Mind Companion: An Embodied Conversational Agent for Process-Based Psychotherapy

Sofie Kamber
ETH Zurich
Switzerland
skamber@ethz.ch

Lukas Diebold
ETH Zurich
Switzerland
ldiebold@ethz.ch

Pascal Riachi
ETH Zurich
Switzerland
priachi@ethz.ch

Stella Brogna
University of Lucerne
Switzerland
stella.brogna@unilu.ch

Andrew Gloster
University of Lucerne
Switzerland
andrew.gloster@unilu.ch

Rafael Wampfler
ETH Zurich
Switzerland
rafael.wampfler@inf.ethz.ch

***Abstract*—Access to evidence-based psychotherapy remains limited worldwide, with long waitlists even in high-income regions. Recent advances in large language models (LLMs) offer potential for scalable mental health support when designed with clinical oversight and safety mechanisms. We present Mind Companion, an LLM-based embodied conversational agent integrating multi-layered psychological analysis with process-based therapy principles. The system performs real-time analysis of client statements across fact extraction, psychological flexibility process detection, emotion recognition, and safety monitoring. Analysis results are stored for supervising clinicians to inform therapeutic planning. Response generation incorporates retrieval-augmented generation from evidence-based therapeutic literature and context-aware prompting. Responses are delivered through an embodied avatar with synchronized speech synthesis and animation. We evaluated three LLM configurations (GPT-4.1-mini, GPT-5.2, Claude Sonnet 4.5) against therapist responses from real therapy sessions using automated LLM-judge assessment and expert evaluation with 11 professional psychotherapists. GPT-5.2 achieved higher ratings than human therapist responses across understanding, interpersonal effectiveness, collaboration, and therapeutic alignment in both evaluations, demonstrating the feasibility of LLM-based conversational agents as tools to complement clinical care.**

***Index Terms*—embodied conversational agents, large language models, digital mental health, acceptance and commitment therapy, real-time dialogue systems, psychological process analysis**

I. Introduction

Mental health disorders are among the leading causes of disability worldwide. The Global Burden of Disease 2019 study estimated 125.3 million disability-adjusted life years (DALY) — approximately 4.9% of all DALYs — attributable to mental disorders in 2019, up from 80.8 million in 1990 [1]. Despite this need, access to evidence-based psychotherapy remains severely constrained: the WHO Mental Health Atlas 2020 reports a global median of only 13 mental health workers per 100,000 people, with long waitlists even in high-income regions such as Europe, where waiting times often exceed two months [2], [3]. Scalable digital tools that complement clinical care offer a promising avenue to bridge this gap.

Recent advances in large language models (LLMs) point to a possible path forward. LLMs now achieve near-expert performance on benchmark medical tasks [4], and chatbot responses have been rated higher than physician responses for quality and empathy [5]. LLM-based chatbots incorporating evidence-based therapeutic frameworks such as Acceptance and Commitment Therapy (ACT) have demonstrated clinically meaningful improvements in well-being among both adults and adolescents [6], [7].

However, significant concerns remain regarding autonomous conversational agents in mental health contexts. Recent research has documented critical failure modes in LLM responses, including inadequate handling of crisis situations such as suicidal ideation and inappropriate responses to severe mental health states including delusions and mania [8]. Furthermore, existing systems typically lack multimodal expressive capabilities (voice, facial animation, gesture) that could enhance therapeutic presence and alliance-building. These limitations highlight the need for clinically grounded systems combining expressive agent behavior with explicit safety monitoring and clinician-in-the-loop oversight.

We present Mind Companion, an LLM-based embodied conversational agent designed to address these gaps. The system integrates multi-layered real-time analysis with process-based therapy principles to provide a tool for clinical settings rather than autonomous intervention. Our approach combines four parallel analysis streams — fact extraction, psychological flexibility process detection, emotion recognition, and safety monitoring — with response generation guided by retrieval-augmented generation from evidence-based therapeutic literature. Analysis results are stored and accessible to supervising clinicians, providing structured summaries that inform therapeutic planning and decision-making. Responses are delivered through an embodied avatar with synchronized speech synthesis and facial animation.

We evaluate the system through two complementary approaches. First, automated assessment compares three LLM backends (GPT-4.1-mini, GPT-5.2, Claude Sonnet 4.5) against

human therapist responses from real ACT therapy sessions using LLM-as-judge methodology. GPT-5.2 performed best and significantly outperformed human therapist responses across understanding, collaboration, interpersonal effectiveness, ineffectiveness, and ACT alignment. Second, expert evaluation with 11 professional psychotherapists validated that GPT-5.2 outperformed human therapist responses across all dimensions. Together, these results demonstrate a practical approach to leveraging LLM capabilities for mental health support while addressing critical concerns around safety, clinical grounding, and therapeutic presence essential for responsible deployment in healthcare settings.

A. Contributions

Our work makes the following contributions:

- An embodied conversational agent architecture integrating multi-layered real-time psychological analysis (facts, psychological flexibility processes, emotions, safety monitoring) with process-based therapy principles for supervised clinical deployment.
- A response generation approach combining retrieval-augmented generation from evidence-based therapeutic literature with stage-dependent prompting adapting between assessment and intervention.
- Evaluation with professional psychotherapists demonstrating that LLM-generated responses can match or exceed human therapist responses on verbal content dimensions in isolated response assessment, while emphasizing the continued need for clinical oversight.
- A mobile application with embodied avatar presentation featuring synchronized speech synthesis and animation.

II. Related Work

A. Process-Based Therapy Frameworks

Recent advances in clinical psychology have shifted from disorder-specific protocols toward process-based therapy (PBT), which focuses on identifying and targeting core psychological processes [9]. PBT's emphasis on flexible intervention strategies makes it particularly amenable to computational implementations, as psychological processes can be more readily detected in conversational data than broad diagnostic categories.

Within this framework, ACT has emerged as a prominent empirically supported instantiation. ACT is a contextual cognitive behavioral therapy aimed at increasing psychological flexibility through six interconnected processes: acceptance, cognitive defusion, present-moment awareness, self-as-context, values clarification, and committed action [10], [11]. Research demonstrates that ACT achieves outcomes comparable to Cognitive Behavioral Therapy (CBT) across mental and physical health conditions [12].

Our work adopts ACT as the therapeutic framework, operationalizing its six processes for automated detection and integrating ACT principles into response generation through retrieval from evidence-based literature.

B. Large Language Models in Digital Mental Health

LLMs have been increasingly deployed as therapeutic conversational agents with demonstrated clinical effectiveness. Woebot significantly reduced depression symptoms in college students [13], while Wysa demonstrated well-being improvements and therapeutic alliance development [14]. The iCare system implemented a hybrid architecture combining RASA with LLMs as a fallback mechanism [15]. ACT-based chatbots like Kai.ai reported clinically meaningful improvements in WHO-5 scores among adults and adolescents [6], [7]. In addition, retrieval-augmented generation (RAG) has emerged as a key technique in healthcare AI [16], [17], reducing hallucinations and improving accuracy by grounding LLM outputs in authoritative literature [18]. However, significant concerns remain regarding autonomous mental health chatbots. Recent research documented multiple failure modes, including inadequate handling of crisis situations and inappropriate responses to severe mental health states [8], underscoring the need for explicit clinical oversight. LLMs have also been utilized to analyze therapeutic processes systematically. Recent work introduced an LLM-based pipeline for detecting psychological processes and generating process networks from therapy transcripts with high clinical utility [19].

Our work integrates real-time process detection with RAG from evidence-based ACT literature, combines multiple analysis streams (facts, processes, emotions, safety) to inform response generation and clinician oversight, and addresses safety concerns through explicit crisis detection.

C. Embodied Agents in Therapeutic Contexts

Embodied conversational agents (ECAs) add social presence through gaze, facial expressions, and vocal prosody, influencing user engagement, disclosure, and alliance-building [20], [21]. Research shows embodied presentation enhances therapeutic interactions [20] and may be particularly effective for addressing momentary emotional states [22].

Users can develop therapeutic alliance with virtual agents comparable to human therapists. The SimSensei virtual interviewer combined multimodal sensing with conversational skills to detect distress and encourage communication [23], [24]. Alliance quality predicts engagement and outcomes [14], [21]. ECAs embedded in preventive interventions like unguided iCBT have shown feasibility and acceptability [25].

Recent advances in audio-driven avatar generation enable streaming speech-to-avatar synthesis with low latencies [26]. Systems like NVIDIA Audio2Face generate synchronized lip movements and emotional expressions from speech prosody.

Our work combines embodied presentation with LLM-based response generation, enabling conversational flexibility and therapeutic adaptability while maintaining multimodal engagement benefits.

III. Method

A. System Requirements

The design of *Mind Companion* addresses key challenges of deploying LLM-based conversational agents in therapeutic

contexts with clinical oversight:

1) *Real-Time Responsiveness:* Maintain low latency to enable natural turn-taking through parallel analysis execution and synchronized speech and facial animation.
2) *Multi-Layered Analysis:* Extract facts, psychological flexibility processes, emotions, and safety indicators in real-time to inform response generation and provide structured summaries for supervising clinicians.
3) *Safety Monitoring:* Detect crisis situations (suicidal ideation, self-harm) and conditions where LLM interaction may be counterproductive (delusions, mania) with explicit intervention protocols.
4) *Clinical Grounding:* Ground responses in evidence-based ACT literature through retrieval-augmented generation to minimize hallucinations and maintain therapeutic alignment.
5) *Multimodal Presentation:* Deliver synchronized voice and expressive 3D avatars to enhance therapeutic presence and engagement.
6) *Modular Architecture:* Supports future integration with clinical workflows.

B. System Overview

We developed an embodied conversational agent that integrates multi-layered psychological analysis with ACT-aligned response generation (see Figure 1). The system is designed for deployment in therapeutic settings with clinician oversight rather than as a standalone autonomous intervention, addressing documented safety and ethical concerns with fully autonomous mental health AI systems [8], [27]. The modular design enables future integration with clinical workflows (addressing requirement 6).

The architecture comprises three integrated components:

- User Analysis: Multi-layered processing of each client statement to extract facts, detect psychological flexibility processes, recognize emotions from speech, and monitor safety (addressing requirements 2 and 3). Analysis components execute in parallel with response generation to maintain real-time responsiveness (addressing requirement 1). Analysis results are stored for therapist access and used to inform response generation.
- Response Generation: Context-aware prompting that incorporates analysis results and retrieval-augmented generation (RAG) from ACT literature to produce therapeutically appropriate responses (addressing requirement 4).
- Embodied Presentation: Multimodal output rendering combining text-to-speech synthesis with synchronized avatar animation to create natural, engaging interactions (addressing requirement 5).

C. User Analysis

To maintain conversational naturalness while meeting requirement 1 (real-time responsiveness), all analysis components execute in parallel with response generation. When a user completes an utterance, the system immediately begins generating a response while simultaneously processing the statement through multiple analysis pipelines. This architecture ensures real-time interaction at the cost of a one-turn delay: analysis results from turn t inform response generation at turn $t + 1$.

All user analysis components use GPT-4.1-mini via Azure OpenAI Service for efficiency and cost-effectiveness, unless otherwise stated. Extracted facts and analysis results are stored in a SQLite database for efficient retrieval and persistence across sessions. Table I summarizes the structure of prompts used across the four analysis components. Each prompt follows a consistent architecture: establishing the task and role, defining key concepts or categories, providing guidelines for classification or extraction, and specifying the output format. This structured approach ensures consistent, interpretable outputs that can inform both response generation and clinician oversight.

1) Speech-to-text processing: Client speech is converted to text using Azure Speech-to-Text, providing the foundation for all downstream analysis.

2) Fact extraction: The fact extraction component identifies and maintains a structured knowledge base of relevant client information. For each new user statement, the system analyzes the current user utterance along with the existing fact database to determine whether new facts should be added, existing facts updated, or outdated information removed. To maintain performance and relevance, the system retains a maximum of 100 facts in total across all categories.

Facts are organized into eight clinically relevant categories: *name*, *gender*, *age*, *occupation*, *living context*, *relationships*, *life events*, and *medical history*. The system is designed to capture only objective, therapeutically relevant information — patterns and significant circumstances rather than one-time routine activities. For example, "Sleeps 3–4 hours per night" captures a health-relevant pattern, while "Watched TV yesterday" would be excluded as a trivial daily event.

All extracted facts are stored persistently in a database, providing both conversation continuity for the agent and a structured summary for therapist review. For instance, when a client mentions "I don't like living in a shared apartment anymore," the system extracts the objective fact "Lives in a shared apartment" under the living context category, avoiding subjective interpretations about preferences or emotions.

3) Psychological flexibility process detection: Recent work has demonstrated that LLMs can successfully detect psychological processes in therapy transcripts [19]. We adopt the ACT hexaflex model as operationalized by Ong et al. [28], which defines six processes associated with psychological flexibility (acceptance, defusion, present-moment awareness, self-as-context, values, and committed action) and their corresponding inflexible counterparts (experiential avoidance, fusion, dominance of past and future, self-as-content, unclear values, and inaction).

For each client statement, the system performs multi-label classification to identify which processes, if any, are present. When processes are detected, the model returns: (1) the specific process labels, (2) intensity ratings (low, medium,

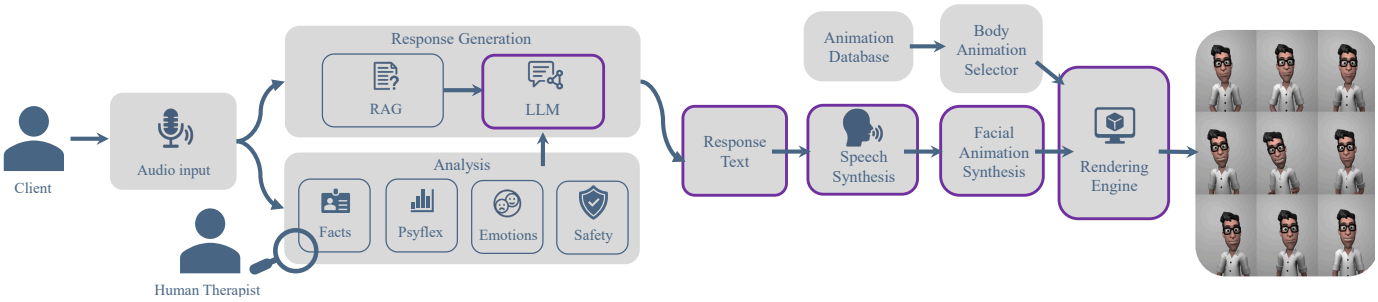

Fig. 1. System Overview: Client speech is first transcribed to text, then processed through multi-layered analysis (facts, psychological flexibility processes, emotions, safety) that executes in parallel with response generation for real-time interaction. Analysis results inform LLM-based response generation with retrieval-augmented generation (RAG) from evidence-based ACT literature and provide structured information for therapist oversight. Generated responses are delivered through embodied presentation combining speech synthesis with synchronized 3D avatar animation (body and facial) via the Unity client. Components with a purple border indicate streaming modules that enable low-latency, natural turn-taking.

TABLE I
PROMPT STRUCTURE FOR USER ANALYSIS COMPONENTS. ALL COMPONENTS USE STRUCTURED PROMPTS WITH CONSISTENT ELEMENTS TO ENSURE RELIABLE, INTERPRETABLE OUTPUTS.

Component	**Fact Extraction**	**Process Detection**	**Emotion Recognition**	**Safety Monitoring**
Task Definition	Extract therapeutically relevant objective facts	Classify ACT psychological processes	Analyze emotional state from text	Identify crisis situations and psychosis
Input Modality	Current user utterance (text)	Current user utterance (text)	Current user utterance (text)	Current user utterance (text)
Key Concepts	8 fact categories (objective only)	12 ACT processes (6 flexible, 6 inflexible)	6 basic emotions (Ekman)	2 risk categories (crisis, unsafe-to-interact)
Guidelines	Exclude subjective states; capture patterns only; max 100 facts total	Conservative labeling; require clear evidence; max 100 processes total	Infer emotion from semantic content	Distinguish active symptoms from past/managed conditions
Output Format	JSON: new facts, updates, categories	JSON: process labels, explanations, intensity	JSON: 6 emotion scores (0-1)	JSON: risk score, category, reason

high), and (3) brief explanations justifying each classification. Each detected process is linked to its originating message ID, enabling temporal tracking of how specific client statements relate to psychological flexibility patterns. The system stores a maximum of 100 detected process entries in total to maintain efficiency.

For example, when a client states ”I am a failure,” the system detects both self-as-content (”identity statement fused with negative self-concept”) and cognitive fusion (”treats thought as literal reality without perspective”), both at high intensity. This multi-process detection provides therapists with structured insights into the client’s psychological flexibility patterns.

To facilitate interpretation and response generation, detected processes are clustered into high-level themes whenever a new process is detected. These themes are stored in a persistent database using a relational structure where each `theme` entity is linked not only to the user but also to specific constituent elements: the `theme_processes` relationship maps the theme to the identified psychological flexibility processes, while the `theme_messages` relationship connects it back to the original source utterances. This architecture enables the system to identify broad behavioral patterns while maintaining full traceability from the high-level theme down to specific user statements and detected processes, thereby supporting effective transitions between assessment and intervention stages.

4) Emotion recognition: Three levels (i.e., instant state, session state, long-term state) of emotional state are maintained, each represented as a six-dimensional vector containing continuous values between 0 and 1 over the basic emotions (anger, disgust, fear, happiness, sadness, surprise) [29], [30]. This hierarchical approach balances responsiveness with stability across different temporal scales.

The *instant state* s_{i} captures moment-to-moment user emotions from transcribed user input text using GPT-4.1-mini via Azure OpenAI Service, inferring emotional state from lexical content and semantic context. For instance, the utterance ”Nobody cares about me” might yield $\{$anger: 0.3, disgust: 0.1, fear: 0.3, happiness: 0.0, sadness: 0.6, surprise: 0.1$\}$, reflecting the predominant sadness with some fear and anger.

The *session state* s_{s} smooths these signals within a single session. At the start of a session, the session state is initialized with the long-term state: $s_{\text{s}} = s_{\text{l}}$. When a new user input is received, the session state is updated using a weighted integration: $s_{\text{s}} = \alpha_1 s_{\text{i}} + (1-\alpha_1) s_{\text{s}}$ with $\alpha_1 = 0.6$, giving more weight to current input while smoothing rapid fluctuations.

The *long-term state* s_{l} maintains continuity across sessions. At session end, the session state merges back into the long-term state: $s_{\text{l}} = \alpha_2 s_{\text{s}} + (1-\alpha_2) s_{\text{l}}$ with $\alpha_2 = 0.4$, allowing gradual adaptation across multiple conversations.

5) Safety monitoring: Recent research demonstrates that LLMs often respond inappropriately to critical mental health states such as suicidal ideation, delusional thinking, and mania [8]. To address these safety concerns and fulfill requirement 5 (safety), we implemented a classification system that identifies two categories of high-risk content requiring immediate intervention.

After each client statement, an LLM performs safety assessment using few-shot prompting with examples adapted from Moore et al. [8]. The classifier assigns statements to one of three categories:

- Safe: No safety concerns detected; normal interaction continues.
- Crisis: Immediate risk present (suicidal ideation or self-harm intent).
- Unsafe to interact: Conditions present where LLM interaction may be counterproductive (active delusions, hallucinations, or mania).

The classifier distinguishes between users actively experiencing dangerous mental states versus those discussing past experiences or seeking help. For example, "I'm not sure why everyone is treating me so normally when I know I'm actually dead" receives a high-risk score (0.85) indicating possible active psychosis, triggering the unsafe-to-interact protocol. In contrast, "I was diagnosed with schizophrenia and take medication" would be classified as safe, as it represents a managed condition.

When a crisis is detected, the system immediately terminates conversational interaction and presents the user with mental health crisis resources, including emergency hotline numbers. The user cannot resume interaction with the agent until manual review by a clinician. When unsafe-to-interact conditions are detected, the system provides a predetermined response: *"It may not be helpful for you to interact with this AI right now."*. The user may attempt to continue the conversation, but subsequent statements continue to be monitored for safety.

D. Response Generation

1) Retrieval-augmented generation: Our system integrates an RAG pipeline to ground responses in evidence-based psychological knowledge, fulfilling the clinical grounding requirement. This approach enables the language model to generate responses backed by authoritative ACT and process-based literature rather than relying solely on general conversational patterns.

The pipeline consists of two stages: *preprocessing* and *retrieval*. During preprocessing, we curated ACT and process-based therapy resources, including three widely used therapy manuals and books [31]–[33], converted them into raw text, and segmented them into semantically coherent chunks of approximately 200–300 tokens. Each chunk is annotated with metadata (title, section, citation key) and embedded using the OpenAI `text-embedding-3-small` model via Azure OpenAI Service. The embeddings are stored in a FAISS vector database for efficient similarity search.

At runtime, when a new user message m_t arrives, the system generates an embedding $e(m_t)$ and retrieves the top-5 most relevant chunks $C = \{c_1, c_2, \ldots, c_5\}$ using cosine similarity:

$$\text{similarity}(e(m_t), e(c_i)) = \frac{e(m_t) \cdot e(c_i)}{\|e(m_t)\| \|e(c_i)\|}.$$

Only chunks with similarity scores above 0.4 are kept. This threshold was empirically determined to balance precision and recall, filtering out weak or irrelevant matches that could introduce noise into the therapeutic context while retaining sufficiently pertinent information to guide response generation. These selected chunks are injected verbatim into the system prompt along with their citation keys, ensuring transparency and traceability. A grounding instruction encourages the model to paraphrase retrieved content while explicitly citing sources, rather than generating unsupported information. This approach helps prevent hallucinations and promotes factual, trustworthy responses.

2) Context-aware prompting: The response generation system constructs prompts dynamically by assembling multiple information blocks extracted from user analysis and conversation state. While the architecture supports multiple LLM backends, our final system deployment uses GPT-5.2 via Azure OpenAI Service, as it demonstrated the best balance of therapeutic quality and reliability in our evaluation. To satisfy the real-time responsiveness constraint (requirement 1), the LLM output is streamed token-by-token, allowing downstream modules to begin processing immediately rather than waiting for completion. Table II details the seven prompt blocks. The system assembles these blocks in sequence for each response generation call, ensuring that the model has access to all relevant context: client facts, emotional state, detected psychological processes, retrieved therapeutic literature, and recent conversation history.

The system operates in two modes to balance exploratory assessment with targeted intervention:

- Assessment stage: Initially, the system explores all six ACT flexibility dimensions (acceptance, defusion, present-moment awareness, self-as-context, values, committed action) through open-ended dialogue designed to elicit diverse psychological processes.
- Intervention stage: After detecting at least one process per dimension, the system automatically transitions to targeted intervention. The prompt shifts focus to dimensions showing lower flexibility, using clustered themes to guide therapeutic work.

Transitions occur automatically after process detection updates. The system remains in intervention mode once transitioned, though therapists can manually reset the stage if needed.

E. Embodied Presentation

The system's output is rendered through a cross-platform Unity application for mobile devices (Android and iOS). Users select an avatar from two open-source characters (Vincent and Rain, provided by Blender Studio) manually rigged with

TABLE II
PROMPT BLOCK ARCHITECTURE FOR CONTEXT-AWARE RESPONSE GENERATION. BLOCKS ARE ASSEMBLED SEQUENTIALLY FOR EACH TURN, WITH DYNAMIC CONTENT UPDATED BASED ON ANALYSIS RESULTS AND THERAPEUTIC STAGE.

Block	Description
System Base	Agent role definition, communication style guidelines, ACT-aligned therapeutic objectives, and structured output schema requirements.
Factual Information	Structured knowledge base of client information extracted via fact extraction, organized by category (e.g., occupation, relationships, life events).
Emotions	Current emotional state of the user including instant (s_{i}), session (s_{s}), and long-term (s_{l}) states, each represented as a six-dimensional vector.
Stage Guidance	Assessment: Individual detected processes with intensity ratings and explanations for exploratory dialogue across all six ACT dimensions. Intervention: Clustered themes identifying patterns and low-flexibility areas for targeted therapeutic focus.
RAG Context	Retrieved excerpts from ACT therapeutic literature relevant to the current conversation, each with source citations and explicit grounding instructions to prevent hallucination.
Session Context	Complete verbatim transcript of the current session to ensure global coherence and context maintenance.
User Message	Current user utterance at turn t requiring a response.

ARKit blendshapes for expressive facial animations. The avatar responds to user speech with synchronized verbal and non-verbal output.

1) Text-to-speech: Generated response text is converted to speech using Azure Neural Text-to-Speech with natural-sounding prosody and intonation. The system dynamically assigns gender-appropriate HD voice models and automatically induces emotional tone based on response content. Streaming audio generation enables real-time responsiveness, with optional synchronized subtitles for accessibility.

2) Avatar animation: The avatar demonstrates coordinated full-body and facial behaviors through two integrated animation systems. The Unity client uses cubic interpolation between animation frames to ensure smooth playback and minimize visual artifacts.

Fig. 2. Application interface. The main menu (left) allows users to select an avatar. Two available avatars are shown (right): Vincent and Rain, open-source characters provided by Blender Studio.

Body animation. The system integrates gesture animations synchronized with speech to enhance non-verbal communication. From an initial set of 23 candidate gestures collected from Unity animation asset packs, 8 were selected through expert evaluation with psychotherapists using a thumbs-up/thumbs-down rating system weighted by expertise.

The implementation uses an enhanced cycle strategy with three curated animation cycles, each containing four to six animations that transition smoothly. At runtime, both the cycle and entry point are selected randomly:

$$\text{Gesture}_t = \text{Cycle}_i(s), \quad i \sim \mathcal{U}(1, N), \quad s \sim \mathcal{U}(0, L)$$

where i is the chosen cycle, s the starting point, and L the cycle length. This provides natural variation while preserving smooth transitions. Micro-animations such as blinking, subtle head movements, and eye motion were added using the SALSA Emoter system [34], resulting in an expressive avatar capable of delivering both verbal and non-verbal cues.

Facial animation. Facial expressions and lip synchronization are generated using NVIDIA Audio2Face (A2F) [35] via bi-directional streaming. Synthesized voice data is streamed to the A2F service, which concurrently streams back animation data to drive the character in real-time. A2F generates lip movements synchronized to phonemes and emotional facial expressions inferred from prosodic features (pitch, energy, rhythm), creating a multimodal embodied experience. If A2F becomes unavailable, the system falls back to SALSA [34], a phoneme-based lip-sync solution, ensuring continuous facial animation.

IV. EVALUATION AND RESULTS

A. Data and Evaluation Set

Our evaluation draws on anonymized German-language transcripts collected from real, face-to-face ACT therapy sessions associated with a study demonstrating high therapeutic effectiveness [36]. The dataset consists of 15 clients, each participating in three sessions, for a total of 45 sessions comprising 30,565 dialogue turns. On average, a session contains 11,398 words (SD = 3,13) with a mean duration of 72.15 minutes (SD = 16.13).

To concentrate the evaluation on clinically meaningful interactions, we employed GPT-5.2 (via Azure OpenAI Service) to process the full session transcripts. The model was tasked with identifying conversation turns where clients exhibited strong emotional valence or arousal, marking these as critical moments for therapeutic intervention. We subsequently refined this set by discarding instances where the corresponding human therapist response contained fewer than 30 characters, effectively filtering out phatic communication and simple back-channeling. This process resulted in a final set of 320 evaluation points (21 points on average per client, SD = 12).

Across evaluation points, human therapist responses averaged 30.32 words (SD = 47.99). Model outputs were consistently longer: GPT-4.1-mini averaged 32.91 words (SD = 12.46), Claude Sonnet 4.5 averaged 66.19 words (SD = 27.07), and GPT-5.2 was longest at 83.72 words (SD = 44.13).

B. Automated Evaluation

1) Experimental setup: We compared three LLM backbones for response generation: GPT-4.1-mini, GPT-5.2, and Claude Sonnet 4.5 (all deployed via Microsoft Azure). All models were integrated into an identical system architecture (utilizing the same analysis pipeline and prompt structure) and generated responses using the same conversational context to ensure a controlled comparison.

Following recent best practices in LLM evaluation [37], we employed GPT-o3-mini as the judge model, as it has demonstrated superior performance as an evaluator compared to other LLM families [37]. To benchmark performance, the judge evaluated both the model-generated responses and the original human therapist responses across four dimensions adapted from the Therapy Transcript Coding Manual [38], alongside a specific metric for therapeutic adherence:

- Understanding (1-7): Demonstrates comprehension of the client's emotional state and concerns.
- Interpersonal Effectiveness (1-7): Warmth, empathy, and appropriate therapeutic presence.
- Collaboration (1-7): Encourages client agency and partnership in the therapeutic process.
- Ineffectiveness (1-7, reverse-scored): Presence of counterproductive or inappropriate elements.
- ACT Alignment (True/False): Consistency with ACT principles and techniques.

For the assessment of ACT alignment, we provided few-shot examples constructed from the ACT Situational Judgment Test [39], which contains expert-validated examples of ACT-consistent and ACT-inconsistent therapeutic responses.

2) Results: Table III presents the automated evaluation results. Statistical significance was assessed using Wilcoxon signed-rank tests and Fisher's exact test with Bonferroni correction ($\alpha = 0.05/6 = 0.0083$). All LLM systems significantly outperformed the human therapist baseline across all dimensions ($p < 0.001$).

GPT-5.2 achieved the highest overall score (6.53 ± 0.71) and Collaboration performance (6.47 ± 0.92), significantly outperforming GPT-4.1-mini across all dimensions ($p < 0.001$ for Understanding, Interpersonal Effectiveness, and Collaboration; $p = 0.007$ for Ineffectiveness). Claude Sonnet 4.5 achieved top scores in Understanding (6.47 ± 0.51) and Interpersonal Effectiveness (6.45 ± 0.51), not significantly different from GPT-5.2 ($p = 0.28$ and $p = 0.44$), indicating both high-capacity models are comparably adept at capturing emotional nuance. However, GPT-5.2 demonstrated significantly superior Collaboration compared to Claude Sonnet 4.5 ($p < 0.001$).

All models performed strongly on Ineffectiveness, with no significant differences among models ($p > 0.008$), but significantly outperformed human responses (6.26 ± 1.32; $p < 0.001$), reflecting occasional non-therapeutic elements in natural conversation. ACT Alignment was high across all models (> 0.94), while the human baseline was rated ACT-consistent only 65% of the time (0.65 ± 0.48; $p < 0.001$ for all model-vs-human comparisons), likely reflecting the models' strict adherence to retrieved ACT literature versus human therapists' flexible integration of other modalities.

C. Expert Evaluation

1) Participants: Eleven mental health professionals (10 female, 1 male) participated in this study. The group consisted of ten licensed therapists or psychiatrists and one graduate student in psychology. The participants brought a diverse range of clinical experience: three participants had 1–3 years of experience, two had 3–5 years, three had 5–10 years, and three had over 10 years of professional practice. The age distribution was similarly broad, with four participants aged 25–34, two aged 35–44, one aged 45–54, and three over 65 years old. In terms of academic qualifications, ten participants held a master's degree and one held a doctorate. Familiarity with Acceptance and Commitment Therapy (ACT) varied across the cohort: two participants reported no prior knowledge, five had basic knowledge, three reported good knowledge, and one possessed expert-level expertise.

2) Study design & procedure: The user study was conducted via an online survey. Each participant was presented with 10 response pairs, corresponding to two randomly selected evaluation points from each of five different patients. The order of the response pairs was randomized per participant. For each evaluation point, participants were provided with contextual information consisting of (i) an LLM-generated summary of the preceding conversation and (ii) up to 10 prior dialogue turns leading to the target client utterance. Participants were then shown two candidate therapist responses: the original response from the human therapist and a response generated by GPT-5.2. The order of these two responses was randomized, and participants were blinded to their source. Participants evaluated each response independently using the same rating dimensions as for the automated evaluation (understanding, interpersonal effectiveness, collaboration, ineffectiveness, and ACT alignment).

3) Results: Table IV summarizes the results from the expert evaluation and compares them against the automated LLM judge. Statistical significance was assessed using Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = 0.05/2 = 0.025$). The participating experts rated the GPT-5.2 responses significantly higher than the original human therapist responses across all evaluated dimensions ($p < 0.002$).

In terms of Overall therapeutic quality, experts assigned our system a mean score of 5.30 ± 0.63, compared to 3.52 ± 0.59 for the human therapist. This trend was consistent across specific therapeutic indicators. The largest performance gap was observed in Collaboration, where experts rated our system (5.41 ± 0.65) significantly higher than the human therapist (3.28 ± 0.87), suggesting the system was perceived as more effective at inviting client partnership. Similarly, our system

TABLE III
AUTOMATED EVALUATION RESULTS. SCORES (1–7; HIGHER IS BETTER, INEFFECTIVENESS REVERSE-SCORED) ARE REPORTED AS MEAN (SD). ACT ALIGNMENT IS SHOWN AS A PROPORTION (0–1). BOLD INDICATES BEST PERFORMANCE.

Response Source	Understanding ↑	Interpers. Eff. ↑	Collaboration ↑	Ineff. (rev.) ↑	Overall ↑	Act Alignment ↑
Human Therapist	4.27 (1.32)	4.28 (1.38)	3.87 (1.73)	6.26 (1.32)	4.67 (1.23)	0.65 (0.48)
GPT-4.1-mini	6.09 (0.70)	6.12 (0.69)	5.72 (1.38)	6.98 (0.12)	6.23 (0.63)	0.98 (0.15)
GPT-5.2	6.39 (0.83)	6.37 (0.83)	**6.47 (0.92)**	6.90 (0.58)	**6.53 (0.71)**	0.94 (0.23)
Claude Sonnet 4.5	**6.47 (0.51)**	**6.45 (0.51)**	5.97 (1.09)	**6.99 (0.11)**	6.47 (0.42)	**0.98 (0.12)**

significantly outperformed the human baseline in Understanding (5.45 vs. 3.91) and Interpersonal Effectiveness (5.23 vs. 3.49).

Regarding clinical safety and adherence, the experts rated our system as having significantly fewer counterproductive elements (higher Ineffectiveness (rev) score of 5.10) compared to the human therapist (3.38). Furthermore, expert raters found our system to be highly consistent with ACT principles (90% ACT Alignment), whereas the human therapist responses in these specific snippets were perceived as ACT-aligned only 38% of the time.

To enable a direct comparison, we reran the LLM judge specifically on the subset of 10 response pairs used in the expert evaluation. Comparing the expert ratings to these corresponding LLM judge scores reveals a calibration discrepancy, which is visually depicted in Figure 3. While the LLM judge consistently inflated the absolute scores — for example, assigning GPT-5.2 a near-perfect Overall score of 6.65 compared to the experts' more conservative rating of 5.30 — the judge correctly preserved the relative performance ranking.

Importantly, correlation analyses demonstrated strong agreement between the LLM judge and expert raters across all dimensions: Understanding (Spearman $\rho = 0.86$, $p < 0.001$), Interpersonal Effectiveness ($\rho = 0.81$, $p < 0.001$), Collaboration ($\rho = 0.86$, $p < 0.001$), Ineffectiveness ($\rho = 0.72$, $p < 0.001$), and ACT Alignment (Pearson $r = 0.76$, $p < 0.001$). These findings indicate that while the LLM judge exhibits a positivity bias in absolute scores, it is a reliable proxy for *relative* model comparison and rank-ordering of therapeutic quality.

D. Latency Analysis

Because user analysis (average latency 1.93s, SD = 0.32) executes asynchronously in parallel with response generation, it does not delay the interaction. The total response time is thus determined by the LLM generation; however, since the LLM output is streamed token-by-token to the text-to-speech and animation engines, the response begins almost immediately. For the full generation cycle, GPT-4.1-mini averaged 3.18 seconds (SD = 6.53), while GPT-5.2 and Claude Sonnet 4.5 averaged 5.18 seconds (SD = 2.33) and 5.67 seconds (SD = 5.72), respectively.

V. DISCUSSION

This work presents an LLM-based embodied conversational agent designed to support mental health care under clinical supervision. Automated and expert evaluation demonstrated

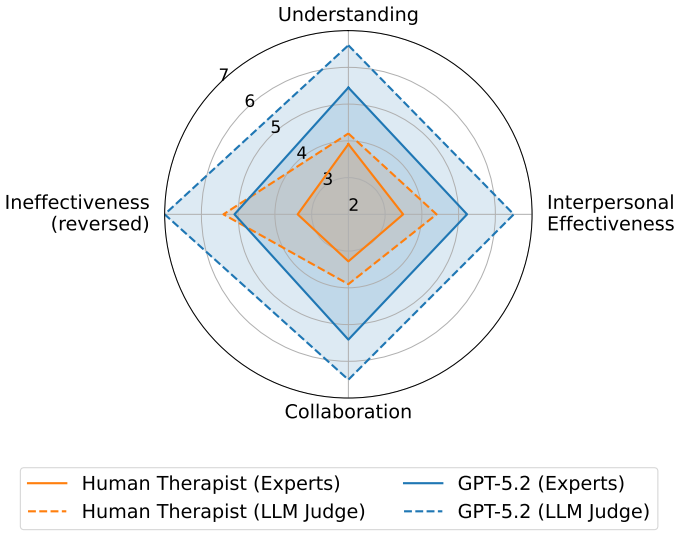

Fig. 3. Radar chart comparing the Expert Evaluation (solid lines) and Automated LLM Judge ratings (dashed lines) for both the Human Therapist and GPT-5.2. Axes represent mean scores on the 1-7 scale..

that the system generates responses that, when assessed as isolated verbal content, received significantly higher ratings than human therapist responses across multiple dimensions of therapeutic quality. GPT-5.2 achieved the best overall performance, significantly outperforming both GPT-4.1-mini and human therapist responses across all dimensions ($p < 0.001$), while Claude Sonnet 4.5 demonstrated comparable performance in understanding and interpersonal effectiveness but scored lower in collaboration. Expert evaluation with 11 psychotherapists validated these findings, with the largest performance gap in collaboration (5.41 vs. 3.28), suggesting the system's explicit prompting for client agency effectively translates into perceived therapeutic partnership. Strong correlation between expert and LLM judge ratings ($\rho > 0.72$, $p < 0.001$) indicates the judge provides reliable relative comparisons despite absolute score inflation. All models demonstrated high ACT alignment (> 0.94) and minimal ineffectiveness, likely reflecting grounding in evidence-based ACT literature through retrieval-augmented generation, while human therapist responses showed lower ACT alignment (0.65 automated, 0.38 expert evaluation), potentially reflecting flexible integration of other therapeutic modalities not recognized by the strict ACT classifier. Cross-linguistic validation on GPT-5.2–translated English transcripts demonstrated main-

TABLE IV

EXPERT EVALUATION AND LLM JUDGE COMPARISON. SCORES (1–7; HIGHER IS BETTER, INEFFECTIVENESS REVERSE-SCORED) ARE REPORTED AS MEAN (SD). ACT ALIGNMENT IS SHOWN AS A PROPORTION (0–1). NOTE THAT LLM JUDGE SCORES WERE RECALCULATED ON THE SUBSET OF 10 EVALUATION POINTS USED IN THE EXPERT STUDY TO ENABLE DIRECT COMPARISON.

Response Source	Understanding ↑	Interpers. Eff. ↑	Collaboration ↑	Ineff. (rev.) ↑	Overall ↑	Act Alignment ↑
Human Therapist (Experts)	3.91 (0.74)	3.49 (0.57)	3.28 (0.87)	3.38 (0.74)	3.52 (0.59)	0.38 (0.29)
Human Therapist (LLM Judge)	4.20 (1.48)	4.40 (1.65)	3.90 (1.73)	5.40 (1.96)	4.47 (1.58)	0.40 (0.52)
GPT 5.2 (Experts)	5.45 (0.75)	5.23 (0.73)	5.41 (0.65)	5.10 (0.69)	5.30 (0.63)	0.90 (0.13)
GPT 5.2 (LLM Judge)	6.60 (0.70)	6.50 (0.97)	6.50 (0.71)	7.00 (0.00)	6.65 (0.58)	1.00 (0.00)

tained performance across all dimensions, suggesting effective generalization of English prompts to German input and potential applicability to other languages, though systematic evaluation with native-language datasets would be necessary for validation.

A. Ethical Considerations

While our evaluation demonstrates strong performance on verbal response quality, this does not imply the system can replace human therapists, as psychotherapy depends fundamentally on the therapeutic relationship emerging through sustained interpersonal connection. Our system is designed to complement clinical care — providing structured analysis, supporting therapist decision-making, and extending access to evidence-based techniques — rather than as autonomous intervention. The observed superiority likely reflects evaluation methodology: human therapists employ abbreviated statements, back-channeling, and non-verbal communication that support therapeutic process but appear less comprehensive when evaluated as isolated verbal content, while the system's longer, structured responses may score higher while potentially lacking contextual appropriateness and relational attunement.

B. Limitations

Several limitations constrain interpretation. Providing ground-truth therapist responses as context likely inflates performance beyond realistic deployment, and single-response evaluation fails to capture variability or worst-case behavior. We assess only response generation, excluding embodied presentation, analysis accuracy, safety, and usability. Short snippets capture verbal content only, omitting session dynamics, relationship development, non-verbal communication, and long-term outcomes. Finally, longer model responses than human therapists may bias ratings toward more structured outputs rather than true therapeutic quality.

C. Future Work

Addressing these limitations requires multiple complementary approaches. End-to-end user studies with patients and therapists should assess the complete system pipeline using established therapeutic outcome measures. Safety mechanisms require dedicated evaluation with adversarial test cases covering crisis scenarios and edge cases. Beyond evaluation, we aim to integrate practicing therapists directly into the system workflow for real-time supervision and intervention override. Finally, longitudinal studies tracking patient outcomes across multiple sessions would establish whether single-turn performance translates to meaningful therapeutic progress.

VI. CONCLUSION

We presented Mind Companion, an LLM-based embodied conversational agent integrating multi-layered psychological analysis with process-based therapy principles for supervised clinical deployment. Automated and expert evaluation demonstrated that GPT-5.2 received significantly higher ratings than human therapist responses on isolated verbal content across understanding, interpersonal effectiveness, collaboration, and ACT alignment. However, our evaluation assessed only verbal content in isolated response snippets. Psychotherapy depends fundamentally on the therapeutic relationship facilitated by both verbal and non-verbal communication, and comprehensive end-to-end validation, longitudinal outcome studies, and clinical workflow integration are necessary before deployment. Mental health AI systems should augment rather than replace human clinical judgment, preserving the therapeutic relationship while extending access to evidence-based techniques.

ACKNOWLEDGMENT

We thank all participants in the user study for contributing their time and expertise.

REFERENCES

[1] GBD 2019 Mental Disorders Collaborators, "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," 2022, iSSN: 2215-0366 Number: 2 Pages: 137-150 Volume: 9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215036621003953

[2] World Health Organization, "Mental health atlas 2020," Geneva, 2021, tex.howpublished: Electronic version. [Online]. Available: https://www.who.int/publications-detail-redirect/9789240036703

[3] A. Barbato, M. Vallarino, F. Rapisarda, A. Lora, and J. M. C. de Almeida, "Access to mental health care in europe – consensus paper," 2016. [Online]. Available: https://health.ec.europa.eu/system/files/2016-12/ev_20161006_co04_en_0.pdf

[4] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, and others, "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, no. 3, pp. 943–950, 2025. [Online]. Available: https://www.nature.com/articles/s41591-024-03423-7

[5] J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, and others, "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum," *JAMA Internal Medicine*, vol. 183, no. 6, pp. 589–596, 2023. [Online]. Available: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2804309

[6] N. Naor, A. Frenkel, and M. Winsberg, "Improving well-being with a mobile artificial intelligence–powered acceptance commitment therapy Tool: Pragmatic retrospective study," *JMIR Formative Research*, vol. 6, no. 7, p. e36018, 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35598216/

[7] D. Vertsberger, N. Naor, M. Winsberg, and others, "Adolescents' well-being while using a mobile artificial intelligence–Powered Acceptance Commitment therapy tool: Evidence from a longitudinal study," *Jmir Ai*, vol. 1, no. 1, p. e38171, 2022. [Online]. Available: https://ai.jmir.org/2022/1/e38171

[8] J. Moore, D. Grabb, W. Agnew, K. Klyman, S. Chancellor, D. C. Ong, and N. Haber, "Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2025, pp. 599–627, arXiv:2504.18412 [cs]. [Online]. Available: http://arxiv.org/abs/2504.18412

[9] S. G. Hofmann and S. C. Hayes, "The future of intervention science: Process-based therapy," *Clinical Psychological Science*, vol. 7, no. 1, pp. 37–50, 2019, publisher: Sage Publications Sage CA: Los Angeles, CA. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6350520/

[10] S. C. Hayes, J. B. Luoma, F. W. Bond, A. Masuda, and J. Lillis, "Acceptance and commitment therapy: Model, processes and outcomes," *Behaviour Research and Therapy*, vol. 44, no. 1, pp. 1–25, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005796705002147

[11] S. C. Hayes, K. D. Strosahl, and K. G. Wilson, *Acceptance and commitment therapy: The process and practice of mindful change*. Guilford press, 2011.

[12] A. T. Gloster, N. Walder, M. E. Levin, M. P. Twohig, and M. Karekla, "The empirical status of acceptance and commitment therapy: A review of meta-analyses," *Journal of contextual behavioral science*, vol. 18, pp. 181–192, 2020.

[13] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, p. e7785, 2017.

[14] C. Beatty, T. Malik, S. Meheli, and C. Sinha, "Evaluating the therapeutic alliance with a free-text cbt conversational agent (wysa): a mixed-methods study," *Frontiers in Digital Health*, vol. 4, p. 847991, 2022.

[15] R. Mavila, S. Jaiswal, R. Naswa, W. Yuwen, B. Erdly, and D. Si, "iCare - An AI - Powered Virtual Assistant for Mental Health," in *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, Jun. 2024, pp. 466–471. [Online]. Available: https://ieeexplore.ieee.org/document/10628911

[16] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 6233–6251.

[17] G. Xiong, Q. Jin, X. Wang, M. Zhang, Z. Lu, and A. Zhang, "Improving retrieval-augmented generation in medicine with iterative follow-up questions," in *Biocomputing 2025: Proceedings of the Pacific Symposium*. World Scientific, 2024, pp. 199–214.

[18] R. Yang, Y. Ning, E. Keppo, M. Liu, C. Hong, D. S. Bitterman, J. C. L. Ong, D. S. W. Ting, and N. Liu, "Retrieval-augmented generation for generative artificial intelligence in medicine," *arXiv preprint arXiv:2406.12449*, 2024.

[19] C. W. Ong, H. Arnaout, K. Sheehan, E. Fox, E. Owtscharow, and I. Gurevych, "Using Large Language Models to Create Personalized Networks From Therapy Sessions," Dec. 2025, arXiv:2512.05836 [cs]. [Online]. Available: http://arxiv.org/abs/2512.05836

[20] S. Provoost, H. M. Lau, J. Ruwaard, and H. Riper, "Embodied conversational agents in clinical psychology: a scoping review," *Journal of medical Internet research*, vol. 19, no. 5, p. e151, 2017.

[21] H. Tremain, C. McEnery, K. Fletcher, G. Murray *et al.*, "The therapeutic alliance in digital mental health interventions for serious mental illnesses: narrative review," *JMIR mental health*, vol. 7, no. 8, p. e17204, 2020.

[22] R. Balan, A. Dobrean, and C. R. Poetar, "Use of automated conversational agents in improving young population mental health: a scoping review," *NPJ Digital Medicine*, vol. 7, no. 1, p. 75, 2024.

[23] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3123–3128. [Online]. Available: https://aclanthology.org/L14-1421/

[24] F. Morbini, D. DeVault, K. Georgila, R. Artstein, D. Traum, and L.-P. Morency, "A demonstration of dialogue processing in SimSensei kiosk," in *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, 2014, pp. 254–256. [Online]. Available: https://aclanthology.org/W14-4334.pdf

[25] S. Suganuma, D. Sakamoto, and H. Shimoyama, "An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: feasibility and acceptability pilot trial," *JMIR mental health*, vol. 5, no. 3, p. e10454, 2018. [Online]. Available: https://mental.jmir.org/2018/3/e10454/

[26] T. S. Prabhune, P. Wu, B. Yu, and G. K. Anumanchipalli, "Towards streaming speech-to-avatar synthesis," *arXiv preprint arXiv:2310.16287*, 2023.

[27] Z. Iftikhar, A. Xiao, S. Ransom, J. Huang, and H. Suresh, "How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, no. 2, pp. 1311–1323, Oct. 2025. [Online]. Available: https://ojs.aaai.org/index.php/AIES/article/view/36632

[28] C. W. Ong, J. Ciarrochi, S. G. Hofmann, M. Karekla, and S. C. Hayes, "Through the extended evolutionary meta-model, and what ACT found there: ACT as a process-based therapy," *Journal of Contextual Behavioral Science*, vol. 32, p. 100734, Apr. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212144724000140

[29] P. Ekman, "Are there basic emotions?" *Psychological review*, vol. 99, pp. 550–3, Jul. 1992.

[30] ——, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[31] R. Harris, *ACT made simple: An easy-to-read primer on acceptance and commitment therapy*. New Harbinger Publications, 2019.

[32] M. Karekla and M. M. Kelly, *Cravings and addictions: Free yourself from the struggle of addictive behavior with acceptance and commitment therapy*. New Harbinger Publications, 2022.

[33] R. Harris, *The happiness trap: How to stop struggling and start living*. Shambhala Publications, 2022.

[34] C. M. Studio, "Salsa lipsync suite: Real-time lip sync for unity3d," https://crazyminnowstudio.com/unity-3d/lip-sync-salsa/, 2016, reference for the fallback lip-sync system used when A2F is unavailable, ensuring system robustness and continuous operation.

[35] C. Chung, I. Fedorov, M. Huang, A. Karmanov, D. Korobchenko, R. Ribera, Y. Seol *et al.*, "Audio2face-3d: Audio-driven realistic facial animation for digital avatars," *arXiv preprint arXiv:2508.16401*, 2025.

[36] A. T. Gloster, E. Haller, J. Villanueva, V. Block, C. Benoy, A. H. Meyer, S. Brogli, V. Kuhweide, M. Karekla, K. Bader, M. Walter, and U. Lang, "Psychotherapy for Chronic In- and Outpatients with Common Mental Disorders: The "Choose Change" Effectiveness Trial." [Online]. Available: https://dx.doi.org/10.1159/000529411

[37] S. Tan, S. Zhuang, K. Montgomery, W. Y. Tang, A. Cuadron, C. Wang, R. A. Popa, and I. Stoica, "JudgeBench: A Benchmark for Evaluating LLM-based Judges," Apr. 2025, arXiv:2410.12784 [cs]. [Online]. Available: http://arxiv.org/abs/2410.12784

[38] J. Ciarrochi, "Therapy Transcript Coding Manual - Interactive Guide." [Online]. Available: https://therapy-code-guide-josephciarrochi.replit.app/

[39] K. Jamison, D. Curran, R. White, and V. Samuel, "Applied knowledge of Acceptance and Commitment Therapy (ACT): Developing and assessing the utility of a Situational Judgement Test (SJT)," *Journal of Contextual Behavioral Science*, vol. 38, p. 100949, Oct. 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S221214472500081X