Original Paper

Modeling of Therapist-Client Dynamics in Psychotherapy Using Large Language Model–Based Assessments: Retrospective Observational Study of 1573 Psychotherapy Sessions

Angela H Chen, PhD; Siwei Jin*, MSc; Canwen Wang*, MSc; Holly A Swartz, MD; Tongshuang Wu, PhD; Robert E Kraut, PhD; Haiyi Zhu, PhD

Carnegie Mellon University, Pittsburgh, PA, United States

*these authors contributed equally

Corresponding Author:
Haiyi Zhu, PhD
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA, 15213
United States
Phone: 1 412 268 6523
Email: haiyiz@andrew.cmu.edu

Abstract

Background: Psychotherapy is a primary treatment for many mental health conditions, yet the dynamic interplay among therapist behaviors, client responses, and the therapeutic relationship remains difficult to disentangle at scale. Traditional process research relies on labor-intensive human coding, limiting the ability to model moment-to-moment therapeutic dynamics across large datasets.

Objective: This study aimed to develop and validate a computational framework for modeling therapist-client interaction processes and to examine how therapist behaviors and relational qualities relate to client self-disclosure and emotional expression in psychotherapy sessions.

Methods: We developed automated assessment pipelines using large language models (LLMs) to measure therapist behaviors (empathy or exploration), relational quality (rapport), and client outcomes (self-disclosure, self-directed negative emotion, or outward-directed negative emotion). Model-generated scores were validated against human annotations. We then applied these measures to approximately 2000 hours of psychotherapy transcripts from the Alexander Street corpus. This yielded an estimation sample of 119,233 client utterances from 1573 sessions and 37 clients. Because utterances are nested within sessions and sessions within clients, we estimated a saturated multivariate structural equation model with SEs clustered at the session level. Results were confirmed in a multilevel regression model with random intercepts for session and client.

Results: Following comparative validation of 6 candidate language models, automated measures showed good to excellent agreement with human annotations (intraclass correlation coefficient [ICC]=0.60-0.81) for rapport (ICC=0.81), self-disclosure (ICC=0.78), empathy components (ICC=0.60-0.75), and emotion measurements (ICC=0.66-0.80). Structural equation modeling analyses revealed that therapist empathy and exploration were each associated with increased client disclosure (empathy: β=.086, 95% CI 0.078-0.095; exploration: β=.086, 95% CI 0.078-0.094; both P<.001) and increased expression of self-directed negative emotions, comprising sadness, anxiety, fear, and depression (empathy: β=.118, 95% CI 0.110-0.127; exploration: β=.078, 95% CI 0.070-0.087; both P<.001). Both behaviors were also associated with reduced expressions of enjoyment (empathy: β=–.037; exploration: β=–.041; both P<.001). Surprisingly, rapport in the prior session showed no association with client self-disclosure or negative emotion expression.

Conclusions: This study demonstrates that LLM-based measurement combined with structural modeling can capture core therapeutic processes at scale. Empathy and exploration showed small but consistent turn-level associations with client disclosure and emotional expression, whereas prior-session rapport did not show a robust association. These findings provide a computational foundation for precision modeling of psychotherapy processes and may offer actionable insights for scalable therapist training and AI-supported clinical education systems.

(JMIR Ment Health 2026;13:e95389) doi: 10.2196/95389

XSL•FO
RenderX

KEYWORDS

AI-supported training; automated behavioral coding; client disclosure; digital mental health; large language models; psychotherapy process modeling; structural equation modeling; therapeutic alliance; therapist empathy

Introduction

Psychotherapy is an effective, evidence-based treatment for many common mental health conditions, yet access and quality remain persistent challenges. Digital mental health technologies—including teletherapy platforms [1], conversational agents [2], clinician-facing dashboards [3], and training or supervision tools [4]—have expanded the reach of care and created new opportunities to monitor and improve treatment delivery. However, improved access alone does not guarantee effective care: clinical benefit depends heavily on *how* therapeutic interactions unfold, including the clinician's moment-to-moment responses, the relational bond between therapist and client, and clients' in-session engagement and emotional processing [5-7].

A central limitation in both psychotherapy research and digital mental health implementation is that the mechanisms linking therapist behavior to patient outcomes are difficult to measure at scale. Therapeutic effectiveness is multifactorial, and the specific effects of discrete therapist behaviors on patient responses are often not characterized well [8]. In routine care and many digital health studies, evaluation relies on self-report symptom scales and postsession assessments of therapist-client interaction (eg, the Patient Health Questionnaire-9, General Anxiety Disorder-7, Working Alliance Inventory [WAI], and qualitative reflections) [9-12]. While indispensable for capturing patient experience and clinical status, these measures are typically collected infrequently, provide limited resolution about what happens within sessions, can be influenced by recall and rater biases, and offer little actionable information for real-time clinical decision support or training.

A complementary approach is to quantify the therapy process directly from session content. Prior work suggests that patient self-disclosure relates to anxiety and depression symptoms [13], that in-session emotional expression is associated with therapeutic change [14], and that the therapeutic alliance—often conceptualized as bond (rapport), agreement on goals, and collaboration—is robustly linked to treatment success across modalities and populations [5,15]. Yet, measuring these constructs at the granularity needed to support scalable evaluation and feedback remains challenging. Manual behavioral coding is labor-intensive and difficult to deploy in real-world digital care settings, motivating computational approaches that can extract clinically meaningful signals from therapy data at scale [16-19].

In this study, we develop and evaluate a transcript-based computational framework for quantifying therapist-client dynamics in psychotherapy, with an emphasis on constructs that are clinically interpretable and potentially actionable in digital care contexts. We focus on three classes of variables: (1) therapist behaviors (empathic responses, reflective listening, interpretations, and exploratory queries), (2) therapeutic relationship quality (rapport as a bond-based indicator of therapeutic alliance), and (3) client responses (self-disclosure and emotional states). Our objective is twofold: first, to develop accurate and scalable measures of these constructs from real-world psychotherapy transcripts; and second, to empirically model how therapists' behaviors and relationship context are associated with clients' immediate responses during sessions.

To address measurement, we used large language models (LLMs) to score therapist empathy components (interpretation, exploration, emotional reaction, and reflection [17]), therapist-client rapport [20,21], and client responses including self-disclosure [22] and emotions (anger, contempt, disgust, enjoyment, fear, sadness, surprise, anxiety, and depression) [23]. We applied these measures to psychotherapy transcripts from the Alexander Street (AS) corpus and validated automated measurements against ratings from trained human annotators. Across constructs, LLM-derived scores aligned well with human judgments (with intraclass correlation coefficient [ICC] ranging from 0.60 to 0.81), supporting the feasibility of using LLMs as a scalable measurement layer for clinically grounded process analysis.

To model the associations between therapist behaviors and therapist-client rapport with client behaviors, we used structural equation modeling (SEM) to test theory-motivated pathways linking therapist behavior, rapport, and client outcomes. Prior work has used SEM to examine causal and temporal relationships among communication variables and alliance-related constructs [24]. Building on this tradition, we evaluated whether therapists' immediate empathic and exploratory behaviors are associated with clients' subsequent disclosure and emotional expression, and whether cumulative rapport from prior sessions adds explanatory value beyond moment-to-moment behaviors. We tested the following four hypotheses: (H1) therapist empathy and exploration (ie, encouragement of deeper discussion) in the immediately preceding turn are associated with client self-disclosure; (H2) cumulative rapport built prior to the current session is associated with client self-disclosure; (H3) higher levels of therapist empathy and exploration are associated with greater expression of negative emotions; and H4) higher levels of cumulative rapport between therapist and client are associated with greater expression of negative emotions. Our findings indicate that therapist empathy and exploration are associated with increased client disclosure and changes in emotional expression, whereas rapport in the prior session shows no association with client self-disclosure or negative emotion expression.

By combining validated, transcript-derived process measures with theory-driven modeling, this work contributes both measurement methods and empirical evidence that are directly relevant to digital medicine and health care delivery. The framework supports scalable assessment of therapeutic process variables that are difficult to capture with conventional symptom scales alone, and it enables empirical tests of hypotheses about how clinician behaviors relate to patient responses during care.

XSL•FO
RenderX

These capabilities can inform digital mental health platforms, quality improvement initiatives, clinician training and supervision tools, and future clinician-facing decision support systems that require fine-grained, interpretable signals about therapeutic interactions.

Methods

Dataset

AS [25] is a trusted source of psychological counseling materials for academics, as it offers access to real clinical interactions that are typically hard to gather in large numbers due to practical limitations in clinical environments, including privacy concerns. The dataset comprises one-on-one therapy sessions transcribed by professional human transcribers who listened to the original audio recordings.

All transcripts were carefully anonymized following ethical research standards and Health Insurance Portability and Accountability Act (HIPAA) rules. Personal details were removed to protect client privacy while maintaining clinical relevance. Therapist names are unknown, but clients likely worked with the same therapist across sessions, while therapists probably worked with multiple clients.

We used strict selection criteria to maintain dataset consistency and meet our research goals. We kept only one-on-one counselor-client sessions and removed all group therapy, family therapy, and sessions with multiple participants. The final processed dataset includes 1610 unique therapy sessions from 37 clients, covering 118 presenting conditions (eg, anxiety, depression, and burnout; clients could present with more than one condition) and a range of topics such as childhood trauma, sexual abuse, and social anxiety. The dataset comprises 243,407 utterances, counting both therapist and client turns, with a mean session length of 151.2 (SD 101.6) utterances. The analyses reported below model clients' behavior in a sample of 119,233 utterances from 1573 sessions and 37 clients. We exclude the first session from each client because, by definition, it has no rapport value measured during a prior session. In this study, an utterance refers to a single speaking turn in the therapy dialogue—an instance in which either the client or the therapist speaks before the conversational floor switches. The dataset also allows us to study concepts that can only be captured in a chunk of the conversation or across multiple turns, rather than at the single turn level.

Ethical Considerations

This paper used anonymized AS therapy transcripts under a nonexclusive license, in a manner consistent with US fair use provisions and international law, and solely for research, educational, or other strictly noncommercial uses that do not involve public performance. All coauthors are authorized users of the dataset. Per National Institutes of Health guidance, secondary research using data for which the research team has no access to identifiers does not meet the definition of human subjects research. The transcripts were deidentified before release, and the research team had no access to identifiers or to any key linking transcripts to individuals. This study therefore does not constitute human subjects research, and institutional review board review was not required. The generative AI API we use to measure clinical behaviors does not use data submitted through the API to train or improve its models. Therefore, our use of generative AI models does not result in the AS dataset being used for model training.

Measurements of Key Therapy-Related Constructs

Overview

We first describe the constructs we measure (constructs). We then describe the full development pipeline and validation studies (prompt design and iteration; validation with human annotators; and model selection). The validation is conducted on a stratified subset of the AS dataset (see Stratified Sampling). The SEM modeling testing the relationships among rapport, therapist empathy, and client outcomes uses the full AS dataset.

Constructs

Overview

We automatically measure the following 3 types of constructs: client-side (ie, self-disclosure and emotions), therapist-side (ie, empathy skills), and the relationship (ie, rapport).

We required each construct to satisfy 3 conditions. First, it had to be operationalized by an existing validated framework, so that our prompts encode an established definition rather than a new one. Second, it had to be observable in transcribed text alone, since our corpus contains no audio or video. Third, it had to be measurable at the granularity our modeling requires, meaning turn level for therapist behaviors and client responses and session level for relational quality. Applying these criteria yields the constructs described below.

Constructs Related to Therapist Behavior

Overview

To ensure clinical validity and interpretability, each construct was grounded in established psychological theory and prior empirical frameworks. Because the dataset lacks reliable information about the treatment modalities or specific intervention techniques used by individual psychologists, we do not attempt to infer or evaluate modality-specific methods. Instead, we focus on cross-cutting psychotherapy "common factors"—relational processes shared across approaches, such as empathy and rapport [26,27]. These theoretical and empirical foundations guided prompt design for each construct:

Empathy

We model empathy through 4 interrelated components. The first 3 are drawn from the Empathy in Text-Based, Asynchronous Mental Health Conversations (EPITOME) framework [17], which conceptualizes empathic communication as comprising: emotional reactions (expressing emotions such as warmth, compassion, and concern), interpretations (communicating an understanding of the client's feelings and experiences), and explorations (attempting to understand the client's feelings and experiences through questions and probes). Each mechanism is scored on a 3-point scale (0=not present, 1=weak, 2=strong), with total scores ranging from 0 to 6. The fourth dimension, reflection, follows definitions from motivational interviewing [28-30] and is treated as a related but

XSL•FO
RenderX

distinct construct that reinforces empathy. Reflection is defined as a therapist response that mirrors or builds upon the client's preceding statement. It is typically categorized as *simple* (paraphrasing client statements) or *complex* (inferring deeper meaning or emotion). Following prior computational work such as the prompt-aware margin ranking framework [31], we adopt a binary reflection vs nonreflection scheme, using the simple-reflection/complex-reflection distinction as guiding examples within the prompt to calibrate model judgment and improve interpretability.

Constructs Related to the Therapist-Client Relationship (Rapport)

The therapeutic alliance that the therapist and client develop over time is a robust predictor of improvements in clients' psychological well-being [26,27]. Here, we focus on *rapport*—the felt bond and harmonious responsiveness that undergirds alliance [20]. Modeled as the "bond" component of the therapeutic alliance, rapport reflects mutual liking, confidence, trust, and affective attunement. We adopt 4 items from the Working Alliance Inventory-Observer Short version (WAI-O-S) Bond subscale [32] to measure rapport on a 7-point Likert scale. Because rapport is presumed to develop slowly over time, we measure it at the prior session rather than the prior utterance.

Constructs Related to Client Behavior

Overview

We measure 2 utterance-level components of client behavior: self-disclosure and their emotions. Together they characterize openness and affect as they unfold in session. Self-disclosure is the verbal revelation of personal information, experiences, or feelings. Emotion is central to therapeutic interaction; clients are more likely to improve when they reveal their emotions [14,33].

Self-Disclosure

Grounded in intimacy theory [34,35], we classify utterances on a 3-level scale: general (G) for factual or impersonal content, medium (M) for personally relevant but nonsensitive expressions, and high (H) for intimate or emotionally vulnerable disclosures [36].

Emotion

We assess emotional expression using 9 subconstructs. Our taxonomy integrates Ekman's 7 universal emotions [23]—enjoyment, sadness, anger, disgust, contempt, fear, and surprise—augmented with 2 clinically relevant categories: anxiety and depression. The extent to which each of the 9 emotions is expressed in each client utterance is rated on a 5-point Likert scale ranging from "very slightly or not at all" to "extremely."

Prompt Design and Iteration

Each construct is operationalized through a dedicated LLM-based prompt aligned with validated psychological definitions (Table 1). All prompts share a common structure specifying the evaluator's role, output format, and evaluative constraints, while rubrics and rating scales are tailored to each construct (Table 2). For maximal generalizability [37], we primarily used zero-shot prompting, in which the prompts assume general knowledge and provide no examples. Rubrics were derived from clinical theory for all constructs (eg, [17,28,32,34,35]; Table 1), with inline examples added only for constructs requiring subtle distinctions [38]. Following recent work on inference-time computation [39,40], the empathy and self-disclosure prompts additionally use chain-of-thought reasoning, asking the model to justify its rating by citing specific elements of the utterance [41]. For each construct, the model first generated a rationale grounded in conversational evidence before producing the final rating. To limit run-to-run variability, every utterance was scored in 3 independent runs, and the ratings were averaged. Full prompt wordings appear in Multimedia Appendix 1.

We used construct-specific context windows, supplying only as much preceding dialogue as was needed to disambiguate relational or emotional cues and adapting window length to each construct's evaluative demands (details shown in Table S1 in Multimedia Appendix 2).

We refined each construct prompt iteratively, evaluating variants against the annotated datasets summarized in Table S2 in Multimedia Appendix 2. Early exploration revealed consistent differences across foundation models in stability and sensitivity to prompt phrasing; GPT-4o-mini was selected as the primary model for its agreement with human annotators, its stability across prompt revisions, and the cost of annotating 243,407 counselor and client utterances; this initial choice was subsequently reexamined against 5 further models, as reported in the Model Selection section below. Two lessons recurred in the iterations: requiring the model to cite specific conversational cues improved scoring consistency, and simplifying definition phrasing reduced noise more effectively than elaborating it.

XSL•FO
RenderX

Table 1. Summary of background clinical theories.

Category and constructs	Key clinical theory and framework
Therapist-side behaviors	
Empathy-emotional reaction/interpretation/exploration	EPITOME[a] framework [17]
Empathy-reflection	Motivational interviewing criteria [28]
Client-side behaviors	
Self-disclosure	Intimacy framework [34,35]
Emotion	Ekman's Basic Emotions [23]; PANAS[b] scale [42]
Therapeutic relationship	
Rapport	WAI-O-S[c] Bond items [32]: items 3, 5, 7, 9

[a]EPITOME: Empathy in Text-Based, Asynchronous Mental Health Conversations.

[b]PANAS: Positive and Negative Affect Schedule.

[c]WAI-O-S: Working Alliance Inventory-Observer Short version.

Table 2. Key components of each prompt, including the evaluator role, core definition phrase, and output format.

Construct	Assigned role (quoted from full prompts[a])	Definition (quoted from full prompts[a])	Output format
Empathy			
Emotional reaction/interpretation/exploration	"...a professional evaluator specializing in assessing therapist empathy..."	"Emotional reactions: expressing warmth, compassion, or concern. —Strong (2): explicitly specifies emotions. —Weak (1)..."	3 mechanisms (0-2 each)
Reflection	"...a professional evaluator specializing in motivational interviewing..."	"Reflection: a therapist's response that demonstrates understanding by restating or adding insight, or maintaining alignment with the client's statements..."	Binary: reflection/nonreflection
Self-disclosure	"...a professional psychologist specializing in communication analysis and self-disclosure..."	"M (medium disclosure): the client shares non-sensitive personal experiences, plans, or general life updates (e.g., 'I started a new job last week...')..."	Ordinal: general/medium/high
Emotion	"...an expert conversation evaluator... skilled in identifying subtle emotional cues..."	"Assess the central utterance using nine emotions. Rate each emotion on a 5-point Likert scale..."	9 emotions × 5-point Likert
Rapport	"...an expert conversation evaluator assessing the therapeutic alliance..."	"Assess bond quality across mutual liking, confidence, appreciation, and trust. Each rated on a 7-point Likert scale..."	4 aspects + overall (1-7)

[a]Full prompts are in Multimedia Appendix 1.

Validation With Human Annotators

To evaluate the effectiveness and reliability of our LLM-based annotation framework, we conducted human validation on a stratified sample of model-labeled utterances and segments.

Stratified Sampling

Our full dataset included 243,407 utterances drawn from 1610 unique therapy sessions. Given this scale, we used stratified sampling strategies tailored to each construct to select representative and label-balanced subsets for validation.

Recruitment and Participation

Annotators were native English-speaking university students who were either psychology majors or had completed at least 2 psychology- or counseling-related courses. Utterance-level classification tasks (self-disclosure and the empathy components) were each rated by 3 annotators; rapport and emotions, which involve longer context windows and multidimensional rating scales, were each rated by 6. In total, 27 annotators took part, each contributing between 15 and 17 hours of annotation.

Training

All annotators participated in structured training workshops tailored to each construct. Each session included definition reviews, guided examples, practice annotations, and collaborative discussions to ensure alignment in interpretation.

XSL•FO
RenderX

Model Selection

We applied identical prompts, with no model-specific tuning, to 6 models from 3 vendors: GPT-4o-mini, GPT-5.5, Claude Opus 4.8, Claude Sonnet 4.6, Gemini 3.5 Flash, and Gemini 3.1 Flash-Lite, scoring each against the same human-annotated validation sets. Exact version strings and API access dates are reported in Table S4 in Multimedia Appendix 2 and pinned in our released analysis code.

Agreement with human annotations was broadly consistent across models (Table S3 in Multimedia Appendix 2): overall ICC ranged from 0.713 to 0.801 (SD 0.032), and every model exceeded 0.70 despite spanning roughly an order of magnitude in cost. We take this as evidence that measurement accuracy rests on the prompts and their grounding in established coding frameworks rather than on any particular vendor's model.

The exception was emotion. Under GPT-4o-mini, 3 of the 9 emotion categories fell below ICC 0.50 (fear: ICC=0.453, anxiety: ICC=0.465, depression: ICC=0.489) and the mean emotion ICC was 0.606 (SD 0.124), which we did not regard as adequate for the outcome measures carrying H3 and H4. On the same validation sets, the best-performing model reached a mean emotion ICC of 0.743 (SD 0.053), with every category above 0.65 (Table S3b in Multimedia Appendix 2). We therefore measured all therapist behaviors (empathy components, exploration, and reflection), rapport, and client self-disclosure with GPT-4o-mini, and rederived the nine emotion measures with Gemini 3.5 Flash.

Interpretation of the Validation Results

Table 3 shows that, according to established guidelines [43], all of our human-human and LLM-human ICC values fall in the "good" (0.60-0.74) to "excellent'' (0.75-1.00) ranges.

Each of the 4 empathy dimensions was validated on 50 utterances by 3 raters, and all fell in the "good" to "excellent" range. Agreement was highest for interpretation (ICC=0.749) and exploration (ICC=0.732), and lower for emotional reactions (ICC=0.601) and reflection (ICC=0.658)—the same 2 dimensions that showed the lowest human-human reliability, suggesting they involve more subjective affective inference that is harder for both humans and LLMs to judge consistently.

Self-disclosure (150 utterances, 3 raters) and rapport, measured at the segment level (200 samples, 6 raters), both showed good to excellent agreement (ICC=0.781 and 0.806, respectively), indicating that the model captures both the depth of client expression and multiturn relational signals.

Emotion was measured with Gemini 3.5 Flash and validated on 450 utterances spanning 9 affect categories, annotated by 6 raters. Agreement averaged ICC 0.743 (SD 0.053) across the 9 categories. Emotions with distinct verbal signatures were rated most reliably—disgust (ICC=0.799), enjoyment (ICC=0.798), anger (ICC=0.796), and contempt (ICC=0.771)—while more introspective or ambiguous states proved harder: anxiety (ICC=0.748) and sadness (ICC=0.733) still reached good agreement, but fear (ICC=0.706), surprise (ICC=0.677), and depression (ICC=0.659) were the weakest in our framework, reflecting the difficulty of inferring complex affect from text alone, without vocal tone or facial expression.

Overall, agreement with trained lay annotators ranged from good to excellent. Our annotators were undergraduate psychology majors or students who had completed at least 2 relevant courses; they were trained on our coding scheme but are not clinicians, so these figures should not be read as agreement with expert clinical judgment. We also note that an ICC above 0.75 is the conventional standard for measures intended to inform clinical decisions, and that our measures do not uniformly meet it.

XSL•FO
RenderX

Table 3. Validation results comparing human-human reliability and model-human agreement across constructs.[a]

Measurement	Human-human, ICC[b]$_{2,k}$ (95% CI)	LLM[c]-human, ICC$_{2,k}$ (95% CI)	LLM-human, Pearson *r*
Empathy			
Emotional reaction	0.634 (0.59-0.72)	0.601 (0.52-0.67)	0.633
Interpretation	0.763 (0.70-0.82)	0.749 (0.69-0.80)	0.814
Exploration	0.826 (0.78-0.87)	0.732 (0.67-0.79)	0.686
Reflection	0.654 (0.60-0.74)	0.658 (0.60-0.71)	0.754
Rapport	0.720 (0.61-0.81)	0.806 (0.75-0.86)	0.676
Self-disclosure	0.749 (0.67-0.81)	0.781 (0.70-0.84)	0.621
Emotions			
Anger	0.799 (0.73-0.85)	0.796 (0.745-0.837)	0.679
Contempt	0.657 (0.55-0.73)	0.771 (0.718-0.812)	0.710
Disgust	0.719 (0.62-0.79)	0.799 (0.744-0.842)	0.711
Enjoyment	0.838 (0.80-0.87)	0.798 (0.732-0.849)	0.673
Fear	0.672 (0.56-0.75)	0.706 (0.626-0.768)	0.576
Sadness	0.719 (0.60-0.79)	0.733 (0.676-0.782)	0.615
Surprise	0.702 (0.61-0.77)	0.677 (0.583-0.749)	0.526
Anxiety	0.757 (0.67-0.81)	0.748 (0.695-0.792)	0.629
Depression	0.710 (0.62-0.77)	0.659 (0.591-0.719)	0.602

[a]Empathy, reflection, rapport, and self-disclosure values were produced with GPT-4o-mini. All 9 emotion values were rederived with Gemini 3.5 Flash following the comparative validation reported in Table S3 in Multimedia Appendix 2; the corresponding GPT-4o-mini emotion values are reported in Table S3b in Multimedia Appendix 2. Because most constructs—empathy mechanisms, self-disclosure, emotions, and rapport—use ordinal or continuous scales, we report the ICC$_{2,k}$, a 2-way random-effects coefficient that captures both absolute agreement and consistency in rating magnitudes [44,45]. Reflection is binary; we report ICC$_{2,k}$ for it as well so that all constructs appear on a common scale. Prompt calibration and model agreement with an external benchmark dataset are reported in Table S2 in Multimedia Appendix 2.

[b]ICC: intraclass correlation coefficient.

[c]LLM: large language model.

Modeling of Therapist-Client Dynamics in Psychotherapy

We used SEM to analyze the AS data in order to capture temporal patterns unfolding between therapist and client behaviors across sessions. To reduce model complexity, we first conducted principal component analysis (PCA) for constructs such as emotion and empathy, which were composed of multiple components [46]. We then specified and fit SEMs using R's *lavaan* package [47]. Because utterances are nested within sessions and sessions within clients, we estimated a saturated multivariate SEM with SEs clustered on a unique session identifier constructed as the combination of client and session [48]. To verify that our conclusions did not depend on this choice of statistical method, we also fitted a multilevel model with random intercepts for session and for client. The 2 specifications agree on the direction and approximate magnitude of every association; the only discrepancy concerns rapport, whose association with self-directed negative emotion is marginal in the SEM and marginally significant in the multilevel model, which is the basis for our statement that the rapport effect is not robust.

Results

PCA Results

Emotions

As shown in Table 4, across 9 client emotion indicators (n=122,939), PCA produced a 4-component rotated structure explaining 78.9% of the total variance. The first component, self-directed negative emotions, showed strong positive associations with sadness (loading=0.88), anxiety (loading=0.85), fear (loading=0.74), and depression (loading=0.88). This component captures inwardly oriented emotional distress; averaging these 4 emotions leads to a highly reliable self-directed negative emotions scale (Cronbach α=0.87). The second component, outward-directed negative emotions, was defined by high associations with contempt (loading=0.91), disgust (loading=0.79), and anger (loading=0.79), representing externally oriented aversive reactions directed toward some external target (eg, "angry at..."). Averaging these 3 emotions leads to a highly reliable outward-directed negative emotions scale (Cronbach α=0.79). The third component corresponded almost exclusively to surprise, which showed minimal cross-loadings, indicating that surprise functions as an independent expressive signal rather

XSL•FO
RenderX

than belonging to either negative-emotion grouping. Similarly, the fourth component corresponded almost exclusively to enjoyment, which showed an extremely high loading (loading=0.98) and low cross-loadings. We treat these 4 components as distinct outcomes: self-directed negative emotions, outward-directed negative emotions, surprise, and enjoyment, each of which functions as an independent expressive signal.

Table 4. Principal component analysis loadings of emotion and empathy.[a]

Item	Self-directed negative emotions	Outward-directed negative emotions	Surprise	Enjoyment
Panel A. Emotions				
Anger	0.236	0.788	0.066	–0.098
Contempt	0.038	0.905	0.026	–0.021
Disgust	0.062	0.788	0.049	0.003
Enjoyment	–0.137	–0.045	0.061	0.984
Fear	0.744	–0.056	0.279	–0.101
Sadness	0.879	0.167	–0.136	–0.077
Surprise	–0.007	0.075	0.929	0.076
Anxiety	0.848	0.065	0.218	–0.147
Depression	0.880	0.129	–0.173	–0.073
Panel B. Empathy				
Reaction	0.494	–0.331	—[b]	—
Interpretation	0.624	0.002	—	—
Exploration	0.209	0.942	—	—
Reflection	0.568	–0.060	—	—

[a]Panel A reports varimax-rotated loadings from the 4-component solution on the Gemini 3.5 Flash emotion measures (n=122,939), which together explain 78.9% of the variance. Panel B reports the therapist-behavior loadings (n=120,277), which together explain 75% of the variance. Both PCA sample sizes exceed the estimation sample of 119,233 because they include utterances from each client's first session, which the models drop for lack of a prior-session rapport value; the empathy solution is estimated on therapist utterances rather than client utterances.

[b]Not applicable.

Empathy

For the 4 therapist-behavior items (n=120,277), PCA revealed a 2-component solution explaining roughly 75% of the variance. The first component, *general empathy*, showed strong associations with interpretation (loading=0.62), reflection (loading=0.57), and reaction (loading=0.49), representing empathic understanding and responsive attunement. Averaging these resulted in a highly reliable general empathy scale (Cronbach α=0.71). The second component, *exploration*, emerged as a distinct factor defined primarily by the exploration item (loading=0.94), indicating a separate behavioral dimension focused on directive inquiry and probing.

SEM Results

Overview

Based on these PCA results, we specified a multivariate SEM in which 5 utterance-level client-related outcome variables (disclosure and 4 emotional expression variables—self-directed negative emotion, outward-directed negative emotion, surprise, and enjoyment) were regressed on utterance-level therapist behaviors (exploration and empathy), session-level rapport from the previous session, and the number of sessions clients have had with their therapist (log-transformed session index).

We estimated a structural equation model in which the 5 client outcomes in a given utterance were regressed simultaneously on the therapist's behaviors in the preceding utterance, client-therapist rapport from the previous session, and the log-transformed session index, with SEs clustered on a unique session identifier constructed as the combination of client and session. Because every outcome is regressed on every predictor and all residual covariances are free, the model is saturated and has zero degrees of freedom; fit indices are consequently uninformative and are not reported. The SEM estimates are shown in Table 5 and Figure 1.

As a robustness check, we conducted parallel analyses using multilevel models. Utterances are nested within sessions and sessions within clients, and we therefore estimated random intercepts for both session and client. Results are reported in Table S7 in Multimedia Appendix 2 and are consistent with the SEM estimates. Because these models are not saturated, likelihood-based fit comparisons are informative, and we additionally report marginal and conditional R^2 for each outcome (Tables S8 and S9 in Multimedia Appendix 2).

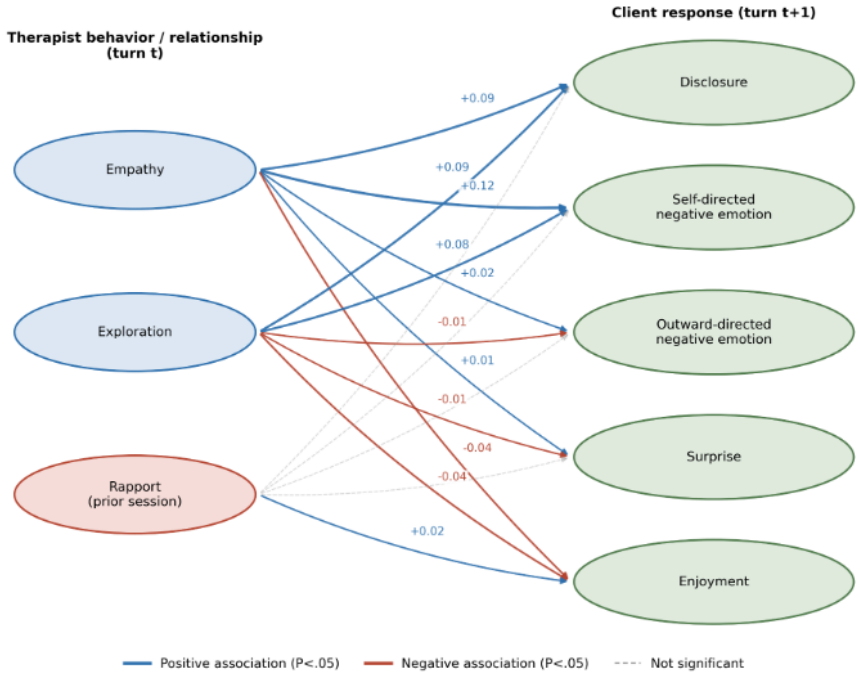
XSL•FO
RenderX

Table 5. Structural equation modeling results.[a]

Predictor	Client's disclosure	Client's emotion expression			
		Self-directed negative emotion (sadness to fear to anxiety to depression)	Outward-directed negative emotion (anger to contempt to disgust)	Surprise	Enjoyment
Log session ID					
Estimates (95% CI)	0.008 (–0.009 to 0.025)	0.002 (–0.020 to 0.024)	0.019 (–0.001 to 0.040)	0.011 (–0.001 to 0.023)	0.008 (–0.011 to 0.028)
P value	.35	.85	.06	.07	.40
Rapport (prior session)					
Estimates (95% CI)	–0.003 (–0.017 to 0.012)	–0.017 (–0.035 to 0.001)	–0.004 (–0.020 to 0.012)	0.002 (–0.008 to 0.012)	0.017 (0.001 to 0.033)
P value	.73	.07	.62	.66	.04
Exploration					
Estimates (95% CI)	0.086 (0.078 to 0.094)	0.078 (0.070 to 0.087)	–0.011 (–0.017 to –0.004)	–0.013 (–0.019 to –0.007)	–0.041 (–0.048 to –0.034)
P value	<.001	<.001	.002	<.001	<.001
Empathy					
Estimates (95% CI)	0.086 (0.078 to 0.095)	0.118 (0.110 to 0.127)	0.020 (0.013 to 0.028)	0.010 (0.004 to 0.017)	–0.037 (–0.043 to –0.030)
P value	<.001	<.001	<.001	.002	<.001

[a]n=119,233 client utterances from 1573 sessions and 37 clients, with SEs clustered on a unique session identifier constructed as the combination of client and session.

Figure 1. Path diagram of associations between therapist behavior, prior-session rapport, and client responses in the following turn. Estimates are standardized coefficients from the saturated structural equation modeling reported in Table 5, with SEs clustered at the session level (n=119,233). Line thickness is proportional to effect magnitude. Solid blue indicates a significant positive association and solid red a negative one at *P*<.05; dashed gray indicates a nonsignificant association. CIs are reported in Table 5. The log-transformed number of sessions the client already had was included as a covariate in the structural equation modeling (SEM) analysis reported in Table 5, but omitted in the path diagram. All associations are small in absolute terms (|β|≤.12).



XSL•FO
RenderX

Hypothesis 1: Therapist Behaviors and Client Self-Disclosure

Hypothesis 1 is supported: a client's self-disclosure is associated with the counselor's immediately preceding behavior, particularly their expressed empathy and exploratory probing. As shown in Table 5, therapist behaviors in the preceding turn were associated with client self-disclosure. Both empathy (β=.086, 95% CI 0.078-0.095; *P*<.001) and exploration (β=.086, 95% CI 0.078-0.094; *P*<.001) showed small positive associations. The 2 coefficients were nearly identical in magnitude, indicating that expressed empathy and exploratory questioning in the prior therapist turn were similarly associated with clients' likelihood of disclosing personal information in the subsequent turn. These findings support hypothesis 1.

Hypothesis 2: Therapist-Client Relationship (Rapport) and Client Self-Disclosure

Hypothesis 2 is not supported: a client's self-disclosure during a session is unrelated to cumulative rapport prior to the current session. *Rapport* from the previous session showed no meaningful association with disclosure in the current session (β=–.003, 95% CI –0.017 to 0.012; *P*=.73). The effect was essentially zero and statistically nonsignificant, providing no evidence that cumulative rapport increases the amount of information clients choose to reveal. Thus, hypothesis 2 is not supported.

Hypothesis 3: Therapist Behaviors and Client Negative Emotions

Hypothesis 3 is partially supported: higher levels of therapist empathy and exploration are associated with greater expression of self-directed negative emotions. Therapist empathy in the prior turn was positively associated with both types of negative emotion, with the association for self-directed negative emotion (β=.118, 95% CI 0.110-0.127; *P*<.001) approximately 6 times larger than for outward-directed negative emotion (β=.020, 95% CI 0.013-0.028; *P*<.001). Exploration showed the same pattern for self-directed negative emotion (β=.078, 95% CI 0.070-0.087; *P*<.001) but was associated with slightly lower outward-directed negative emotion (β=–.011, 95% CI –0.017 to –0.004; *P*=.002). Hypothesis 3 therefore holds for self-directed negative emotion but not uniformly across both emotion groups.

This asymmetry aligns with the functional distinction between the 2 emotion groups. Empathic responses may accompany a relational context in which clients feel safer acknowledging emotional distress, which would be consistent with the observed pattern of greater sadness, anxiety, and depression alongside less enjoyment. Outward-directed emotions such as anger, contempt, and disgust, by contrast, are more often reactive to external circumstances and may be relatively less dependent on empathic attunement for their expression.

Hypothesis 4: Therapist-Client Relationship (Rapport) and Client Negative Emotions

Hypothesis 4 is not supported: higher rapport between therapists and clients is not associated with greater expression of negative emotions. Contrary to hypothesis 4, greater rapport in the previous session was associated with lower levels of self-directed negative emotion in the current session, although this association was not significant (β=–.017, 95% CI –0.035 to 0.001; *P*=.07). The association with outward-directed emotion was also nonsignificant (β=–.004, 95% CI –0.020 to 0.012; *P*=.62).

We found no evidence that higher rapport is associated with greater expression of internally directed negative emotions (sadness, fear, anxiety, and depression) or outward-directed negative emotions (anger, contempt, and disgust). However, higher rapport is associated with greater enjoyment (β=.017, 95% CI 0.001 to 0.033; *P*=.04).

Discussion

Summary of Findings and Interpretations

This study examined how therapist behaviors and prior-session rapport relate to what clients say and feel during therapy. By analyzing each utterance and separating negative emotions into 2 groups, self-directed and outward-directed, we were able to see how different therapist actions are associated with different types of emotional expression and client disclosure.

We found that therapist empathy and exploration showed small positive associations with client disclosure (β=.086 for each). When therapists responded with understanding or asked clients to elaborate, clients became more willing to share personal information in the next turn. Empathy was also associated with greater negative emotional expression, although the association was approximately 6 times larger for self-directed negative emotions (sadness, fear, anxiety, and depression) than for outward-directed emotions (anger, contempt, and disgust). Overall, these findings are consistent with literature reporting that empathy is associated with greater emotional accessibility and tolerance, which may help clients approach and articulate difficult feelings [49,50].

Contrary to our expectations, rapport was not associated with greater self-disclosure or outward-directed negative emotions. A methodological gap may help clarify the mixed findings for rapport. Our hypotheses assumed that rapport represented the client's subjective experience of trust and safety, but in this study, the LLM measurement of rapport was based on observer ratings (Working Alliance Inventory-Observer version [WAI-O]). While prior research shows that both the outside observer and client version of the WAI predict positive treatment outcomes, such as improvements in depression, anxiety, and posttraumatic stress, the effect sizes are larger for alliance measured from a client's perspective [15,51]. However, the client and observer perspectives are not strongly related to each other. If the rapport scores reflect what observers see in the session, rather than how clients personally experience the relationship, then the weaker or unexpected effects become easier to understand. Future work should include measures based on client-rated rapport to better test our hypotheses.

Limitations

Small Effect Sizes

We acknowledge the small effect sizes of all the associations we report and that practical use of our empirical research would

XSL•FO
RenderX

require evidence of stronger association. The weak effect size problem may be methodological and may require better automated measures of the constructs such as empathy and self-disclosure. Or the low associations may be more theoretical, for instance, identifying the aggregation of turns and lags to show stronger associations. For example, our measure of rapport looks only at rapport as measured in the last session, but meta-analytic work [51] suggests measuring rapport over a longer span and measured later in therapy. Similarly, the relationship between empathy and self-disclosure may be stronger if empathy were measured over a longer period of time and not just at the immediately preceding turn.

Dataset Scope and Generalizability

Our analyses draw on a single corpus of professionally transcribed, one-on-one therapy sessions from the AS dataset. While this corpus provides high-quality, naturalistic clinical interactions, it reflects primarily Western therapeutic contexts. Linguistic, cultural, and religious backgrounds play an important role in shaping how clients express emotion, disclose personal information, and build rapport. Cultural norms influence preferred communication styles, expectations of therapist-client roles, and the degree to which emotional experiences are expressed directly or indirectly [52,53]. In many cultural contexts, emotional expression is modulated by sociocultural rules (eg, display norms), and self-disclosure is shaped by collective vs individualistic value systems and culturally grounded understandings of distress [54]. These differences can affect how therapeutic processes unfold and how constructs such as empathy, disclosure, or rapport manifest in session dialogue. In addition, this dataset is text-only, focusing on verbal content rather than multimodal cues. Because many therapeutic constructs are conveyed through language, the corpus supports meaningful measurement of therapist behaviors and client processes, though future work incorporating audio or multimodal data could further enrich assessments of emotional nuance and relational attunement. Therefore, the generalizability of our findings to other modalities, populations, or cultural settings remains an open question.

We were also unable to conduct subgroup analyses by clinical status or participant characteristics, which could have revealed meaningful heterogeneity. Three features of the data prevent this. The corpus is anonymized, so patient names, gender, and other identifying details are unavailable. The recorded presenting conditions are vague and are associated at the client level, leaving roughly one to 2 clients per condition, far too few to estimate a separate model per subgroup. Most fundamentally, all measures in our models are standardized within client, which removes between-client variance by construction, so the model as specified cannot address between-client questions of any kind. Datasets that retain clinical and demographic metadata would permit the analyses this limitation forecloses.

Contributions and Limitations of the LLM Prompts

Overview

Table S3 in Multimedia Appendix 2 shows that our prompts have the potential to generalize and are not an artifact of one vendor's model. If a specific model is updated or withdrawn, the prompts can be reapplied to an alternative model and yield comparable measurements. Researchers can also consult Table S3 in Multimedia Appendix 2 to choose a model appropriate to their budget and construct of interest. However, we want to acknowledge 3 limitations:

Portability of the Prompts

For example, the ICCs for the emotion measures ranged from 0.606 to 0.743. For these constructs, model choice materially affects measurement, and researchers substituting one model for another should expect different results.

Vendor-Driven Behavioral Drift

We recorded the exact model version strings and API access dates for all 6 models, pinned those versions in the released analysis code, and released all prompts. However, vendor-driven behavioral drift—the changes in behavior caused by shifts in the underlying models—is a limitation that we want to acknowledge [55].

Lack of Transparency

Although the scoring prompts/rubrics can be directly inspected and critiqued and are released with the paper, the model's reasoning remains opaque [56].

Ethical Considerations in LLM-Based Assessments for Therapeutic Applications

The use of LLMs to assess therapeutic interactions introduces important ethical considerations that must be carefully addressed to prevent harm and ensure responsible deployment. Although this study focuses on designing LLM-based measurement of generic therapist behavior types (eg, empathy) and high-level client outcomes (eg, self-disclosure and emotional expression), we also acknowledge that LLMs may inadvertently reproduce or amplify harmful behaviors, particularly those affecting marginalized or underrepresented groups. In the context of our study, these biases may manifest in the inappropriate or inaccurate evaluation of therapist behaviors or their relationships with client outcomes due to a limited understanding of the complex and culturally situated needs for different demographic groups. For example, models may systematically misinterpret communication styles common in particular cultures as avoidance, or misinterpret culturally normative emotional restraint as disengagement [57,58]. Similarly, expressions of emotional state and symptoms from marginalized groups may be flattened or pathologized, as has been the case across history when it comes to cultural differences for mental health treatment outcomes [59,60]. Limitations of the current model include insufficient cultural responsiveness in non-Western therapeutic dyads [61], a lack of nuanced understanding of clinical context (eg, the presence of underlying psychotic processes or approach-avoidance patient behaviors driven by insecure attachment), and a lack of external validation demonstrating clear value for patient outcomes.

Future Work

These considerations highlight several directions for future research. First, future studies should examine sessions conducted by therapists with varying levels of experience, as expertise may influence the variability of rapport. Second, future work

XSL•FO
RenderX

should test sequential mediation and moderation directly. Our data could not test moderation: the rapport main effect was not robust across specifications, and 37 clients is too few to estimate a cross-level interaction with useful precision. Future work should use multilevel dynamic SEM or related state-space/continuous-time approaches with distributed lags, random slopes, and cross-level interactions to capture heterogeneity across therapists and clients. Future work with datasets that include therapist identifiers could extend this work by examining how individual therapists differentially enact these general skills, but this lies beyond the scope of our present aims. Future analyses could also look for nonlinear patterns (such as thresholds or inverted-U shapes) and use mixture models to uncover different types of process patterns that might be hidden when averaging across all sessions.

Beyond these methodological directions, 2 ethical requirements should guide any future use of LLM-based automatic assessment. First, external validation before deployment. Our measures have been validated against trained lay annotators, not against clinical experts or patient outcomes. Demonstrating that these constructs predict clinically meaningful change is a prerequisite for any decision-support role. Until such validation exists, the appropriate use of this framework is research at scale, not clinical guidance.

Second, requirements for prospective deployment. Should systems of this kind eventually offer assessment or direct feedback to clinicians in training, that step would require safeguards our archival study does not need and therefore does not test: postdeployment monitoring for drift in the underlying commercial models, crisis-escalation protocols for sessions where risk indicators appear, rigorous data accountability for highly sensitive session content, and evaluation of differential performance across the populations on which the system would be used.

Conclusion and Broader Implications

This study provides a process-level account of how clinicians' in-session behaviors relate to clients' next-turn responses. Across outcomes, therapist empathy and exploration showed small but consistent associations with client disclosure and with the direction of emotional expression, while prior-session rapport showed no association that was robust across model specifications. Whether rapport conditions the association between therapist behavior and client response is a question our design cannot answer, and we identify it as a priority for future work. These findings provide a computational foundation for precision modeling of psychotherapy processes and may offer actionable insights for scalable therapist training and AI-supported clinical education systems.

Acknowledgments

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2026), the following tasks were delegated to GenAI tools under full human supervision: (1) reproducibility testing, (2) proofreading and editing, and (3) reformatting. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. Using large language models as measurement instruments is described in Methods and is separate from this disclosure.

Data Availability

The Alexander Street database [25] is available through academic, public, and K-12 institutional library subscriptions. The measurement prompts and all information necessary to reproduce the analysis are provided in the supplementary materials.

Funding

This work is funded by the National Institute of Mental Health (NIMH), "SCH: Training Mental Health Supporters with Virtual Patients and Automated Feedback" (R01 MH139114), and a fellowship provided by the Center for Machine Learning and Health at Carnegie Mellon University.

Authors' Contributions

Conceptualization: AHC, HZ, REK, HAS, TW
Methodology: AHC, SJ, CW, HZ
Formal analysis: AHC, HZ, REK
Data curation: AHC, SJ, CW
Funding acquisition: HZ, REK, HAS, TW
Supervision: HZ
Validation: SJ, CW
Writing–original draft: AHC, HZ
Writing–review and editing: HZ, REK, HAS, TW

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1

Construct prompts.
[DOCX File , 3002 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Specific context window configurations, prompt calibration datasets, intraclass correlation, estimates from multilevel models, variance decomposition, and fit statistics.
[DOCX File , 4620 KB-Multimedia Appendix 2]

References

1. Robledo Yamamoto F, Voida A, Voida S. From therapy to teletherapy: relocating mental health services online. Proc ACM Hum Comput Interact. 2021;5(CSCW2):1-30. [doi: 10.1145/3479508]
2. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry. 2019;64(7):456-464. [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
3. Berardi C, Antonini M, Jordan Z, Wechtler H, Paolucci F, Hinwood M. Barriers and facilitators to the implementation of digital technologies in mental health systems: a qualitative systematic review to inform a policy framework. BMC Health Serv Res. 2024;24(1):243. [FREE Full text] [doi: 10.1186/s12913-023-10536-1] [Medline: 38408938]
4. Kenny P, Parsons T, Gratch J, Leuski A, Rizzo A. Virtual patients for clinical therapist skills training. 2007. Presented at: 7th International Working Conference, IVA 2007; 2007 September 17-19:197-210; Paris, France. [doi: 10.1007/978-3-540-74997-4_19]
5. Eaton TT, Abeles N, Gutfreund MJ. Therapeutic alliance and outcome: impact of treatment length and pretreatment symptomatology. Psychotherapy (Chic). 1988;25(4):536-542. [doi: 10.1037/h0085379]
6. Opland C, Torrico TJ. Psychotherapy and Therapeutic Relationship. Treasure Island, Florida. StatPearls Publishing; 2024.
7. D'Alfonso S, Lederman R, Bucci S, Berry K. The digital therapeutic alliance and human-computer interaction. JMIR Ment Health. 2020;7(12):e21895. [FREE Full text] [doi: 10.2196/21895] [Medline: 33372897]
8. Hill CE, Clara JC. Psychotherapy Skills and Methods That Work. London. Oxford University Press; 2023.
9. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-613. [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
10. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
11. Fossey E, Harvey C, Mokhtari MR, Meadows GN. Self-rated assessment of needs for mental health care: a qualitative analysis. Community Ment Health J. 2012;48(4):407-419. [doi: 10.1007/s10597-011-9433-6] [Medline: 21691820]
12. Horvath AO, Greenberg LS. Development and validation of the working alliance inventory. J Couns Psychol. 1989;36(2):223-233. [doi: 10.1037/0022-0167.36.2.223]
13. Kim M, Hong JE, Ban M. Mediating effects of emotional self-disclosure on the relationship between depression and quality of life for women undergoing in-vitro fertilization. Int J Environ Res Public Health. 2021;18(12):6247. [FREE Full text] [doi: 10.3390/ijerph18126247] [Medline: 34207782]
14. Peluso PR, Freund RR. Therapist and client emotional expression and psychotherapy outcomes: a meta-analysis. Psychotherapy (Chic). 2018;55(4):461-472. [doi: 10.1037/pst0000165] [Medline: 30335458]
15. Horvath AO, Symonds BD. Relation between working alliance and outcome in psychotherapy: a meta-analysis. J Couns Psychol. 1991;38(2):139-149. [doi: 10.1037/0022-0167.38.2.139]
16. Yang W, Fang A, Shah RS, Mathur Y, Yang D, Zhu H, et al. What makes digital support effective? How therapeutic skills affect clinical well-being. Proc ACM Hum Comput Interact. 2024;8(CSCW1):1-29. [doi: 10.1145/3641029]
17. Sharma A, Miner A, Atkins D, Althoff T. A computational approach to understanding empathy expressed in text-based mental health support. 2020. Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 November 16-20:5263-5276; Online. [doi: 10.18653/v1/2020.emnlp-main.425]
18. Chikersal P, Belgrave D, Doherty G. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. 2020. Presented at: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020 April 25-30:1-16; Honolulu HI USA. [doi: 10.1145/3313831.3376341]
19. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. TACL. 2016;4:463-476. [doi: 10.1162/tacl_a_00111]
20. Leach MJ. Rapport: a key to treatment success. Complement Ther Clin Pract. 2005;11(4):262-265. [doi: 10.1016/j.ctcp.2005.05.005] [Medline: 16290897]
21. Elvins R, Green J. The conceptualization and measurement of therapeutic alliance: an empirical review. Clin Psychol Rev. 2008;28(7):1167-1187. [doi: 10.1016/j.cpr.2008.04.002] [Medline: 18538907]
22. Hill CE, Knox S, Pinto-Coelho KG. Self-disclosure and immediacy. In: Psychotherapy Relationships That Work: Volume 1: Evidence-Based Therapist Contributions. London. Oxford University Press; 2019:379-420.

XSL•FO
RenderX

23. Ekman P. Basic emotion. In: Handbook of Cognition and Emotion. New Jersey. John Wiley & Sons; 1999:45-60.
24. Vail K, Girard JM, Bylsma LM, Cohn JF, Fournier J, Swartz HA, et al. Toward causal understanding of therapist-client relationships: a study of language modality and social entrainment. Proc ACM Int Conf Multimodal Interact. 2022;2022:487-494. [FREE Full text] [doi: 10.1145/3536221.3556616] [Medline: 36913231]
25. Counseling and psychotherapy transcripts, client narratives, and reference works. Alexander Street. 2026. URL: https://search.alexanderstreet.com/psyc [accessed 2026-10-17]
26. Laska KM, Gurman AS, Wampold BE. Expanding the lens of evidence-based practice in psychotherapy: a common factors perspective. Psychotherapy (Chic). 2014;51(4):467-481. [doi: 10.1037/a0034332] [Medline: 24377408]
27. Wampold BE. How important are the common factors in psychotherapy? An update. World Psychiatry. 2015;14(3):270-277. [FREE Full text] [doi: 10.1002/wps.20238] [Medline: 26407772]
28. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change and Grow. New York. The Guilford Press; 2023.
29. Shah RS, Holt F, Hayati SA, Agarwal A, Wang Y, Kraut RE, et al. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. Proc ACM Hum Comput Interact. 2022;6(CSCW2):1-24. [doi: 10.1145/3555640]
30. Wu Z, Balloccu S, Kumar V. Anno-MI: a dataset of expert-annotated counselling dialogues. 2022. Presented at: ICASSP 2022-IEEE International Conference on Acoustics, Speech and Signal Processing; 2022 May 22-27:6177-6181; Singapore. [doi: 10.1109/ICASSP43922.2022.9746035]
31. Min DJ, Pérez-Rosas V, Resnicow K, Mihalcea R. PAIR: prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. 2022. Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 December 7-11:148-158; Abu Dhabi, United Arab Emirates. [doi: 10.18653/v1/2022.emnlp-main.11]
32. Tichenor V, Hill CE. A comparison of six measures of working alliance. Psychotherapy (Chic). 1989;26(2):195-199. [doi: 10.1037/h0085419]
33. Greenberg L. Emotion and cognition in psychotherapy: the transforming power of affect. Can Psychol. 2008;49(1):49-59. [doi: 10.1037/0708-5591.49.1.49]
34. Vondracek S, Vondracek F. The manipulation and measurement of self-disclosure in preadolescents. Merrill Palmer Q. 1971;17(1):51-58. [FREE Full text]
35. Bak J, Lin C, Oh A. Self-disclosure topic model for classifying and analyzing twitter conversations. 2014. Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 October 25-29:1986-1996; Doha, Qatar. [doi: 10.3115/v1/d14-1213]
36. Balani S, De CM. Detecting and characterizing mental health related self-disclosure in social media. 2015. Presented at: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems; 2015 April 18–23:1373-1378; South Korea. [doi: 10.1145/2702613.2732733]
37. Cheng X, Pan C, Zhao M. Revisiting chain-of-thought prompting: zero-shot can be stronger than few-shot. arXiv. Preprint posted online on June 17, 2025. [FREE Full text] [doi: 10.18653/v1/2025.findings-emnlp.729]
38. Hatanpää C, Smith N, Kumar S. On distributional robustness of in-context learning for text classification. OpenReview. URL: https://openreview.net/forum?id=fJgOCX5rDp [accessed 2026-10-23]
39. Parashar S, Olson B, Khurana S. Inference-time computations for LLM reasoning and planning: a benchmark and insights. arXiv. Preprint posted online on February 18, 2025. [FREE Full text]
40. Liu F, Chao W, Tan N, Liu H. Bag of tricks for inference-time computation of LLM reasoning. 2025. Presented at: Advances in Neural Information Processing Systems 38 (NeurIPS 2025); 2025, December 2-7; San Diego, California. URL: https://papers.neurips.cc/paper_files/paper/2025/hash/d3cb4d3573edb4404d39054c04e2b3c1-Abstract-Datasets_and_Benchmarks_Track.html
41. Kojima T, Gu S, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. 2022. Presented at: NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 November 28:22199-22213; New Orleans, LA. [doi: 10.52202/068431-1613]
42. Crawford JR, Henry JD. The Positive and Negative Affect Schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. Br J Clin Psychol. 2004;43(Pt 3):245-265. [doi: 10.1348/0144665031752934] [Medline: 15333231]
43. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284-290. [doi: 10.1037/1040-3590.6.4.284]
44. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163. [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
45. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1(1):30-46. [doi: 10.1037/1082-989X.1.1.30]
46. Jolliffe IT. Principal Component Analysis. Cham. Springer; 2002.
47. Rosseel YL. Lavaan: an R package for structural equation modeling. J Stat Soft. 2012;48(2):1-36. [doi: 10.18637/jss.v048.i02]
48. Kline RB. Principles and Practice of Structural Equation Modeling. New York City. Guilford Press; 2016.

XSL•FO
RenderX

49. Pascual-Leone A, Greenberg LS. Emotional processing in experiential therapy: why "the only way out is through.". J Consult Clin Psychol. 2007;75(6):875-887. [doi: 10.1037/0022-006X.75.6.875] [Medline: 18085905]
50. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: an updated meta-analysis. Psychotherapy (Chic). 2018;55(4):399-410. [FREE Full text] [doi: 10.1037/pst0000175] [Medline: 30335453]
51. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy (Chic). 2018;55(4):316-340. [FREE Full text] [doi: 10.1037/pst0000172] [Medline: 29792475]
52. Sue DW, Sue D. Counseling the Culturally Diverse: Theory and Practice. New Jersey. John Wiley & Sons; 2016.
53. Markus HR, Kitayama S. Culture and the self: implications for cognition, emotion, and motivation. Psychol Rev. 1991;98(2):224-253. [doi: 10.1037/0033-295X.98.2.224]
54. Kleinman A. Rethinking Psychiatry: From Cultural Category to Personal Experienc. Mumbai. Free Press; 1988.
55. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? Harv Data Sci Rev. 2024;6(2). [doi: 10.1162/99608f92.5317da47]
56. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e750. [FREE Full text] [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
57. Eng JS. Emotion regulation and culture: the effects of cultural models of self on western and east Asian differences in suppression and reappraisal [Dissertation]. ProQuest. Berkeley. University of California; 2012. URL: https://www.proquest.com/openview/2e8c91f89c3a07b04f03cd480549b7e1/1?pq-origsite=gscholar&cbl=18750 [accessed 2026-10-26]
58. Lim N. Cultural differences in emotion: differences in emotional arousal level between the East and the West. Integr Med Res. 2016;5(2):105-109. [FREE Full text] [doi: 10.1016/j.imr.2016.03.004] [Medline: 28462104]
59. Jimenez DE, Park M, Rosen D, Joo JH, Garza DM, Weinstein ER, et al. Centering culture in mental health: differences in diagnosis, treatment, and access to care among older people of color. Am J Geriatr Psychiatry. 2022;30(11):1234-1251. [FREE Full text] [doi: 10.1016/j.jagp.2022.07.001] [Medline: 35914985]
60. Pederson AB. Management of depression in Black people: effects of cultural issues. Psychiatr Ann. 2023;53(3):122-125. [FREE Full text] [doi: 10.3928/00485713-20230215-01] [Medline: 37396678]
61. Yang J, Liu T, Luo YT, Niu T, Pang P, Xiang A, et al. Exploring the application boundaries of LLMs in mental health: a systematic scoping review. Front Psychol. 2025;16:1715306. [FREE Full text] [doi: 10.3389/fpsyg.2025.1715306] [Medline: 41836276]

Abbreviations

AS: Alexander Street
EPITOME: Empathy in Text-Based, Asynchronous Mental Health Conversations
HIPAA: Health Insurance Portability and Accountability Act
ICC: intraclass correlation coefficient
LLM: large language model
PCA: principal component analysis
SEM: structural equation modeling
WAI: Working Alliance Inventory
WAI-O: Working Alliance Inventory-Observer version
WAI-O-S: Working Alliance Inventory-Observer Short version

Edited by J Torous; submitted 16.Mar.2026; peer-reviewed by K Satasiya, I Benzi, C-Y Chen; comments to author 05.Jun.2026; revised version received 13.Oct.2026; accepted 13.Oct.2026; published 05.Nov.2026

Please cite as:
Chen AH, Jin S, Wang C, Swartz HA, Wu T, Kraut RE, Zhu H
Modeling of Therapist-Client Dynamics in Psychotherapy Using Large Language Model–Based Assessments: Retrospective Observational Study of 1573 Psychotherapy Sessions
JMIR Ment Health 2026;13:e95389
URL: https://mental.jmir.org/2026/1/e95389
doi: 10.2196/95389
PMID:

©Angela H Chen, Siwei Jin, Canwen Wang, Holly A Swartz, Tongshuang Wu, Robert E Kraut, Haiyi Zhu. Originally published in JMIR Mental Health (https://mental.jmir.org), 05.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution,

XSL•FO
RenderX

and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX