	0.0 _{0.0}	1.8 × 10 ⁻⁸	-0.3
Reflections on Strengths	33.9 _{47.5}	1.6 × 10 ⁻¹²	0.8	28.3 _{45.5}	6.2 × 10 ⁻⁴	0.6	24.8 _{43.3}	1.8 × 10 ⁻⁶	0.5	20.4 _{40.4}	1.1 × 10 ⁻³	0.4	0.0 _{0.0}	1.0 × 10 ⁻¹²	-0.4	0.0 _{0.0}	1.0 × 10 ⁻¹²	-0.4	0.0 _{0.0}	1.0 × 10 ⁻¹²	-0.4	0.0 _{0.0}	1.0 × 10 ⁻¹²	-0.4
Questions on Experiences	44.3 _{49.9}	1.0 × 10 ⁰	0.0	35.8 _{48.4}	1.0 × 10 ⁰	-0.2	52.3 _{50.1}	1.0 × 10 ⁰	0.2	50.3 _{50.1}	1.0 × 10 ⁰	0.1	0.0 _{0.0}	4.9 × 10 ⁻¹¹²	-1.3	0.0 _{0.0}	4.9 × 10 ⁻¹¹²	-1.3	0.0 _{0.0}	4.9 × 10 ⁻¹¹²	-1.3	0.0 _{0.0}	4.9 × 10 ⁻¹¹²	-1.3
Questions on Perspectives	0.9 _{9.3}	1.0 × 10 ⁰	-0.1	0.0 _{0.0}	1.0 × 10 ⁰	-0.1	6.7 _{25.1}	2.3 × 10 ⁻¹	0.3	5.4 _{22.6}	1.0 × 10 ⁰	0.2	0.0 _{0.0}	2.7 × 10 ⁻²	-0.2	0.0 _{0.0}	2.7 × 10 ⁻²	-0.2	0.0 _{0.0}	2.7 × 10 ⁻²	-0.2	0.0 _{0.0}	2.7 × 10 ⁻²	-0.2
Questions on Emotions	0.9 _{9.3}	2.5 × 10 ⁻¹	-0.3	5.7 _{23.3}	1.0 × 10 ⁰	-0.1	1.3 _{11.5}	1.4 × 10 ⁻¹	-0.3	4.8 _{21.4}	1.0 × 10 ⁰	-0.1	0.0 _{0.0}	4.0 × 10 ⁻¹⁵	-0.4	0.0 _{0.0}	4.0 × 10 ⁻¹⁵	-0.4	0.0 _{0.0}	4.0 × 10 ⁻¹⁵	-0.4	0.0 _{0.0}	4.0 × 10 ⁻¹⁵	-0.4
Problem-Solving	35.7 _{48.1}	7.9 × 10 ⁻⁸	0.6	49.1 _{50.5}	8.0 × 10 ⁻¹¹	1.0	36.2 _{48.2}	4.9 × 10 ⁻¹⁰	0.6	34.1 _{47.6}	5.9 × 10 ⁻⁹	0.6	0.0 _{0.0}	6.8 × 10 ⁻²⁹	-0.6	0.0 _{0.0}	6.8 × 10 ⁻²⁹	-0.6	0.0 _{0.0}	6.8 × 10 ⁻²⁹	-0.6	0.0 _{0.0}	6.8 × 10 ⁻²⁹	-0.6
Planning	17.4 _{38.1}	1.3 × 10 ⁻²	0.4	26.4 _{44.5}	4.6 × 10 ⁻⁵	0.7	32.2 _{46.9}	4.8 × 10 ⁻¹⁸	0.8	22.8 _{42.1}	3.3 × 10 ⁻⁸	0.5	0.0 _{0.0}	4.9 × 10 ⁻¹²	-0.4	0.0 _{0.0}	4.9 × 10 ⁻¹²	-0.4	0.0 _{0.0}	4.9 × 10 ⁻¹²	-0.4	0.0 _{0.0}	4.9 × 10 ⁻¹²	-0.4
Normalizing	30.4 _{46.2}	5.9 × 10 ⁻⁴	0.5	49.1 _{50.5}	1.0 × 10 ⁻⁹	1.0	15.4 _{36.3}	1.0 × 10 ⁰	0.0	14.4 _{35.2}	1.0 × 10 ⁰	0.0	0.0 _{0.0}	7.5 × 10 ⁻²⁶	-0.6	0.0 _{0.0}	7.5 × 10 ⁻²⁶	-0.6	0.0 _{0.0}	7.5 × 10 ⁻²⁶	-0.6	0.0 _{0.0}	7.5 × 10 ⁻²⁶	-0.6
Psychoeducation	5.2 _{22.2}	1.0 × 10 ⁰	0.0	3.8 _{19.2}	1.0 × 10 ⁰	-0.1	9.4 _{29.3}	1.0 × 10 ⁰	0.2	3.0 _{17.1}	1.0 × 10 ⁰	-0.1	0.0 _{0.0}	3.0 × 10 ⁻¹²	-0.4	0.0 _{0.0}	3.0 × 10 ⁻¹²	-0.4	0.0 _{0.0}	3.0 × 10 ⁻¹²	-0.4	0.0 _{0.0}	3.0 × 10 ⁻¹²	-0.4
Client behavior: Gained insights																								
n	258			303			258			124			258			232			258			167		
Reflections on Needs	18.5 _{38.9}	1.0 × 10 ⁰	-0.2	22.6 _{42.0}	1.0 × 10 ⁰	-0.1	12.9 _{33.6}	9.5 × 10 ⁻³	-0.4	19.8 _{39.9}	1.0 × 10 ⁰	-0.2	0.0 _{0.0}	2.6 × 10 ⁻¹⁸	-0.8	0.0 _{0.0}	2.6 × 10 ⁻¹⁸	-0.8	0.0 _{0.0}	2.6 × 10 ⁻¹⁸	-0.8	0.0 _{0.0}	2.6 × 10 ⁻¹⁸	-0.8
Reflections on Emotions	6.3 _{24.3}	1.0 × 10 ⁰	-0.2	8.9 _{28.5}	1.0 × 10 ⁰	-0.1	6.5 _{24.6}	1.0 × 10 ⁰	-0.2	9.0 _{28.7}	1.0 × 10 ⁰	-0.1	0.0 _{0.0}	2.1 × 10 ⁻⁶	-0.5	0.0 _{0.0}	2.1 × 10 ⁻⁶	-0.5	0.0 _{0.0}	2.1 × 10 ⁻⁶	-0.5	0.0 _{0.0}	2.1 × 10 ⁻⁶	-0.5
Reflections on Values	3.3 _{17.9}	1.0 × 10 ⁰	-0.2	0.8 _{9.0}	4.6 × 10 ⁻¹	-0.3	4.7 _{21.3}	1.0 × 10 ⁰	-0.1	3.6 _{18.7}	1.0 × 10 ⁰	-0.2	0.0 _{0.0}	5.3 × 10 ⁻⁴	-0.4	0.0 _{0.0}	5.3 × 10 ⁻⁴	-0.4	0.0 _{0.0}	5.3				

Table S19. Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset⁷¹. Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 14$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 exhibits POSITIVE EMOTIONS by 1.5% while human therapists responds by 1.0% on average ($P = 2.8 \times 10^{-19}$, Cohen’s $d = 0.2$, two-sided Student’s t -test).

Linguistic Attributes	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ($n = 7705$)			GPT-3.5 ($n = 7135$)			Llama2-70b ($n = 7461$)			Llama2-13b ($n = 7464$)			Average ($n = 5446$)	High-quality ($n = 3907$)	Low-quality ($n = 1539$)
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	freq.	freq.
Positive Emotions	1.5 _{2,1}	2.8×10^{-19}	0.2	1.1 _{1,8}	3.3×10^{-2}	0.1	1.0 _{2,1}	1.0×10^0	0.0	1.0 _{2,5}	1.0×10^0	0.0	1.0 _{4,2}	1.2 _{4,3}	0.7 _{3,6}
Negative Emotions	1.1 _{2,0}	1.3×10^{-16}	0.2	1.2 _{2,0}	1.1×10^{-26}	0.2	1.2 _{2,5}	6.1×10^{-20}	0.2	1.1 _{2,5}	2.3×10^{-10}	0.1	0.7 _{3,6}	0.7 _{3,4}	0.6 _{4,3}
BigWords	24.6 _{7,2}	0.0×10^0	1.6	20.9 _{7,7}	0.0×10^0	1.2	23.7 _{11,3}	0.0×10^0	1.2	21.0 _{12,0}	0.0×10^0	0.9	10.7 _{9,9}	11.0 _{9,9}	10.4 _{10,1}
Self	1.5 _{1,9}	6.3×10^{-139}	-0.5	1.8 _{2,3}	1.9×10^{-84}	-0.4	1.4 _{2,4}	2.9×10^{-146}	-0.5	1.4 _{2,4}	7.1×10^{-144}	-0.5	3.1 _{5,1}	2.5 _{4,5}	3.8 _{6,1}
Other	8.9 _{3,4}	1.0×10^0	0.0	8.4 _{3,6}	3.2×10^{-7}	-0.1	7.5 _{4,5}	6.2×10^{-41}	-0.2	7.0 _{5,0}	2.4×10^{-70}	-0.3	8.9 _{7,1}	8.7 _{6,9}	9.2 _{7,4}
Health	2.0 _{2,6}	9.4×10^{-2}	0.0	2.1 _{3,0}	6.0×10^{-5}	0.1	1.5 _{3,1}	2.1×10^{-8}	-0.1	1.3 _{3,0}	5.1×10^{-16}	-0.2	1.8 _{4,3}	1.2 _{3,2}	2.5 _{6,1}
Wellness	0.7 _{1,4}	3.0×10^{-41}	0.2	0.6 _{1,5}	1.0×10^{-13}	0.1	0.3 _{1,0}	1.0×10^0	0.0	0.3 _{1,1}	1.0×10^0	0.0	0.3 _{2,1}	0.2 _{1,3}	0.4 _{3,5}
Behavioral Activation	6.0 _{4,0}	7.7×10^{-225}	0.6	5.7 _{4,4}	2.0×10^{-163}	0.5	4.8 _{4,5}	6.6×10^{-70}	0.3	4.5 _{5,3}	1.1×10^{-38}	0.2	3.2 _{5,9}	3.5 _{5,9}	2.9 _{5,9}
Meaning	3.6 _{2,6}	0.0×10^0	0.7	3.2 _{2,6}	2.2×10^{-232}	0.6	2.2 _{2,8}	5.5×10^{-48}	0.3	2.0 _{2,7}	1.1×10^{-27}	0.2	1.4 _{3,5}	1.4 _{3,4}	1.4 _{3,7}
Purpose	3.9 _{2,7}	0.0×10^0	0.8	3.6 _{2,9}	1.5×10^{-275}	0.7	2.4 _{2,9}	6.3×10^{-50}	0.3	2.2 _{3,2}	4.9×10^{-28}	0.2	1.5 _{3,6}	1.5 _{3,5}	1.5 _{3,8}
Motivation	3.4 _{2,5}	0.0×10^0	0.7	3.3 _{2,7}	6.9×10^{-272}	0.7	2.0 _{2,5}	1.2×10^{-38}	0.2	1.9 _{2,9}	3.1×10^{-23}	0.2	1.3 _{3,3}	1.3 _{3,3}	1.3 _{3,4}
Sadness	1.0 _{1,6}	1.2×10^{-126}	0.4	1.1 _{1,8}	1.5×10^{-134}	0.5	0.8 _{1,9}	2.9×10^{-69}	0.3	0.9 _{2,1}	8.0×10^{-65}	0.3	0.3 _{1,5}	0.3 _{1,5}	0.2 _{1,6}
Sympathy	0.3 _{0,8}	1.5×10^{-29}	0.2	0.4 _{1,2}	9.2×10^{-36}	0.2	0.3 _{1,0}	8.0×10^{-25}	0.2	0.4 _{1,4}	2.8×10^{-29}	0.2	0.1 _{1,0}	0.2 _{1,1}	0.1 _{0,8}
Stress	1.5 _{2,2}	3.6×10^{-16}	0.2	1.7 _{2,3}	1.4×10^{-30}	0.2	1.8 _{2,9}	6.4×10^{-33}	0.2	1.6 _{2,9}	4.6×10^{-18}	0.2	1.1 _{3,5}	1.1 _{3,2}	1.1 _{4,2}

Table S20. Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset⁷¹ for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where m is the number of t-test per one model ($m = 13$). d is the cohen’s d when compared with human therapist average. For instance, GPT-4 with single response simulation responds with POSITIVE EMOTIONS by 2.1% and GPT-4 with full conversation simulation responds by 1.3% when compared with the human therapists responds by 1.0% on average from Table S19. (Single: $P = 2.8 \times 10^{-24}$, Cohen’s $d = 0.3$; two-sided student t-test) (Full: $P = 4.7 \times 10^{-8}$, Cohen’s $d = 0.1$; two-sided student t-test).

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)														
	GPT-4 ($n = 6158$)			GPT-3.5 ($n = 6144$)			Llama2-70b ($n = 6174$)			Llama2-13b ($n = 6282$)			GPT-4 ($n = 1547$)			GPT-3.5 ($n = 991$)			Llama2-70b ($n = 1287$)			Llama2-13b ($n = 1182$)					
	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d	freq.	p	d
Positive Emotions	2.1 _{1,7}	2.8×10^{-24}	0.3	1.7 _{1,6}	1.3×10^{-7}	0.2	1.3 _{1,5}	3.3×10^{-2}	0.1	1.7 _{2,2}	6.2×10^{-8}	0.2	1.3 _{2,2}	4.7×10^{-8}	0.1	1.0 _{1,8}	1.0×10^0	0.0	0.9 _{2,2}	1.0×10^0	0.0	0.9 _{2,5}	1.0×10^0	0.0	0.9 _{2,5}	1.0×10^0	0.0
Negative Emotions	0.5 _{1,0}	1.0×10^0	-0.1	0.6 _{1,1}	1.0×10^0	0.0	0.7 _{1,1}	1.0×10^0	0.0	0.7 _{1,2}	1.0×10^0	0.0	1.3 _{2,1}	1.6×10^{-24}	0.2	1.3 _{2,1}	2.2×10^{-32}	0.2	1.3 _{2,7}	2.1×10^{-23}	0.2	1.1 _{2,6}	1.4×10^{-12}	0.1	1.1 _{2,6}	1.4×10^{-12}	0.1
BigWords	28.3 _{3,0}	0.0×10^0	1.9	25.6 _{6,7}	0.0×10^0	1.6	31.8 _{10,2}	0.0×10^0	2.1	31.2 _{16,0}	0.0×10^0	1.8	23.7 _{7,4}	0.0×10^0	1.5	20.2 _{7,6}	0.0×10^0	1.1	22.1 _{10,8}	0.0×10^0	1.1	19.1 _{10,1}	0.0×10^0	0.8	19.1 _{10,1}	0.0×10^0	0.8
Self	2.3 _{1,9}	4.0×10^{-8}	-0.2	2.7 _{2,4}	9.0×10^{-2}	-0.1	1.5 _{2,3}	6.2×10^{-28}	-0.3	1.5 _{1,9}	1.5×10^{-25}	-0.3	1.3 _{1,8}	8.7×10^{-147}	-0.5	1.6 _{2,3}	3.0×10^{-92}	-0.4	1.3 _{2,4}	1.5×10^{-129}	-0.5	1.3 _{2,5}	8.6×10^{-128}	-0.5	1.3 _{2,5}	8.6×10^{-128}	-0.5
Other	9.5 _{2,3}	3.5×10^{-2}	0.1	8.1 _{2,8}	3.1×10^{-3}	-0.1	5.9 _{3,2}	9.4×10^{-50}	-0.5	5.6 _{8,4}	1.4×10^{-49}	-0.5	8.8 _{3,6}	1.0×10^0	0.0	8.4 _{3,7}	2.5×10^{-5}	-0.1	7.9 _{4,6}	3.0×10^{-20}	-0.2	7.3 _{4,9}	8.0×10^{-49}	-0.3	7.3 _{4,9}	8.0×10^{-49}	-0.3
Health	2.3 _{2,1}	1.3×10^{-4}	0.1	2.7 _{2,5}	6.6×10^{-9}	0.2	1.7 _{1,9}	1.0×10^0	0.0	1.6 _{3,3}	1.0×10^0	0.0	1.9 _{2,7}	1.0×10^0	0.0	2.1 _{3,0}	4.0×10^{-2}	0.1	1.4 _{3,3}	7.8×10^{-9}	-0.1	1.2 _{3,0}	5.1×10^{-18}	-0.2	1.2 _{3,0}	5.1×10^{-18}	-0.2
Wellness	1.1 _{1,3}	3.3×10^{-40}	0.4	1.2 _{1,5}	1.2×10^{-31}	0.4	0.6 _{1,1}	1.2×10^{-6}	0.2	0.6 _{1,4}	5.7×10^{-5}	0.1	0.6 _{1,4}	1.7×10^{-22}	0.2	0.5 _{1,4}	6.3×10^{-5}	0.1	0.2 _{0,9}	1.7×10^{-1}	0.0	0.2 _{1,1}	6.0×10^{-2}	-0.1	0.2 _{1,1}	6.0×10^{-2}	-0.1
Behavioral Activation	7.8 _{3,0}	4.4×10^{-181}	0.9	8.7 _{4,2}	4.3×10^{-165}	1.0	6.8 _{3,8}	8.5×10^{-94}	0.7	5.7 _{4,0}	1.2×10^{-43}	0.5	5.6 _{4,1}	1.1×10^{-137}	0.5	5.3 _{4,3}	1.7×10^{-102}	0.4	4.4 _{4,5}	2.4×10^{-35}	0.2	4.3 _{5,5}	6.0×10^{-24}	0.2	4.3 _{5,5}	6.0×10^{-24}	0.2
Meaning	3.9 _{1,9}	2.6×10^{-159}	0.8	3.6 _{2,0}	8.7×10^{-82}	0.7	2.3 _{1,7}	4.9×10^{-19}	0.3	2.1 _{1,7}	1.8×10^{-9}	0.2	3.5 _{2,7}	5.7×10^{-286}	0.7	3.1 _{2,6}	5.2×10^{-193}	0.6	2.2 _{3,0}	4.0×10^{-39}	0.2	2.0 _{2,9}	2.8×10^{-23}	0.2	2.0 _{2,9}	2.8×10^{-23}	0.2
Purpose	4.6 _{2,0}	1.4×10^{-217}	0.9	4.8 _{2,5}	3.0×10^{-159}	1.0	2.7 _{1,9}	8.2×10^{-31}	0.4	2.5 _{2,8}	1.5×10^{-16}	0.3	3.7 _{2,8}	4.1×10^{-275}	0.7	3.4 _{2,9}	3.5×10^{-210}	0.6	2.3 _{3,1}	3.7×10^{-36}	0.2	2.1 _{3,2}	1.2×10^{-21}	0.2	2.1 _{3,2}	1.2×10^{-21}	0.2
Motivation	4.2 _{2,0}	1.8×10^{-221}	1.0	4.5 _{2,5}	5.7×10^{-171}	1.0	2.3 _{1,8}	2.5×10^{-25}	0.3	2.2 _{2,7}	2.4×10^{-15}	0.3	3.2 _{2,5}	1.4×10^{-246}	0.6	3.1 _{2,7}	1.4×10^{-204}	0.6	1.9 _{2,6}	9.2×10^{-27}	0.2	1.8 _{2,9}	2.3×10^{-17}	0.2	1.8 _{2,9}	2.3×10^{-17}	0.2
Sadness	0.6 _{0,9}	1.8×10^{-9}	0.2	0.6 _{1,0}	2.2×10^{-8}	0.2	0.6 _{1,0}	1.3×10^{-9}	0.2	1.0 _{2,0}	7.9×10^{-39}	0.4	1.1 _{1,7}	8.3×10^{-141}	0.5	1.1 _{1,9}	6.7×10^{-144}	0.5	0.9 _{2,0}	9.5×10^{-72}	0.3	0.8 _{2,1}	2.4×10^{-56}	0.3	0.8 _{2,1}	2.4×10^{-56}	0.3
Sympathy	0.2 _{0,4}	9.0×10^{-1}	0.1	0.2 _{0,5}	2.1×10^{-1}	0.1	0.3 _{0,7}	2.1×10^{-6}	0.2	0.6 _{1,9}	1.9×10^{-37}	0.4	0.3 _{0,8}	5.4×10^{-34}	0.2	0.4 _{1,2}	9.8×10^{-39}	0.2	0.3 _{1,1}	1.8×10^{-23}	0.2	0.3 _{1,3}	3.5×10^{-21}	0.2	0.3 _{1,3}	3.5×10^{-21}	0.2
Stress	0.7 _{1,2}	8.4×10^{-4}	-0.1	0.9 _{1,3}	6.4×10^{-1}	-0.1	1.3 _{1,5}	3.1×10^{-1}	0.1	1.2 _{1,5}	1.0×10^0	0.0	1.7 _{2,3}	2.4×10^{-28}	0.2	1.8 _{2,4}	6.3×10^{-40}	0.3	1.9 _{3,1}	3.4×10^{-36}	0.2	1.7 _{3,1}	8.0×10^{-21}	0.2	1.7 _{3,1}	8.0×10^{-21}	0.2

Table S21. Dataset statistics.

Dataset	# of Conversations	# of Utterances		Words per Utterance	
		Therapist	Client	Therapist	Client
High-Low Quality ⁴⁵	High quality: 155 Low quality: 104	3753	3790	31.8 (std = 34.7)	27.3 (std = 33.1)
HOPE ⁷¹	212	6070	6081	24.0 (std= 31.9)	21.7 (std = 32.3)

Table S22. Performance of classification models on the tasks of (1) identifying therapist behavior across 13 different therapists codes and (2) identifying client behavior across 6 different client codes. We create five random train-test splits of our annotated dataset and report the average macro-precision (P), macro-recall (R), and macro-F1 (F1) scores with standard deviations as subscripts. def.: Definitions; ex.: Examples. Best-performing models are **bolded**. We find that GPT-4-based prompting methods that make use of psychotherapy-based definitions and examples achieve the highest macro-F1 performance. We adopt the prompting (multi-label) method for classifying therapist behavior while the prompting (binary-label) method for classifying client behavior (both highlighted in lightblue).

Method	Model	Therapist Behavior			Client Behavior		
		P	R	F1	P	R	F1
Random	Uniform Distribution	11.8 _{0.7}	17.6 _{2.9}	14.1 _{1.3}	23.2 _{5.1}	29.6 _{8.0}	25.9 _{6.2}
Finetuning (multi-label)	GPT-3	51.6 _{3.2}	47.9 _{3.7}	49.6 _{2.8}	45.2 _{5.9}	36.8 _{6.0}	40.4 _{5.2}
	GPT-3.5-turbo	45.9 _{3.7}	31.1 _{3.1}	37.1 _{3.4}	44.5 _{8.0}	48.8 _{6.4}	46.3 _{6.3}
Prompting (binary-label)	GPT-4 (w/ def. & ex.)	49.4 _{2.6}	52.2 _{2.1}	50.7 _{1.8}	73.4_{5.9}	80.3 _{7.2}	76.7_{6.3}
Prompting (multi-label)	GPT-4 (w/ def.)	45.4 _{1.9}	57.8 _{3.8}	50.8 _{1.9}	65.6 _{7.3}	81.6_{7.5}	72.6 _{6.8}
	GPT-4 (w/ def. & ex.)	56.0_{1.7}	59.5_{3.7}	57.7_{2.5}	72.6 _{6.7}	76.4 _{10.2}	74.4 _{8.4}

Table S23. Therapist Conversational Behavior Definition and Corresponding Examples based on techniques identified by researchers^{43,44}. We used these behavior categories in our annotation dataset. See more details in Methods.

Behavior	Therapist	
	Definition	Example
REFLECTIONS ON		
NEEDS	Identifies an implied or background need for the client.	It sounds like you've realized that maintaining a balance between your work and personal life is essential for your overall well-being.
EMOTIONS	Identifies an implied or background emotion for the client	So it seems like you have been feeling a little overwhelmed and anxious about all the moving parts in your new job.
VALUES	Identifies an implied or background value or set of values for the client.	Being respected by others is a significant value for you.
CONSEQUENCES	Identifies consequences the client experience or could experience	Whenever you overspend on luxury items, you struggle to pay your bills at the end of the month.
CONFLICT	Identifies an implied or background emotional or situational conflict for the client.	You're striving to improve your health, but your demanding job leaves you with little time for exercise and nutrition.
STRENGTH	Identifies an implied or background strength or resource that the client exhibits.	Your ability to adapt and overcome adversity really shows your resilience and determination.
QUESTIONS ON		
EXPERIENCES	More information about a specific event or statement is sought	You mentioned you are trying to eat healthier. What changes did you make to your diet?
PERSPECTIVES	Client is asked to consider an experience from a different perspective or vantage point.	That's fantastic, now let's focus on the goals you want to accomplish. Can you visualize any particular approach or strategy you'd like to implement to achieve these goals?
EMOTIONS	Asks client to express how they are feeling in the immediate present about something that just happened in the therapy.	Would you like to talk more about what that feels like for you right now?
SOLUTIONS		
PROBLEM-SOLVING	Therapist offers possible solutions to a client problem.	It may help to create a routine for daily relaxation techniques, such as deep breathing or meditation. This could assist in managing your anxiety levels.
PLANNING	Therapist works with client to construct a specific plan of action.	Let's create a meal plan together. Try to follow it for the next two weeks and note down any changes you notice in your energy levels and overall well-being.
NORMALIZING		
NORMALIZING	The therapist acknowledges and validates the client's experience as "normal" or expectable, sympathizes with their challenges, and provides reassurance to foster a supportive and encouraging therapeutic atmosphere.	I hear you, it's perfectly normal to feel overwhelmed given your circumstances.
PSYCHOEDUCATION		
PSYCHOEDUCATION	Therapeutically relevant information about psychological principles is provided.	Cognitive behavioral therapy aids in altering detrimental thought patterns.

Table S24. Client Conversational Behavior Definition and Corresponding Examples based on expressions from clients identified by researchers⁴³. We used these behavior categories in our annotation dataset. See more details in Methods.

Behavior	Client	
	Definition	Example
CHANGING UNHEALTHY BEHAVIOR	Showing intention or action taken on changing unhealthy behavior.	I've tried to quit drinking, but I end up drinking more than I try to drink less.
SUSTAINING UNHEALTHY BEHAVIOR	Showing intention or action taken on sustaining unhealthy behavior	I am smoking around 20 cigarettes a day for the past couple of years. I know it's bad for my health, but I'm not ready to quit yet.
SHARING NEGATIVE FEELING OR EMOTION	Clients describe discomfort or suffering without a specific object, or explicitly acknowledge specific negative emotion.	Life has been really challenging lately, I am feeling lost.
SHARING POSITIVE FEELING OR EMOTION	Client describes enjoyment without a specific object, or explicitly acknowledges specific positive emotion.	I was so thankful when I received that news.
GAINED INSIGHT	Client expresses that they learned something new about themselves or about their situation.	I hadn't considered how much I avoid confrontations.
SHARING LIFE EVENT OR SITUATION	Client shares the details of their basic background, their life events, the situation they faced or the changes in their life.	I lost my job due to the pandemic, and I've now had to move back in with my parents at age 35 which is a significant change for me.

Table S25. Explanations generated from our behavior prediction model. We ask the model to explain its predictions and examine the differences between true and predicted behavior.

Utterance	True Behavior	Predicted Behavior	Explanation
I'm sorry to hear that, John. Have you tried applying ice to the affected area to reduce swelling and pain? Additionally, you might want to consider elevating your foot to help with the swelling. If the pain persists, it might be best to schedule an appointment with your primary care physician or a specialist to further assess the injury.	Normalizing, Problem-Solving	Problem-Solving, Planning	The utterance provides advice or solutions to a problem ("Have you tried applying ice to the affected area to reduce swelling and pain? Additionally, you might want to consider elevating your foot to help with the swelling.") which is indicative of the "Problem-Solving" intent. It also suggests a course of action ("If the pain persists, it might be best to schedule an appointment with your primary care physician or a specialist to further assess the injury.") which aligns with the "Planning" intent.
okay well hopefully you won't have any issues like that with this one I'll be sure to go over this with you okay	Normalizing	Problem-Solving	In the utterance, the speaker hints at having a solution to a problem that might have occurred previously. The speaker implies that they will attempt to prevent the same issues from happening with this new situation by going over it with the listener. This aligns with the "Problem-Solving" intent, where the speaker offers a possible solution or way to avoid a problem.
That's a great idea. Staying hydrated with water is important for your overall health, including your dental health. It's also a great way to keep yourself from reaching for a soft drink when you're thirsty. You could also try snacking on foods that are less harmful to your teeth, such as fruits or vegetables instead of reaching for candy or sugary snacks. How does that sound?	Problem-Solving, Evokes Concrete Elaboration	Problem-Solving, Teaching / Psychoeducation	The utterance is providing solutions to a potential problem (the client's possible tendency to reach for soft drinks or sugary snacks), hence it falls under the "Problem-Solving" category. It's also providing therapeutically relevant information about the importance of hydration and healthier snack alternatives for dental health, which is "Teaching/Psychoeducation".
Great, everything seems to be in order. Did you have any questions or concerns about your medication?	Evokes Concrete Elaboration	Evokes Concrete Elaboration	The phrase "Great, everything seems to be in order. Did you have any questions or concerns about your medication?" is seeking for more information about a specific topic, which is the client's medication. This aligns with the intent "Evokes Concrete Elaboration" as the speaker is asking for more details or clarification on a particular subject.
You're stuck between the demands of being a mother and of being a woman.	Makes Conflict Explicit	Makes Conflict Explicit	The statement "You're stuck between the demands of being a mother and of being a woman" shows that the speaker is acknowledging a conflict the listener is experiencing. The conflict is between the roles and responsibilities of being a mother and those of being a woman, possibly referring to personal, professional, or social aspirations or duties that the listener, as a woman, might have. This aligns with the "Makes Conflict Explicit" intent, which is defined as identifying an implied or background emotional or situational conflict for the client.

Table S26. First Part of the Illustrative Example of Single-Response Simulation Method by LLMs. This method involves the actual responses of both therapist and client from a reference dataset, which are written in *italic*, and simulated responses of LLM Therapists. This full example is the simulated conversation from GPT-4 on one of the conversations using High-Low quality dataset as reference.

Role	Utterances	Dataset
Therapist	<i>I can I help you</i>	Reference
	Hello, I'm glad you're here today to discuss any concerns or behaviors you'd like to change. Let's start by identifying what exactly you'd like to work on. Is it quitting smoking, alcohol consumption, or something else?	Simulated
Client	<i>yeah I hope so um my workplace is just gone smoke-free hey I'm not not very happy about it I'm going to keep smoking but I was worrying if there's anything that can help be like get through the day it's like you know go get my beer and smoke cigarettes after work</i>	Reference
Therapist	<i>so your workplace just went smoke free you're not too pleased huh</i>	Reference
	I understand that the change at your workplace has been challenging for you. It's important to recognize that there are different strategies that can be helpful in coping with this adjustment. Firstly, let's explore your reasons for continuing to smoke and how motivated you are to make changes in your smoking behavior. On a scale of 1 to 10, where 1 is not motivated at all and 10 is extremely motivated, how motivated are you to quit or cut back on smoking?	Simulated
Client	<i>no I'm not going to quit smoking oh my god I like it it helps me relax so I mean I don't see what the big deal is</i>	Reference
Therapist	<i>okay tell me a little bit about how smoking fits in your day it'll help with the</i>	Reference
	I understand that you enjoy smoking and find it relaxing. It's natural to feel frustrated when a change like this is imposed on you. Let's explore some strategies that might help you manage the situation at work while still maintaining your personal preferences. How do you feel about discussing this further?	Simulated
Client	<i>I mean I suppose a little less than a pack a day it's not too bad um the I probably smoked for the last five years but I I mean I could quit if I wanted to but I don't really want to right now</i>	Reference
Therapist	<i>okay so you feel like right now smoking is a pretty important part of your life it's not something you're ready to change but if the time came when you did decide to quit you feel pretty confident that you could go ahead and do that</i>	Reference
	I can understand that quitting smoking is a personal decision, and it's important that you feel confident in your choices. It sounds like smoking is a part of your daily routine that you find helpful in managing stress. Since your workplace has gone smoke-free, we can explore some strategies to cope while at work. Are you open to discussing some potential ways to manage without cigarettes during the day?	Simulated
Client	<i>sure I can quit</i>	Reference
Therapist	<i>okay what makes you so confident</i>	Reference
	Alright, it's great to hear that you believe in yourself and your ability to quit if you wanted to. Now, since your workplace has gone smoke-free and you're looking for ways to cope during the day, let's discuss some alternatives or strategies to help you manage any cravings or stress you might encounter. Have you ever tried using any nicotine replacement options like gum or patches? They can be helpful in reducing cravings while at work without actually smoking. Alternatively, we can also explore some stress management techniques, such as deep breathing exercises or even a short walk outside the workplace during breaks, to help you relax and refocus. What do you ...	Simulated
Client	<i>well I mean I don't see why I couldn't although I have not ever tried to quit I mean it's just something that I do for fun you know I'm breaks at work in the morning and have to work</i>	Reference
Therapist	<i>mm-hmm and so in terms of a readiness to actually go ahead and make a quit attempt where do you think you fall on the scale of one to ten where one is I'm not considering it at all and ten is I'm definitely ready I want to quit</i>	Reference
	It sounds like smoking is something you enjoy and is a part of your routine. It's important for me to understand where you're coming from so I can better help you. Since your workplace has gone smoke-free, have you noticed any changes in how you feel or act during the day?	Simulated
Client	<i>I am probably - I haven't even thought about quitting and until this great announcement at work today I've been totally not gonna quit smoking</i>	Reference
Therapist	<i>okay what made you say two instead of one</i>	Reference
	I understand, it sounds like the change at work has brought up the idea of quitting, but right now you're still at a 2 and not really considering it. In the meantime, your main concern is finding ways to manage stress without smoking since the workplace has gone smoke-free. Is that correct?	Simulated
Client	<i>well I mean anything's possible I mean I I guess I could maybe be convinced yeah is there something I mean I are there things I can try that would make me like want to quit smoking</i>	Reference
Therapist	<i>well there are a lot of products available now that can actually help you curb that urge to smoke we have products that can help to replace nicotine which can help to decrease withdrawal symptoms that you might experience when quitting or cutting down and also some medications that your doctor could prescribe for you that might take away even some of the desire to smoke</i>	Reference
	well, it's great that you're open to the idea of considering quitting smoking, even if it's not something you're actively pursuing right now. There are many reasons people choose to quit, such as personal health benefits, financial savings, and reducing exposure to secondhand smoke for others. If you were to consider quitting, what do you imagine could be a motivating factor for you?	Simulated
Client	<i>okay well that's good information I mean I'm not gonna do anything today I just need to know my options but that's encouraging if I decided to quit smoking I do have have at least tools to help me quit</i>	Reference

* continue on next page

Table S27. (Continue Table S26) Second Part of the Illustrative Example of Single-Response Simulation Method by LLMs. This method involves the actual responses of both therapist and client from a reference dataset, which are written in *italic*, and simulated responses of LLM Therapists. This full example is the simulated conversation from GPT-4 on one of the conversations using High-Low quality dataset as reference.

Role	Utterances	Dataset
Therapist	<i>sure and back to that skill you mentioned you said you're a - what do you think it would take to move you up to a four or five</i>	<i>Reference</i>
	Absolutely, it's great to know your options. And just by talking about it today, you're already taking a step towards considering your choices. Remember that change takes time and it's completely normal to not feel ready to quit just yet. If there comes a time when you'd like to explore this further or if you have any questions, don't hesitate to reach out. In the meantime, regarding your current situation at work, you might consider finding alternative ways to manage stress or take breaks without smoking. This could involve quick breathing exercises, going for short walks, or finding another enjoyable activity during breaks. If you have any other concerns or ...	Simulated
Client	<i>I don't know I guess if my friends like if one of my friends decided like quit with me that would probably help too but I all my friends love to smoke so I don't see that happening in the near future</i>	<i>Reference</i>
Therapist	<i>sure I wonder if now that your workplace is going smoke-free if there will be some people at work who are planning to quit</i>	<i>Reference</i>
	I understand that having a support system can be helpful in making a change. It sounds like right now, you're not quite ready to quit smoking, but you're open to considering it if you had the right support or motivation. Remember that quitting is a personal decision, and when you decide that you're ready to make that change, there are tools and resources available to help you along the way, like we discussed earlier. In the meantime, it might be helpful to think about the reasons you might want to quit in the future, such as health benefits or saving money. Take some time to evaluate what's important to you and what role ...	Simulated
Client	<i>that's a good point I hadn't really thought about that that's a good idea</i>	<i>Reference</i>
Therapist	<i>okay well it sounds like you're really here just for information today I'm happy to help you pick out some of these products if and when you decide that the time is right for you</i>	<i>Reference</i>
	yeah sometimes having a support system can make all the difference in making a change like this. So I encourage you to keep an open mind and maybe talk to some coworkers, see if any of them are planning to quit. It could be a good opportunity for you to explore quitting together and having that support. In the meantime, feel free to reach out to me if you have any questions or want to discuss your progress. And remember, quitting is a personal decision, so ultimately it's up to you to decide when and if you're ready for that change. I am here to help and support you throughout the process. Take care.	Simulated
Client	<i>okay well thanks a lot</i>	<i>Reference</i>

Table S28. Illustrative Example of Full Conversation Simulation Method by LLMs. The reference conversations have been fed into LLM to simulate both therapist and client. This full example is the simulated conversation from Llama-70b on one of the conversations using High-low quality dataset as reference.

Role	Utterances
Client	Listen, I'm in a bit of a pickle here, just can't shake the feeling that I'm not entirely prepared to quit just now. You understand, right?
Therapist	Absolutely! And that's completely normal. Quitting an addiction is never easy and it takes time for people to feel ready. Can you tell me what specific concerns are holding you back from quitting? Maybe we can tackle those together so that you'll be more confident when deciding to stop.
Client	I guess I'm just dealing with a lack of faith in myself... fearing that I'll cave and revert back to my old ways. It just feels like such a gigantic hurdle.
Therapist	That self-doubt can indeed make things seem insurmountable at times, but keep this in mind - it's part of human nature to waver between wanting positive change yet being scared of failing at it. Now, let's consider strategies to boost your confidence. Have you tried visualization exercises? They often prove useful for identifying and fortifying one's personal motivations to overcome challenges like these. Would you care to give them a shot together with me now?
Client	Yes, I'm open to anything really but visualization doesn't sound like it'll do the trick. You have to remember, I've been through tough situations before. My divorce was a real battle.
Therapist	That's understandable; visualizations may not work equally well on everyone. Let us think about other approaches, then. Perhaps examining past successes could offer some insight into how capable you truly are. Could you recall instances where, despite facing difficulties, you were still able to rise above them? How did you achieve success during those trying periods? What inner strengths played a role there? Building upon and applying those same strengths might empower you in quitting successfully.
Client	That's true, I definitely was able to get through the divorce... But it was different, you know? It was a different kind of struggle. I'm just not certain that I can apply the same tactics to this situation.
Therapist	That makes sense. Each person's journey is unique, and comparing yourself to others or past experiences isn't always helpful. Instead, let's explore some practical tools you can use when cravings strike. Mindfulness practices, deep breathing exercises, or healthy distractions (such as taking walks) tend to work effectively for many individuals battling addictions. Are any of these methods something you'd want to try out whenever you encounter triggers or yearnings? Also, please keep in mind that setbacks occur along the road to recovery; they don't define failure. Instead of viewing slip-ups as disasters, consider them learning chances and adjust your plan accordingly. Having a backup strategy will strengthen your commitment. Does that resonate with you?
Client	Yes, I like the idea of having some practical tools and a backup plan. I just don't want to let myself or my family down. They're rooting for me, and I couldn't bear disappointing them or myself. I don't think I'm ready to go cold turkey yet.
Therapist	That's perfectly fine. Gradual reduction tends to work better than attempting abrupt withdrawal anyway. You could gradually lessen the number of cigarettes or drinks until reaching none at all. This way, you ease into it instead of putting excessive pressure on yourself. Moreover, nicotine replacement medication exists for those finding it hard to cease smoking rapidly. Do any of these suggestions appeal to you? We're steadily uncovering potential solutions that suit your preferences, which helps build up confidence as you progress.
Client	Yes, I've had some luck in the past with cutting back day by day... And I've used nicotine gum, it's just that the taste was unpleasant. But it did help a bit.
Therapist	Great! Nicotine gum might aid you again since you found it effective previously, albeit with an unpleasant flavor. Another option is prescription medicines like bupropion (Zyban), proven to reduce withdrawal symptoms and cravings. Of course, we'd need to discuss this further with your physician first before exploring that route. Slowly reducing consumption seems agreeable to you, too. As we proceed, you can adapt strategies according to your comfort level. Realize that your loved ones desire nothing more than seeing you happy and healthy; if setbacks arise, communicate that with them and seek their assistance. Their encouragement stays with you throughout this process. Is there anything else I can assist you within this conversation today? Has our discussion given you hope or any ideas regarding moving forward?
Client	Definitely, I agree with the gradual reduction strategy, I think pacing myself is the way to go. I'll also look into the gum again, though I can't promise anything about the taste! Your support is reassuring; I definitely feel less alone in this battle.
Therapist	Wonderful! Embracing gradual reduction indicates wisdom and maturity—a wise decision considering your background. Keep investigating various methods, including the gum or medicinal alternatives available following consultation with a doctor. Feel free to reach out whenever you require additional reinforcement or guidance since knowing you possess the necessary capabilities means overcoming this challenge surely. Remember, baby steps count toward lasting achievements. Believe in yourself and acknowledge the efforts made up to this point. Continue drawing upon internal resilience; it served you well during past obstacles. Please stay connected and remember, together, we can navigate anything life presents. Best wishes on your path ahead!

Table S29. Illustrative Example of Full Conversation Simulation Method by LLMs. The reference conversations have been fed into LLM to simulate both therapist and client. This full example is the simulated conversation from GPT-3.5 on one of the conversations that used HOPE dataset as a reference.

Role	Utterances
Client	I've been considering making a change in my diet and cutting out fast food altogether. You know, like having more structured meals and focusing on healthier options. I guess I should start by going to the grocery store more often.
Therapist	Thank you for sharing your thoughts about making a change in your diet. It's wonderful to hear that you want to take a positive step toward a healthier lifestyle. It's great that you recognize the importance of having structured meals and focusing on healthier options. It sounds like you're considering a gradual shift towards healthier habits, which is a good approach!
Client	Yes, that's true. Recently, I've just been getting food at work or picking up fast food on my way home. Okay. It's been difficult to find the motivation and energy to go grocery shopping, especially since I've been working extra hours to save up for the baby. Scheduling time for grocery shopping has been a challenge. I used to go every Sunday and even had a shopping list ready beforehand. Perhaps getting back to that routine could help. On Sundays, I don't have to work until the afternoon, so maybe I can try going around 10 in the morning. To motivate myself, I've found that giving myself a small reward works well. There's a fantastic bakery on my way back from the grocery store, so maybe I could treat myself there if I stick to my shopping plan.
Therapist	It's great to hear how you are trying to motivate yourself to stick to your new structure around grocery shopping. It seems like you are thinking about the challenges you might face and coming up with good strategies for how to work around them. Having a shopping list is a clear and helpful way to stay focused on buying only the foods you need. And I like that you are rewarding yourself for sticking to your plan! It's important to find ways to positively reinforce yourself for good behavior. If you don't mind me asking, how do you think cutting out fast food and eating healthier will impact your life?
Client	I feel that cutting out fast food and adopting healthier eating habits will make a positive impact on my life in various ways. Okay. For starters, I believe it will improve my overall health and well-being, which is critical, especially with a baby on the way. I also expect better energy levels and perhaps even some weight loss in the long run. Additionally, I think that by sticking to a structured grocery shopping routine, I'll develop better time management skills and it might even save me some money. Developing these good habits now will not only be beneficial for me but also for my growing family as we can all strive for a healthier lifestyle.
Therapist	Absolutely! It's great to focus on the benefits of making healthy changes, not only for yourself but for your growing family too. You bring up great points about benefits related to improved health and well-being, better energy levels, weight loss, and even better time management skills. These are all great things to look forward to and can help you find the motivation to keep making progress towards your goals. Do you have any specific steps in mind to help you stay on track with your healthy eating plan?
Client	Yes, I've been thinking about some specific steps I can take to stay on track with my healthy eating plan. Okay. Firstly, I want to establish a weekly meal plan that includes nutritious options and a variety of fruits, vegetables, lean proteins, and whole grains. Having a meal plan will help me create my shopping list, so I know exactly what to buy when I go to the grocery store. Also, preparing meals in advance can save me time during the week, making it less tempting to grab fast food. Another step I'd like to take is to find healthier alternatives for my favorite fast food items so I can still enjoy the flavors I like in a more nutritious way. Lastly, I think it's essential to have a support system, so I plan to involve my partner in these changes and share our progress to keep each other accountable.
Therapist	These are all great steps towards a healthier lifestyle. Establishing a weekly meal plan that includes nutritious options is a smart way to stay on track, and preparing meals in advance is an excellent time-saving strategy. Finding healthier alternatives for your favorite fast food items is a great way to stay motivated and still have the flavors you enjoy. Involving your partner in these changes is a brilliant idea too, and having someone for accountability can be very helpful. Change can be challenging, but with strategies like these, you are setting yourself up for success. Do you think you need some support in sticking to this plan?
Client	I definitely think having some support in sticking to this plan will be helpful. Okay. As I mentioned, my partner will be a significant source of support since we're in this together. We can encourage each other and celebrate our progress, which makes the journey more enjoyable. Besides my partner, sharing my goals and plans with close friends and family members could also provide me with some external encouragement and motivation. Maybe joining a group or forum focused on healthy eating could be of help too, as it would connect me with people who are going through similar experiences, and we can learn from each other's successes and challenges. Do you have any other suggestions for support that I can consider?
Therapist	It sounds like you have thought of some great sources of support already. Your partner, close friends, and family members can all offer encouragement and accountability. Joining a group or forum dedicated to healthy eating in your community or online can be helpful in finding like-minded people who can offer understanding and support. Additionally, if you find that you're having difficulties sticking to your plan, you can always consider working with a registered dietitian or health coach who can provide expert guidance and support in a more structured way. Remember, it's okay to ask for help along the way, and you're not alone in making these changes.
Client	Thank you for the suggestion of working with a registered dietitian or health coach. That's something I'll definitely consider if I find it challenging to make progress on my own. Okay. It's reassuring to know that there are different options for support available. I realize that making these changes won't be an entirely smooth process, and I might encounter setbacks, but having a support network can help me stay focused on my goals. I appreciate your insights and encouragement in helping me take these steps towards a healthier lifestyle.
Therapist	You're very welcome! Remember, making lasting changes to our habits takes time, effort, and commitment, so don't get discouraged if it doesn't happen overnight. Keep focusing on the benefits of adopting healthier eating habits and use the strategies and resources we've talked about to help you stay on track. I am glad to have been able to offer some support and perspective, and I look forward to hearing about all the positive changes you make moving forward. Don't hesitate to reach out if you need additional support or guidance.

Figure S1. Different prompts variations for classifying conversational behavior embedded in both therapist and client utterance in the counseling setting. We explored two variations of prompting – 1) Prompting with definitions of different conversational behaviors, with the sources of two studies^{43,44}. 2) Prompting with the definitions and also examples from annotated dataset. For 2), we explored two ways of asking questions, which are the multi-label setting and binary-label setting. We experimented with different variations of these prompts (e.g., using different language for the instructions) but observed similar results. See the detailed methodology description in Methods.

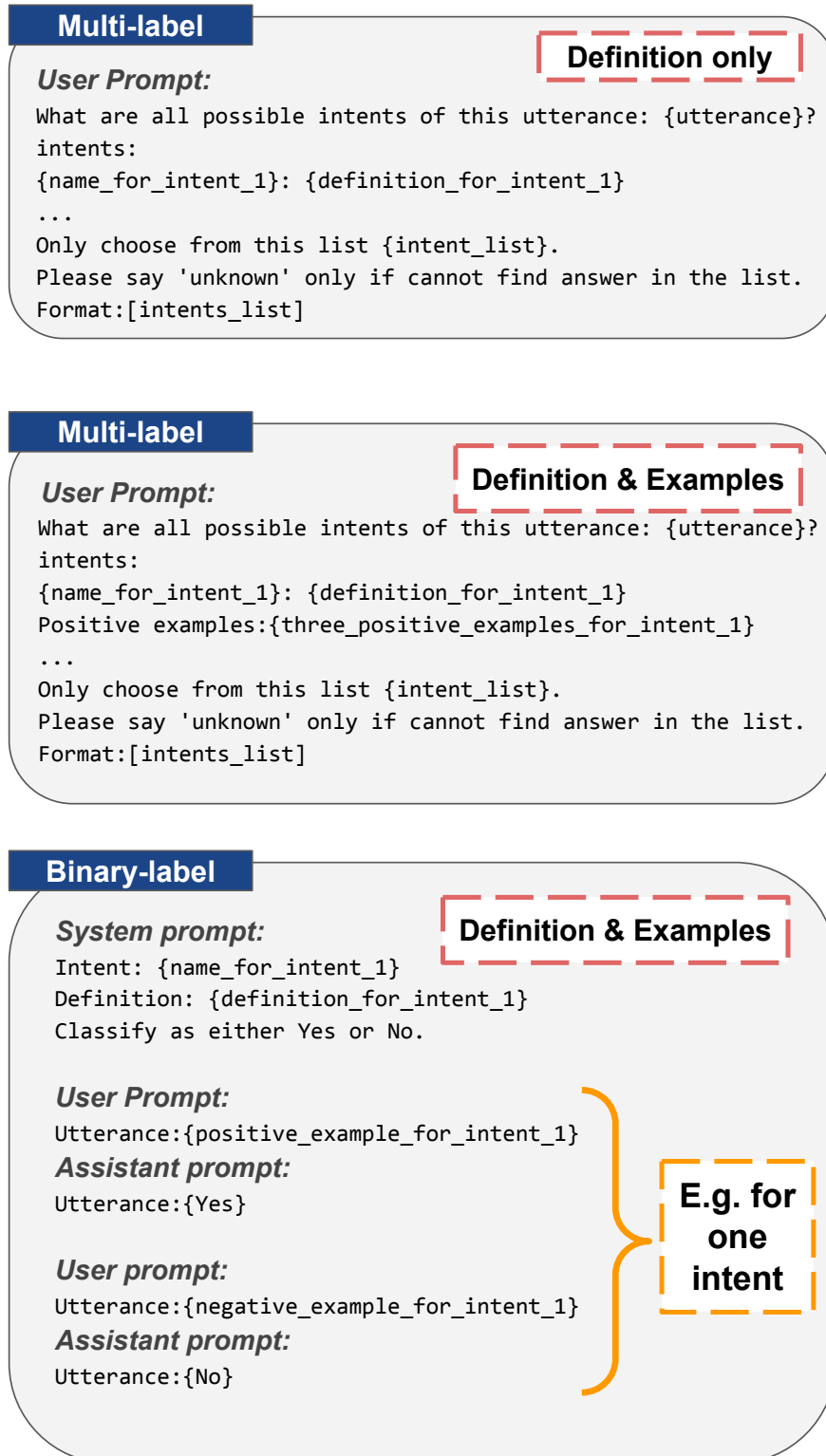


Figure S2. Instruction prompts given to LLMs for simulating counseling conversation using two types of methods (single response and full conversation): 1) using LLMs to write a single response within an ongoing conversation between human therapist and client. 2) creating a full and multi-turn conversation between LLMs and clients. See the detailed description in Methods.

Therapist

Single Response & Full Conversation

User Prompt:

Act as if you're a professional therapist. You provide evidence-based therapy, utilizing motivational interviewing techniques, to help clients in making behavioral changes, such as quitting smoking or alcohol consumption. You should maintain your therapist persona while responding. Communicate in a conversational style, mirroring the style of previous therapist responses. You should conclude your session by saying "###END###"

For High-low Quality as reference dataset only

Client

Full Conversation

User Prompt:

You will be provided with a conversation between a client and a therapist. Your task is to simulate the same client talking to a different therapist in a parallel universe. You can ignore that the previous chat ever happened. While the context of the previous conversation should not influence this session, it should guide you on how the client communicates, including their tone of speech, sentence structure, and the manner in which they address particular topics or concerns. Essentially, you're creating a new conversation but with the client's life situation and their response pattern maintained. Only generate the client utterances.