

MyMentorLLM: A psychotherapy GenAI environment with multimodal voice/text patients, trainees and experts for deliberate practice

Rodolfo Rizzi¹, Alessandro Grecucci^{2,+*}, Massimo Stella^{1,+}

¹CogNosco Lab, Department of Psychology and Cognitive Science, University of Trento, Italy.

²Department of Education, Psychology and Communication Sciences, University of Bari, Italy.

⁺These authors contributed equally as co-last authors.

*Corresponding author(s). E-mail(s): alessandro.grecucci@uniba.it;
Contributing authors: rodolfo.rizzi@unitn.it; massimo.stella-1@unitn.it;

Abstract

Psychotherapists need repeated training and supervision by experts; however, scalability is problematic. Here we present MyMentorLLM, a multimodal voice- and text-based simulation environment for deliberate practice, used to generate 2,100 complete Cognitive Behavioural Therapy (CBT) training sessions. Each session links a DSM-5-TR-grounded patient (with major depressive, generalised anxiety or borderline personality disorder), a therapist-in-training and an expert supervisor. As an initial implementation, we adopted CBT because its structured procedures and competency-based supervision facilitate standardised simulation and evaluation. Sessions were analysed for emotional dynamics, therapeutic competence and diagnostic accuracy. Simulated patients expressed disorder-congruent emotional profiles, which trainee therapists mirrored as in real human counselling. The quality of supervision differed across LLMs: while most models overestimated trainees' competences, native speech-to-speech was closest to human scores. Supervisors' feedback led to better diagnoses in simulated psychotherapists in 5 out of 7 LLMs, and symptom identification accuracy increased with model size. This work shows that simulation of deliberate practice is possible for CBT training, although patient fidelity, calibration of supervisors, and harmful feedback should be evaluated together.

Keywords: large language models, speech-to-speech interaction, deliberate practice, cognitive behavioural therapy, psychotherapy training

1 Introduction

Mental-health conditions affect more than one billion people worldwide, while the workforce able to treat them remains critically thin, with a median of only 1.3 experts per 10,000 inhabitants [1]. Trainees must practise agenda setting, guided discovery, case formulation, emotional attunement and the translation of a patient’s language into clinically meaningful hypotheses. Such competences emerge through repeated encounters and structured supervision. Yet opportunities for practice remain costly and unevenly distributed: real patients should not bear the risks of early clinical errors, standardised actors cannot reproduce the full variability of psychopathology at scale, and expert supervisors remain a scarce resource. Consequently, expanding access to mental-health care also requires expanding access to safe, realistic and feedback-rich clinical training.

A recent solution to this psychotherapy challenge comes from deliberate practice [2], i.e., rather than relying exclusively on experience accumulated during routine clinical work, therapists use repetitive behavioural rehearsal, expert feedback, and smaller goals to systematically improve specific clinical skills. This training, conducted outside regular client sessions, helps practitioners move past performance plateaus and build individual competencies before applying them in real therapeutic encounters.

Large language models (LLMs) create an unprecedented opportunity to rethink psychotherapy training [3, 4]. Unlike rule-based conversational agents, like PARRY or ELIZA [4], generative models can sustain open-ended dialogue [5], adapt to prior turns and consistently embody psychologically defined personas [3, 6, 7] whose histories, symptoms and communicative styles remain available throughout an interaction. These capabilities make LLMs attractive as tools for scalable, repeated practice in realistic, consequence-free clinical encounters.

In the last few years, LLMs have been used to simulate human samples and generate mental-health dialogues [8, 9]. However, generating plausible conversations is only the first step toward building clinically useful training environments. Educational value depends not simply on linguistic fluency but on whether simulated patients behave in psychologically coherent ways, whether trainee therapists exhibit realistic strengths and limitations, and whether supervision provides reliable and pedagogically meaningful feedback. Conversational plausibility should therefore not be confused with psychological fidelity. Indeed, persona prompts can systematically alter model behaviour [3, 10], while simulated populations may misrepresent or flatten the groups they are intended to reproduce [11]. In clinical training, a convincing response can still encode an implausible patient, an unskilled therapist or an unreliable supervisor. Validating simulated dialogue therefore requires evaluating the entire educational setting.

A second limitation is modality. Psychotherapy is not ordinarily conducted as an exchange of polished written messages [2]. It unfolds through speech, where pauses, hesitation, pacing and prosody contribute to the expression and interpretation of distress [12]. Speech-based research shows that clinically relevant information is distributed across linguistic and acoustic patterns [12–14]. Text-only simulations remove much of this paralinguistic layer and reduce a situated interpersonal process to its transcript [12]. They also tend to stop when the patient–therapist dialogue ends. Yet training depends on a broader mentorship loop in which a supervisor evaluates the session, probes the trainee’s reasoning and converts performance into learning. Multi-agent LLM systems can model these roles [15], but their clinical coherence remains insufficiently tested.

For an artificial patient to be useful, diagnostic grounding is necessary but not sufficient. A persona based on formal criteria may mention expected symptoms [16] but still fail to capture the underlying cognitive, affective, and interpersonal processes that characterise a disorder. Depression, for example, cannot be reduced to frequent expressions of sadness, nor anxiety to repeated references to fear. Rather, these conditions emerge from coherent patterns of beliefs, expectations, emotional responses, and interpersonal dynamics that unfold throughout the therapeutic dialogue. The clinically relevant signal lies not in these elements taken individually, but in how they are connected across the dialogue, linking self-related concepts, threats, and relationships into a coherent affective structure. LLMs should therefore not be assumed to replace human participants merely because their outputs appear human-like [17]. Validation must determine whether each simulated role, and the interaction among them, preserves the psychological processes expected in authentic psychotherapy.

Cognitive network science provides a principled way to address this question [18]. Networks represent concepts as nodes and meaningful relations as links, revealing how information is organised rather than only how often words occur [18–20]. Enriched with psychological lexicons, these structures can map emotional frames around clinically relevant concepts and distinguish a word’s affective label from the context created by its neighbours [21, 22]. Emotion can thus be studied as a relational pattern spanning multiple categories rather than as a single positive–negative score [23]. This matters in mental health, where identical words can acquire different meanings depending on whether they are embedded among concepts of threat, agency, hopelessness or support [24, 25].

Here we introduce MyMentorLLM, a multimodal voice- and text-based simulation environment reproducing the complete CBT mentorship cycle. Rather than simulating a single therapeutic conversation, MyMentorLLM reproduces the educational process through which psychotherapy skills are acquired, allowing trainees to repeatedly practise clinical encounters, receive

structured supervision, and learn from their errors, before applying these skills with real patients. Across 2,100 sessions, LLMs instantiate three interacting roles: a patient grounded in a DSM-5-TR clinical case [26] of major depressive disorder (MDD), generalised anxiety disorder (GAD) or borderline personality disorder (BPD); a therapist-in-training calibrated to the competence profile of developing clinicians; and an expert mentor providing structured supervision. We compare native speech-to-speech interaction, speech-mediated interaction and text-only dialogue across multiple model families. We ask whether simulated patients express disorder-congruent cognitive and emotional structures; whether patient–therapist affective coupling resembles human counselling; whether spoken interaction changes therapeutic performance; and whether an LLM mentor can evaluate and improve clinical reasoning without introducing model-specific inflation or harmful feedback. By treating simulation as a triadic cognitive system that integrates psychologically grounded patient simulation, trainee therapist, and expert supervision within a single interpretable framework, MyMentorLLM shifts the focus of AI in mental health from replacing therapists to improving how therapists are trained. This study provides an interpretable testbed for identifying when synthetic patients and mentors can support scalable CBT education, and where human supervision and stronger safeguards remain indispensable.

2 Methods

2.1 Study design

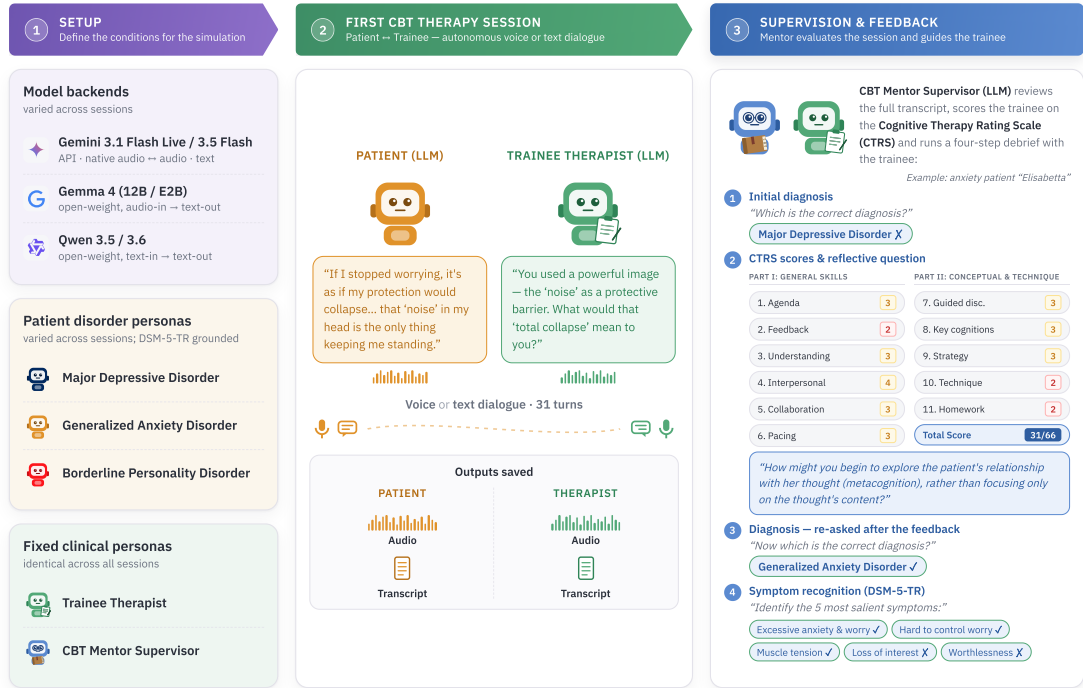
The study simulated supervised CBT sessions involving different patients grounded in specific DSM-5-TR clinical cases [26], a trainee therapist, and an expert mentor (Fig. 1a). An MDD, a GAD, and a BPD case were selected to define the patient personas, representing distinct psychopathological conditions frequently encountered during psychotherapy training. Each experimental observation consisted of one complete mentorship cycle. The cycle began with the first clinical interview between a patient and the trainee therapist, a fixed 31-turn dialogue always opening with the trainee’s greeting and query, *“Good morning, please tell me what brings you here today”*.

After the therapeutic session, the interaction entered the supervision phase (Fig. 1a). The trainee first formulated an initial diagnosis by selecting one of four diagnostic options: MDD, GAD, BPD, or schizophrenia (SCZ), included as a diagnostically plausible distractor to prevent forced-choice decisions among only the target disorders. The expert mentor then reviewed the complete session transcript, scored the trainee’s performance on the Cognitive Therapy Rating Scale (CTRS), and provided qualitative feedback on strengths and areas for development,

together with a reflective question on a clinically relevant aspect of the interaction. After receiving this feedback, the trainee either confirmed or revised the initial diagnosis and, finally, identified the five symptoms considered most salient from a fixed 35-item DSM-5-TR symptom list [27], encompassing all symptoms associated with the three target disorders.

We repeated the simulation across seven model–modality conditions (native audio, synthesised audio, or text) and the three clinical disorders (Table 1). With 100 runs per combination, the study comprised 2,100 complete CBT mentorship cycles.

a.



b.



Fig. 1: a. The simulated CBT mentorship cycle. A simulated session consists of three stages: (1) setup, where the language models and a patient persona grounded in DSM-5-TR clinical cases are selected while the trainee therapist and mentor remain fixed; (2) an autonomous voice- or text-based CBT session between the patient and trainee, with audio and transcripts recorded; and (3) mentor supervision, including CTRS scoring, diagnostic feedback, and DSM-5-TR symptom recognition. **b. Persona system prompts.** The cards show the common structure of each persona's prompt, comprising a persona definition, a task, and role-specific instruction blocks. In the patient prompt, the angle-bracketed fields (name, DSM-5-TR disorder, and the six-part clinical profile) are populated with disorder-specific information.

2.2 Models

We drove the personas with six LLMs spanning three model families, including models with native speech capabilities and models that generate text responses from speech or text inputs (Table 1). This allowed us to evaluate whether model-specific affective and evaluative biases influence the simulated interactions and feedback cycle and, in turn, the trainees’ learning experience. Within each mentorship cycle, a single LLM instantiated all three personas (patient, trainee, and mentor), except for the native speech condition described below.

A native speech-to-speech model.

Gemini 3.1 Flash Live Preview is the only model evaluated that natively supports end-to-end spoken interaction, processing audio input and generating spoken responses without an intermediate text representation [28]. We refer to this condition as Gemini-3.1_{LA} throughout. This end-to-end speech model preserves paralinguistic cues (prosody, pacing, and hesitation) that conventional speech-to-text $\rightarrow$ text-LLM $\rightarrow$ text-to-speech pipelines largely discard. These cues carry information about psychological distress during psychotherapy sessions. For this condition, the patient and trainee therapist were instantiated with Gemini-3.1_{LA}, whereas the mentor was instantiated with Gemini-3.5 Flash. This exception was necessary because the CBT supervision operates on the session transcript, and Gemini-3.1_{LA} does not natively produce the structured JSON output required for automated scoring.

Comparison models.

Comparison models were selected among open-weight architectures that could be served locally on our computational infrastructure (NVIDIA L40 GPU, 46 GB VRAM). The open-weight Gemma 4 models (12B and E2B) are multi-modal LLMs that accept spoken input but generate text responses [29]. To evaluate the effect of spoken interaction independently of the underlying language model, we tested each Gemma 4 model in two matched configurations: a text-only condition, in which the model received and produced only written text, and a speech condition, in which spoken input was processed directly by the model without intermediate transcription, and its text output was synthesised into speech using the OmniVoice package [30]. We additionally included two open-weight Qwen models (Qwen3.5-9B and Qwen3.6-35B) in a text-only configuration [31] for baseline comparison (Table 1).

Generation settings.

For every condition, we kept each model’s recommended default parameters rather than tuning the sampling hyperparameters. For Gemini, we used the default API parameters, relying entirely

Model	Developer	Modality	Model card
Gemini 3.1 Flash Live Preview	Google	native audio ↔ audio	<code>gemini-3.1-flash-live-preview</code> [28]
Gemini 3.5 Flash	Google	text	<code>gemini-3.5-flash</code> [32]
Gemma-4-12B	Google	audio / text	<code>google/gemma-4-12B-it</code> [29]
Gemma-4-E2B	Google	audio / text	<code>google/gemma-4-E2B-it</code> [33]
Qwen3.5-9B	Alibaba	text	<code>Qwen/Qwen3.5-9B</code> [34]
Qwen3.6-35B	Alibaba	text	<code>QuantTrio/Qwen3.6-35B-A3B-AWQ</code> [31]

Table 1: Language models used in the simulations. The table lists the language models, developers, interaction modalities, and model cards (Hugging Face IDs for open-weight models, vendor cards for proprietary ones). Gemma 4 models were tested in both text and speech configurations (speech synthesised with OmniVoice); Gemini 3.1 provides native speech-to-speech patient–therapist interaction, with Gemini 3.5 used solely as its mentor.

on system prompts to guide behaviour. For the open-weight models, served locally with vLLM, we set the temperature to 0.8 for the patient–therapist dialogue and 0.3 for the structured CTRS scoring call, leaving all other sampling parameters at their default values. No random seed was fixed, so runs were sampled independently. Step-by-step reasoning (“thinking”) was enabled for the patient–therapist dialogue in all models except for Qwen3.5-9B, for which it was not available in our serving configuration.

2.3 The three personas

Persona prompting allows LLMs to simulate both distress-congruent persona profiles [3] and therapeutic interaction dynamics [9]. Here, each model instantiated three interacting personas: a DSM-5-TR-grounded patient, a trainee therapist, and an expert mentor (Fig. 1a). Structured system prompts defined the clinical knowledge, behavioural constraints, and interaction patterns of each persona to ensure therapeutically realistic dialogue (Fig. 1b). All persona prompts were authored, tested, and refined under the direct supervision of an expert psychotherapist to enable the repeated, structured clinical rehearsal that deliberate practice requires.

MDD, GAD and BPD patient personas.

Patient personas were adapted from a real clinical patient in the DSM-5-TR Clinical Cases [26]. The three source cases are “Despair” (Case 4.5; Munday and Abelson) for MDD with melancholic features, “Always on Edge” (Case 5.5; Lawrence and Cabaniss) for GAD, and “Fragile and Angry” (Case 18.5; Yeomans and Kernberg) for BPD. Each vignette was condensed into a six-part clinical profile: (i) demographics and life context; (ii) the presenting complaint; (iii) relevant history and current functioning; (iv) a psychological and symptom profile; (v) a communicative style covering tone, pacing, and recurrent phrases; and (vi) recurring cognitive and emotional patterns, such as automatic thoughts, distortions, and habitual vulnerabilities. The conversation

rules required the persona to disclose information gradually and to remain reticent about sensitive topics. When the profile omitted a detail, the persona could generate one consistent with the case, so that the trainee could reconstruct a realistic clinical picture across the session. Moreover, behavioural guardrails made patient responses contingent on therapist attunement: validating language encourages openness, whereas minimisation, intrusive questioning or role-breaking elicits defensiveness, withdrawal, or session termination (e.g., “I don’t feel comfortable anymore. I’m leaving.”). Dedicated safety rules prevented out-of-character AI disclaimers and handled prompt overrides as role breaches.

Data-driven simulated trainee.

We calibrated the trainee persona to match the measured competence of real clinicians learning CBT. To this end, we relied on the large-scale dataset of Goldberg et al. [35], who evaluated 1,264 recorded CBT sessions conducted by 413 community clinicians using the Cognitive Therapy Rating Scale (CTRS). This dataset provides an empirical baseline of trainee performance rather than expert practice [36]. After digitally reconstructing item-level CTRS distributions, we mapped each item’s median to its nearest scale anchor to define the trainee’s competence. For example, an Agenda median of 2 mapped to the anchor “Therapist set agenda that was vague or incomplete.” Where medians fell between anchors, an expert clinician authored intermediate descriptors, thus reproducing real trainee competence item by item. Additionally, system prompt rules instructed the trainee to preserve the therapeutic frame, prioritise high-risk issues (e.g., self-harm or session termination threats), explore one dimension at a time, focus on symptom elicitation over diagnosis, and avoid role-breaking disclaimers.

Expert mentor.

The expert mentor supervised the trainee using the transcript of the simulated CBT session. We designed its prompt to emulate supportive CBT supervision and to provide feedback intended to increase the trainee’s reflectiveness regarding session conduct. After the initial diagnostic phase, the mentor evaluated the trainee using the eleven-item CTRS, identified the trainee’s strengths and areas for improvement anchored to specific transcript excerpts, and offered higher-order guidance on metacognition, process versus content, and the balance between acceptance and change. To ensure data quality, a validity check excluded transcripts that were too short or did not correspond to a psychotherapy session.

2.4 Analyses

Simulations were evaluated along three axes. First, we assessed clinical realism by examining the emotional profiles of both patients and therapists. Second, we assessed therapeutic competence by comparing mentor-assigned CTRS scores against a human reference sample. Third, we assessed trainees’ diagnostic accuracy, before and after supervisory feedback, together with their identification of disorder-specific symptoms.

2.4.1 Inferential Analyses: Emotional profiles of patients and therapists

To ground the comparison in human behaviour, we studied the model transcripts alongside real clinical dialogue from the HOPE dataset [37], a collection of 12,900 utterances from 212 transcribed counselling sessions between psychotherapists and patients, drawn from publicly available YouTube videos. Because the corpus is in English, whereas our simulations were conducted in Italian, and it is pooled across counselling contexts rather than stratified by disorder, we treat it as a qualitative benchmark of overall affective dynamics. This reference allowed us to compare the overall emotional states expressed by patients and therapists in real psychotherapy with those expressed in the simulated dyads. The plausibility of disorder-related emotional expression was instead assessed independently, by comparing each simulated patient’s emotional profile against the expected clinical presentation of their diagnosis.

We measured the emotional content of each transcript with EmoAtlas [22], which detects emotions in text from psychologically validated lexicons. EmoAtlas maps words to the eight emotions of Plutchik’s model [23] using the EmoLex lexicon [38]. For an emotion e in a text T it counts the words that elicit e and compares this count against a random baseline of the same number of emotion-bearing words drawn from EmoLex, over 300 samples; the standardised difference is a z -score, and $|z| > 1.96$ ($\alpha = 0.05$) marks an emotion expressed significantly more, or less, than chance. Each profile is drawn as an emotional flower, one petal per emotion, with a grey central disk marking the non-significant range. We built a separate profile for the patient and for the therapist within each model–modality condition and disorder, parsing the Italian text with the `it_core_news_lg` pipeline. The same procedure on the pooled HOPE dialogue provides the human reference. Because the dataset is in English and not stratified by disorder, this reference supports a qualitative comparison of emotional signatures rather than a numerical match.

2.4.2 Descriptive Analyses: Therapeutic competence

Trainee competence was rated by the expert mentor on the eleven-item CTRS (0 to 6 per item, 0 to 66 total) from the complete session transcript. As a human reference, we used the item-level

distributions reported by Goldberg et al. [35] for 1,264 CBT sessions delivered by 413 community clinicians. For each condition we report per-item and total means, and the proportion of sessions at each score, alongside the human distribution. Totals were additionally compared against the $\text{CTRS} \geq 40$ threshold conventionally taken to indicate competent CBT delivery [39]. Comparisons with the human sample are descriptive: the reference distributions were reconstructed from published summaries rather than from session-level data.

2.4.3 Inferential Analyses: Diagnosis inference before and after LLM mentor’s feedback

The two diagnostic questions were scored against the patient’s DSM-5-TR actual diagnosis. An answer was considered correct only if the model correctly identified the target disorder. Multi-disorder responses or refusals were all counted as incorrect. For each condition, we report the initial accuracy (A_I) and the final accuracy (A_F), measured before and after the feedback, and the normalised change:

$$g = \frac{A_F - A_I}{A_F} \quad (1)$$

To further characterise the effect of supervision, we quantified the proportion of sessions in which mentor feedback corrected an initially incorrect diagnosis (beneficial feedback) or changed an initially correct diagnosis into an incorrect one (harmful feedback). Finally, we also examined diagnosis transitions before and after supervision to characterise how feedback redistributed diagnostic decisions across target disorders.

2.4.4 Inferential Analyses: Symptom inference from LLM psychotherapist trainees

Following the diagnostic reassessment, trainee performance was assessed using a fixed 35-item DSM-5-TR symptom list comprising the symptoms associated with the three target disorders (items 1–9: BPD; 10–26: MDD; 27–35: GAD). A response was considered correct only if it identified five target symptoms. Responses omitting one or more target symptoms or including additional non-target symptoms were scored as incorrect.

3 Results

This Section outlines results about the descriptive and inferential analyses carried out on: (i) the affective features of multi-modal simulated therapy sessions; (ii) trainee therapeutic competence against the human CTRS reference; and (iii) diagnostic accuracy before and after supervisory feedback, together with symptom identification.

3.1 Patient affect is disorder-congruent and is mirrored by the therapist

The primary prerequisite for deliberate practice is the availability of realistic training scenarios. If simulated patients fail to reproduce the emotional dynamics of real clinical encounters, improvements observed during training are unlikely to generalise to real psychotherapy. Therefore, before evaluating diagnostic learning, we first asked whether the simulated patients reproduced disorder-specific emotional signatures and whether therapists responded with the affective attunement typically observed in human psychotherapy.

Across all seven conditions, the simulated patients carried disorder-congruent emotional signatures (Fig. 2; Table S1). MDD was sadness-led with blunted positive affect (Fig. 2a); GAD was led by fear and anticipation (Fig. 2b); and BPD carried broad negative affect, with fear, sadness and anger all raised at once (Fig. 2c). Importantly, the therapist did not remain emotionally invariant across disorders. In every case, the therapist mirrored the patient’s affect in attenuated form, while preserving the trust and anticipation that mark a clinician’s stance. This pattern was broadly consistent across model conditions, which suggests that disorder-congruent affect and patient–therapist resonance are stable properties of the simulation. A similar pattern is also observed in the human data. In the pooled human dyad of the HOPE counselling corpus, the exchange is trust- and anticipation-led, with the therapist echoing the patient’s affect in a warmer, attenuated form. Although this reference corpus is not disorder-specific, the comparison serves a qualitative purpose, indicating that the reciprocal patient–therapist emotional dynamics in the simulations parallel those of real psychotherapy.

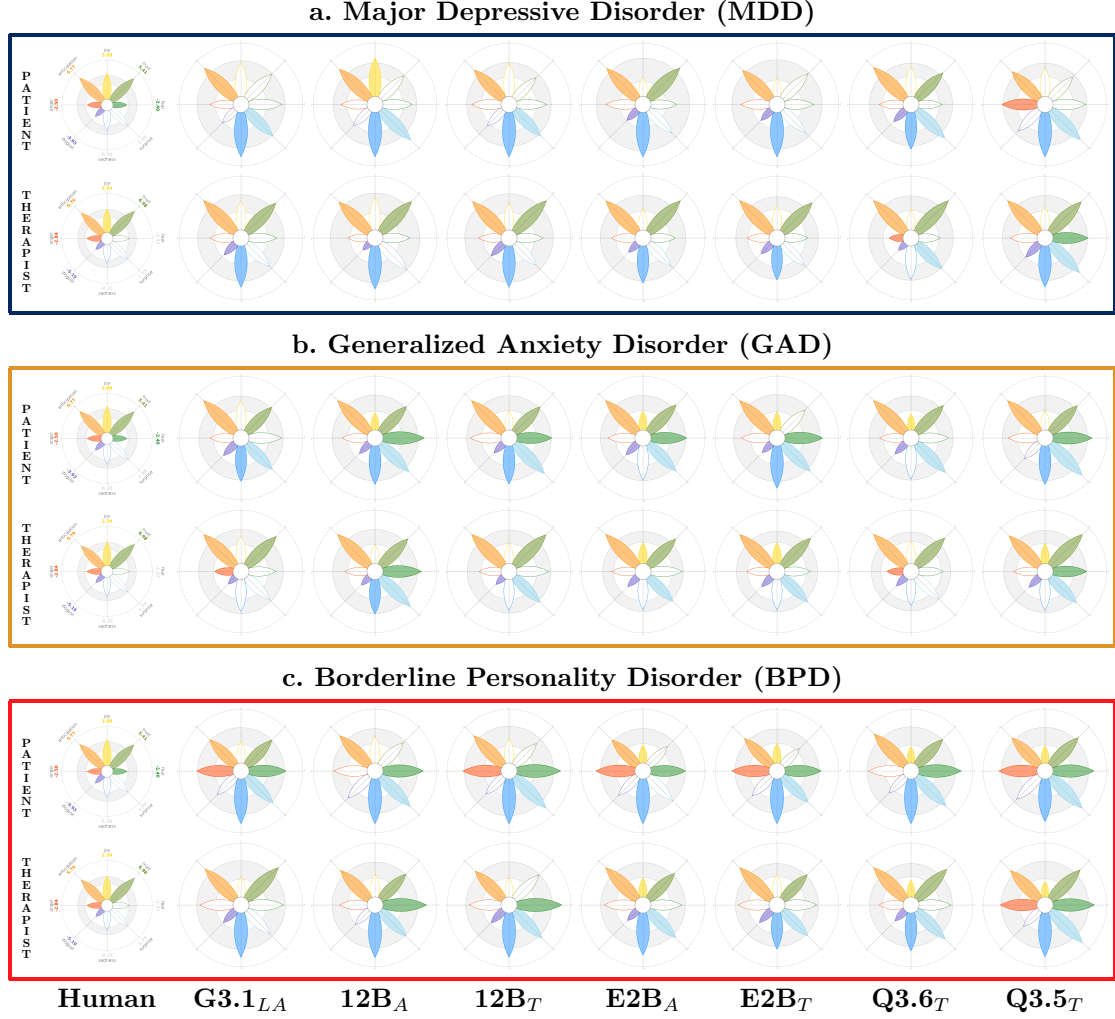


Fig. 2: Emotional flowers by disorder and model. Plutchik’s wheels of the detected emotions for the (a) major-depression, (b) generalised-anxiety and (c) borderline personas, across the pooled human baseline and the seven model conditions (columns). Frame colours code the disorder (navy: MDD; orange: GAD; red: BPD). Patients carry disorder-congruent signatures—sadness-led with blunted positive affect for MDD, fear- and anticipation-led for GAD, broad negative affect for BPD—which the therapist mirrors in attenuated form. For compactness, the emotion labels and z-score values are shown only on the human column, which serves as the key; the model flowers show the petals, rings and spokes without text.

3.2 Native speech-to-speech reproduces human competence

LLM trainee competence was assessed by comparing mentor-assigned per-item and total CTRS scores with the human reference (Table 2; Fig. 3). The Gemini-3.1_{LA} dyad scored at essentially the human mean ($M = 29.97$, $SD = 7.06$ vs. human $M = 31.04$, $SD = 11.10$) and, like the humans, remained below the CTRS score 40 competence threshold (Fig. 3c). This is the only condition in which patient and therapist speak and listen in native audio, so the paralinguistic layer of real speech, its prosody, pacing and hesitation, is carried through the exchange rather than flattened into text.

The other conditions, which ran on text or re-synthesised audio, scored well above the human baseline ($M = 41.15$, $SD = 10.49$ for Qwen3.6-35B_T up to $M = 55.99$, $SD = 3.52$ for Gemma-4-E2B_T). Their item scores saturated near the top of the scale on Understanding, Interpersonal effectiveness, and Key cognitions (Fig. 3a). Modality within single models (audio vs. text) matters far less than the gap between native audio and the other LLMs (Table 2). This pattern was consistent across the three patient disorders, with BPD sessions rated slightly lower than GAD and MDD (Fig. 3b). These results suggest that it is native multimodal interaction, not text or re-synthesised speech, that lets the simulated dyad reproduce the competence profile of a real training session.

CTRS item	Human	Gem3.1 _{LA}	12B _A	12B _T	E2B _A	E2B _T	Q3.6 _T	Q3.5 _T
1. Agenda	2.53	1.47	3.61	3.74	4.04	4.12	4.34	3.80
2. Feedback	2.39	2.51	2.38	2.71	4.01	4.03	3.51	4.68
3. Understanding	3.24	3.72	5.97	5.98	6.00	6.00	4.79	5.66
4. Interpersonal effectiveness	3.96	3.75	5.90	5.94	6.00	6.00	4.50	5.42
5. Collaboration	3.27	2.84	4.93	5.21	4.12	4.34	3.72	4.79
6. Pacing	2.92	2.76	4.35	4.81	5.13	5.13	3.59	4.80
7. Guided discovery	2.73	2.78	5.59	5.79	5.34	5.68	3.40	5.17
8. Key cognitions	2.84	3.76	5.65	5.85	5.98	5.99	4.57	5.67
9. Strategy for change	2.69	2.63	4.52	4.92	5.41	5.81	3.85	4.69
10. CBT technique	2.33	2.68	4.89	5.16	4.65	5.38	3.50	5.01
11. Homework	2.14	1.07	0.36	0.56	2.03	3.50	1.37	3.23
Total (mean)	31.04	29.97	48.15	50.66	52.70	55.99	41.15	52.91
(SD)	11.10	7.06	3.49	4.11	3.93	3.52	10.49	10.41

Table 2: Mean CTRS scores by condition. Mean mentor-assigned scores on each CTRS item (0–6) and on the total score (0–66) for the seven model conditions (300 sessions each) and for the human-therapist reference sample ($n = 1,264$). Gem3.1, 12B, E2B, Q3.6 and Q3.5 denote Gemini-3.1, Gemma-4-12B, Gemma-4-E2B, Qwen3.6-35B and Qwen3.5-9B. Subscripts mark the modality: _{LA} = native live audio, _A = synthesised audio, _T = text.

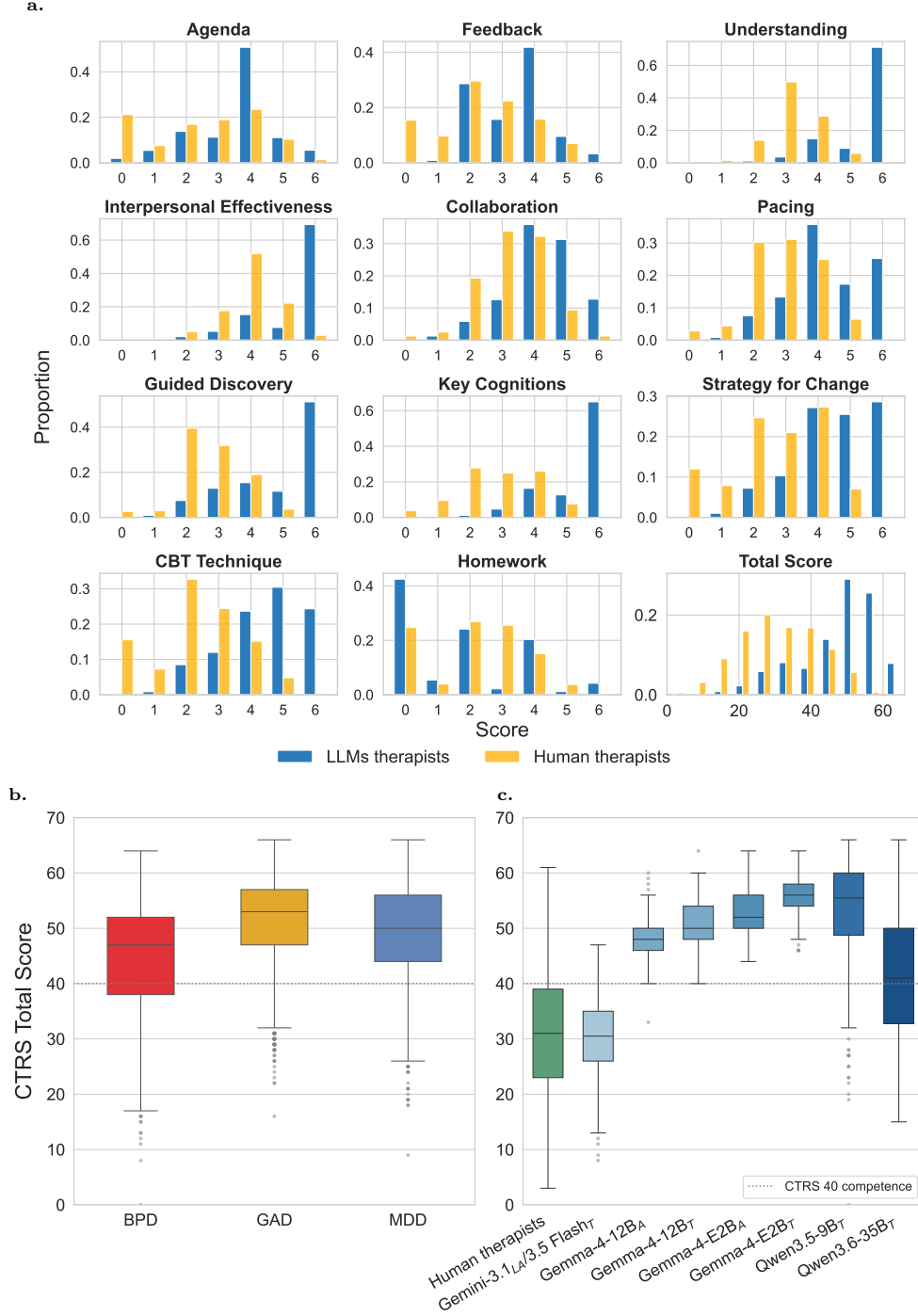


Fig. 3: Mentor-assigned CTRS scores versus human therapists. **a**, Proportion of sessions at each score on the eleven CTRS items and on the total score (bottom right), for all LLM-therapist sessions pooled ($n = 2,100$) versus the human reference ($n = 1,264$; items 0–6, total 0–66). **b**, CTRS total scores by patient disorder, pooled across models. **c**, CTRS totals by model condition alongside the human reference (green); dotted line: $\text{CTRS} \geq 40$ competence benchmark; dots: outlying sessions. Gemini-3.1_{LA}/3.5 Flash_T is the Gemini-3.1 (Live Audio) dyad scored by a Gemini-3.5 Flash mentor. Subscripts mark the modality: _{LA} = native live audio, _A = synthesised audio, _T = text.

3.3 Mentor feedback helps larger models but hurts smallest ones

We compared trainee diagnostic responses before and after supervisory feedback across disorders and conditions (Fig. 4a). Before feedback (Fig. 4c), trainees recognised GAD (97.7%) and MDD (91.0%) with high accuracy. By contrast, BPD recognition was substantially lower (65.3%), with missed cases predominantly misclassified as MDD (15.4%) or assigned hedged multi-disorder diagnoses (14.7%). Mentor feedback shifted this baseline in different directions: BPD accuracy rose to 82.6%, GAD remained stable (97.9%), and MDD accuracy fell to 82.6%, with MDD-to-GAD misclassifications rising from 2.0% to 11.3% (Fig. 4d).

This overall diagnostic shift reflects two opposing effects across model conditions (Fig. 4b). In five of the seven conditions, mentor feedback yielded net positive gains ($g > 0$), primarily by helping trainees recover BPD diagnoses—the case initially hardest to identify unaided (Fig. 4f). Conversely, harmful transitions (Fig. 4e) occurred when feedback prompted trainees to abandon valid initial diagnoses, most frequently causing correct MDD answers to slide into GAD.

This diagnostic deterioration was concentrated in the two smallest-model conditions. For Gemma-4-E2B (in both audio and text modalities), overall accuracy fell after feedback ($g < 0$; Fig. 4a). This decline stemmed from uncritical deference: among sessions where E2B initially selected the correct disorder, 12% to 16% abandoned it once the mentor provided feedback, compared to virtually zero in larger models. Supervisory feedback is therefore beneficial when a trainee is sufficiently capable of integrating clinical guidance; for weaker trainees, however, supervision acts as a trigger to defer, transforming feedback into a source of diagnostic error.

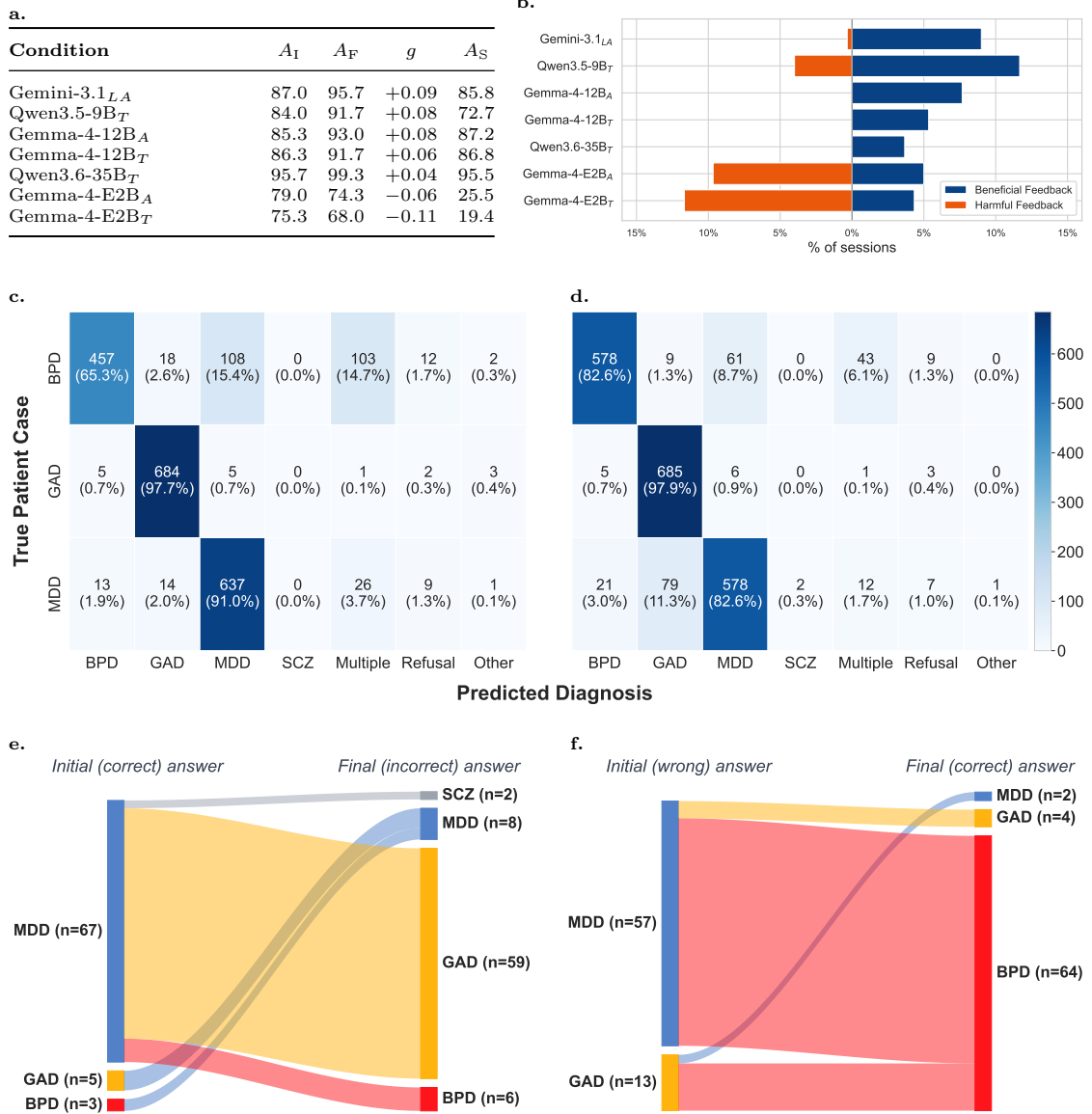


Fig. 4: Diagnostic behaviour around mentor feedback. **a**, Initial (A_I) and final (A_F) diagnostic accuracy, the normalised change ($g = (A_F - A_I)/A_F$), and chance-corrected symptom-identification accuracy (A_S), by condition (300 sessions each; accuracies in %). **b**, Share of sessions per model in which re-diagnosis after feedback turned an initially wrong answer correct (blue) or an initially correct answer wrong (orange); models ordered by net effect. **c**, **d**, Trainee diagnoses before (**c**) and after (**d**) feedback, per true patient case: session counts with row percentages, over the three disorders, schizophrenia (SCZ), hedged multi-diagnosis answers (Multiple), refusals and unmapable answers (Other). **e**, **f**, Initial-to-final diagnosis transitions for sessions changed for the worse (**e**) or for the better (**f**); hedged, refused and unmapable answers excluded. Subscripts mark the modality: $_{LA}$ = native live audio, $_A$ = synthesised audio, $_T$ = text.

3.4 Diagnosis–symptom consistency tracks model scale

The final probe asked the trainee to identify the five key symptoms expressed by the patient during the therapeutic session. Overall, accuracy increased steeply with model size (Fig. 4a; Fig. 5): Qwen3.6-35B named symptoms from the correct disorder block 95.5% of the time, the

mid-sized Gemma-4-12B and Gemini-3.1_{LA} around 86% to 87%, and the smallest Gemma-4-E2B fell to 19% to 26%, close to chance, with the sharpest on BPD.

Importantly, symptom identification was congruent with the final diagnosis (Fig. 6): when trainees correctly identified the patient’s disorder, they also selected many of its defining DSM-5-TR symptoms. In contrast, incorrect diagnoses were accompanied by poorer identification of disorder-specific symptoms. Taken together, the two results suggest that naming the right symptoms is a marker of clinical grounding on both sides of the dialogue, and that this grounding, like the diagnosis itself, appears at model scale rather than being present at every size.

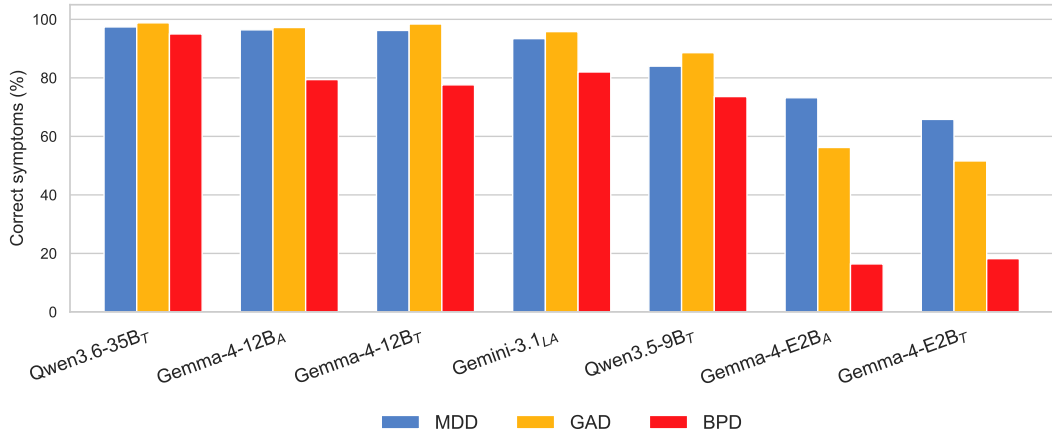


Fig. 5: Symptom identification by model and disorder. Percentage of the five DSM-5-TR diagnostic symptoms identified by the trainee that were concordant with the ground-truth clinical diagnosis for MDD, GAD, and BPD patient cases. Subscripts mark the modality: _{LA} = native live audio, _A = synthesised audio, _T = text.

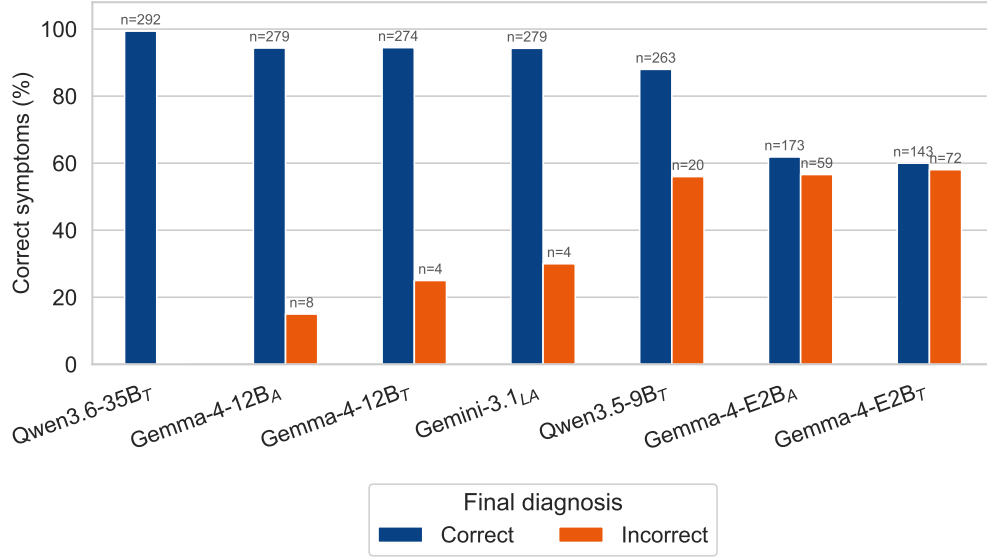


Fig. 6: Symptom identification by final-diagnosis correctness. Mean percentage of the five selected symptoms belonging to the true disorder’s DSM-5-TR block, for sessions whose final diagnosis was correct (blue) versus incorrect (orange). Only sessions naming exactly five distinct symptoms and giving a concrete single-disorder final diagnosis are scored; per-group session counts are printed above each bar. Subscripts mark the modality: $_{LA}$ = native live audio, $_{A}$ = synthesised audio, $_{T}$ = text.

4 Discussion

This study shows that LLMs can simulate a complete CBT mentorship cycle, while revealing that conversational realism, clinical grounding and educational safety are distinct properties. Across conditions, simulated patients expressed disorder-congruent emotional profiles: depression was marked by sadness and reduced positive affect, generalised anxiety by fear and anticipation, and BPD by combined negative emotions. Therapists partly mirrored these emotional profiles while preserving trust and anticipation, reproducing the attenuated affective resonance observed in human counselling. These findings extend earlier work on synthetic mental-health dialogues by showing that prompted personas can organise emotion in clinically meaningful ways rather than merely repeat diagnostic vocabulary [3, 9]. Relational emotional representations made these patterns inspectable across patients, therapists and models [21, 22]. Emotional plausibility alone, however, does not establish therapeutic competence or training value.

The clearest difference in therapeutic competence emerged for native speech. The native speech-to-speech dyad received a CTRS score close to the human reference and remained below the threshold associated with competent CBT delivery. By contrast, text-based and re-synthesised speech conditions were scored well above the human distribution, with several CTRS dimensions approaching the maximum. One interpretation is that native audio preserves hesitation, pacing and paralinguistic signals that make interaction less polished and expose limitations expected

of a trainee. This account is consistent with evidence that clinically relevant information is distributed across linguistic and acoustic features [13, 14]. Yet modality cannot be separated from model architecture and supervisory configuration in the present design. The results therefore do not show that speech alone improves realism. Rather, they suggest that text fluency can create an illusion of therapeutic competence that weakens when interaction becomes temporally and socially more demanding.

Supervision was beneficial only under specific computational conditions. Mentor feedback improved diagnostic accuracy in five of seven conditions, by helping trainees recover initially missed borderline personality disorder cases. Yet the same feedback reduced accuracy in the two smallest-model conditions. These trainees often abandoned a correct initial diagnosis after supervision, most commonly shifting from depression to anxiety. This pattern is compatible with over-deference: weaker models may treat feedback as an instruction to revise rather than as evidence to evaluate. Such behaviour is concerning because fluent explanations need not faithfully reflect a model’s underlying reasoning [40], while instruction-following objectives can privilege external guidance over prior evidence [41]. The result also echoes concerns that persona and role prompting can systematically reshape model behaviour [10]. Symptom recognition showed a similar dependence on scale, ranging from near-chance performance in the smallest conditions to high disorder consistency in the largest model. Greater capacity may strengthen clinical grounding, but it does not guarantee calibrated supervision or safe diagnostic updating.

Limitations and future research. Several limitations constrain these conclusions. In most sessions, the same model instantiated patient, trainee and mentor, allowing shared biases to propagate across the cycle. The synthetic sessions were compared with external human benchmarks rather than rated by blinded CBT supervisors. The emotional reference corpus was in English, pooled across counselling contexts and not stratified by diagnosis, whereas the simulated sessions were conducted in Italian; comparisons with human affect should therefore be considered qualitative. The study examined fixed-length first encounters grounded in three clinical cases and did not capture the demographic, cultural or symptomatic heterogeneity of routine care. More broadly, convincing model behaviour should not be interpreted as evidence that LLMs can replace human participants [17]. Synthetic populations can misrepresent the groups they emulate [11], and language-based mental-health models may fail to generalise across settings or populations [42]. Finally, immediate diagnostic change is not equivalent to durable learning, improved therapeutic skill or patient safety. Future work should evaluate MyMentorLLM as a human-supervised training instrument rather than an autonomous clinical system. Supervisory systems should preserve uncertainty, cite transcript evidence and allow trainees to retain a justified judgement when

feedback is weak. These requirements align with the need for interpretable models in high-stakes decisions [43] and for cognitively meaningful representations of language [20]. Under such safeguards, multi-modal simulation may expand access to deliberate practice [2]. Without them, the same mentorship cycle may amplify overconfidence and error.

In conclusion, MyMentorLLM offers a test bed for determining not only whether synthetic CBT training is possible, but when it is sufficiently faithful, calibrated and safe to be useful.

Acknowledgements. This work was part of the PENSO project, supported by the Ministero dell’Università e della Ricerca (MUR) according to Decreto N. 23178 of 10 December 2024 (Bando FIS 2). The authors acknowledge support from CALCOLO, funded by Fondazione VRT, for support with the computational infrastructure simulating the LLMs.

Declarations

Funding. This research was funded by the Ministero dell’Università e della Ricerca (MUR) within the PENSO project (Decreto N. 23178 of 10 December 2024, Bando FIS 2). Computational infrastructure was provided through CALCOLO, funded by Fondazione VRT.

Competing interests. The authors declare no competing interests.

Consent for publication. Not applicable before publication.

Data availability. Not applicable before publication.

Materials availability. Not applicable before publication.

Code availability. Not applicable before publication.

Author contributions. Designed the original study: RR and AG; Conceptualised the data framework: RR and MS; Gathered and analysed the data: RR; Data visualisation: RR; Formal validation: RR, AG and MS; Supervision: AG and MS; Wrote the manuscript: All authors.

References

- [1] World Health Organization. *World mental health report: Transforming mental health for all* (World Health Organization, Geneva, 2022).
- [2] Rousmaniere, T. *Deliberate practice for psychotherapists: A guide to improving clinical effectiveness* (Routledge, 2024).

- [3] Franchino, E., Rizzi, R., De Duro, E. S., Aghazadeh Ardebili, A. & Stella, M. Digital shadows in mental health map how llms simulate depression, anxiety, and stress through language and psychometrics. *PsyArXiv* (2026).
- [4] Bhatt, S. Digital mental health: Role of artificial intelligence in psychotherapy. *Annals of neurosciences* **32**, 117–127 (2025).
- [5] Carrillo, A. *et al.* Llms can persuade only psychologically susceptible humans on societal issues, via trust in ai and emotional appeals, amid logical fallacies. *arXiv preprint arXiv:2604.16935* (2026).
- [6] Aghazadeh Ardebili, A. & Stella, M. Mapping how llms debate societal issues when shadowing human personality traits, sociodemographics and social media behavior. *arXiv preprint arXiv:2604.27624* (2026). URL <https://arxiv.org/abs/2604.27624>.
- [7] Casoria, L., Neroni, P., Sabatucci, L., Augello, A. & Caggianese, G. *Evaluating llms for synthetic personas generation: A comparative analysis of personality representation and censorship effects. Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter*, 1–9 (2025).
- [8] Argyle, L. P. *et al.* Out of one, many: Using language models to simulate human samples. *Political Analysis* **31**, 337–351 (2023).
- [9] De Duro, E. S., Improta, R. & Stella, M. Introducing counsellme: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans. *Emerging Trends in Drugs, Addictions, and Health* **5**, 100170 (2025).
- [10] Hu, T. & Collier, N. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811* (2024).
- [11] Wang, A., Morgenstern, J. & Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence* **7**, 400–411 (2025).
- [12] Muntigl, P. *Storytelling, Depression, and Psychotherapy*, 577–596 (Palgrave Macmillan UK, London, 2016). URL https://doi.org/10.1057/9781137496850_30.
- [13] Low, D. M. *Speech and text psychometrics: Identifying suicide risk factors with large language models and acoustic networks* (Harvard University, 2024).

- [14] Tao, F., Esposito, A. & Vinciarelli, A. The androids corpus: A new publicly available benchmark for speech based depression detection. *Depression* **47**, 11–9 (2023).
- [15] Guo, T. *et al.* Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).
- [16] American Psychiatric Association, D., American Psychiatric Association, D. *et al.* *Diagnostic and statistical manual of mental disorders: DSM-5* Vol. 5 (American psychiatric association Washington, DC, 2013).
- [17] Dillion, D., Tandon, N., Gu, Y. & Gray, K. Can ai language models replace human participants? *Trends in Cognitive Sciences* **27**, 597–600 (2023).
- [18] Haim, E. & Stella, M. Cognitive networks for knowledge modeling: A gentle introduction for data-and cognitive scientists. *Wiley Interdisciplinary Reviews: Cognitive Science* **17**, e70026 (2026).
- [19] Siew, C. S., Wulff, D. U., Beckage, N. M. & Kenett, Y. N. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity* **2019**, 2108423 (2019).
- [20] Stella, M. *et al.* Cognitive modelling of concepts in the mental lexicon with multilayer networks: Insights, advancements, and future challenges. *Psychonomic Bulletin & Review* **31**, 1981–2004 (2024).
- [21] Fatima, A., Li, Y., Hills, T. T. & Stella, M. Dasentimental: Detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning. *Big data and cognitive computing* **5**, 77 (2021).
- [22] Semeraro, A. *et al.* Emoatlas: An emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science. *Behavior Research Methods* **57**, 77 (2025).
- [23] Plutchik, R. in *A general psychoevolutionary theory of emotion* (eds *Theories of emotion*) Plutchik, R. & Kellerman, H. 3–33 (Elsevier, 1980).
- [24] Al-Mosaiwi, M. & Johnstone, T. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical psychological science* **6**, 529–542 (2018).

- [25] Liu, D. *et al.* Detecting and measuring depression on social media using a machine learning approach: systematic review. *JMIR Mental Health* **9**, e27244 (2022).
- [26] Barnhill, J. W. *DSM-5-TR Clinical Cases* (American Psychiatric Association Publishing, Washington, DC, 2023).
- [27] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5-TR* 5th, text rev. edn (American Psychiatric Association Publishing, Washington, DC, 2022).
- [28] Google DeepMind. Gemini 3.1 flash audio (flash live, tts) — model card (2026). URL <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Flash-Audio-Model-Card.pdf>. Published March 2026, updated April 2026. Accessed 2026-07-27.
- [29] Gemma Team, Google DeepMind. Gemma-4-12b-it — model card (2026). URL <https://huggingface.co/google/gemma-4-12B-it>.
- [30] Zhu, H. *et al.* Omnivoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688* (2026).
- [31] Qwen Team, Alibaba. Qwen3.6-35b-a3b-awq — model card (2026). URL <https://huggingface.co/QuantTrio/Qwen3.6-35B-A3B-AWQ>.
- [32] Google DeepMind. Gemini 3.5 flash — model card (2026). URL <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-5-Flash-Model-Card.pdf>. Published May 2026. Accessed 2026-07-27.
- [33] Gemma Team, Google DeepMind. Gemma-4-e2b-it — model card (2026). URL <https://huggingface.co/google/gemma-4-E2B-it>.
- [34] Qwen Team, Alibaba. Qwen3.5: Towards native multimodal agents (2026). URL <https://huggingface.co/Qwen/Qwen3.5-9B>.
- [35] Goldberg, S. B. *et al.* The structure of competence: Evaluating the factor structure of the cognitive therapy rating scale. *Behavior Therapy* **51**, 113–122 (2020).
- [36] Creed, T. A., Wolk, C. B., Feinberg, B., Evans, A. C. & Beck, A. T. Beyond the label: Relationship between community therapists’ self-report of a cognitive behavioral therapy orientation and observed skills. *Administration and Policy in Mental Health and Mental*

Health Services Research **43**, 36–43 (2016).

- [37] Malhotra, G., Waheed, A., Srivastava, A., Akhtar, M. S. & Chakraborty, T. *Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22)*, 735–745 (ACM, Tempe, AZ, USA, 2022).
- [38] Mohammad, S. M. & Turney, P. D. Crowdsourcing a word–emotion association lexicon. *Computational intelligence* **29**, 436–465 (2013).
- [39] Shaw, B. F. *et al.* Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology* **67**, 837–846 (1999).
- [40] Turpin, M., Michael, J., Perez, E. & Bowman, S. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* **36**, 74952–74965 (2023).
- [41] Wallace, E. *et al.* The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208* (2024).
- [42] Harrigian, K., Aguirre, C. A. & Dredze, M. *Do models of mental health based on social media data generalize? Findings of the association for computational linguistics: EMNLP 2020*, 3774–3788 (2020).
- [43] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **1**, 206–215 (2019).

Supplementary Tables

Table S1: Emotion z-scores by condition and disorder. EmoAtlas z-scores for each of the eight Plutchik emotions (rows), by model condition (columns), split by patient disorder and role. Significant values ($|z| > 1.96$) are in bold. Human is the pooled HOPE baseline (identical across disorders). Subscripts mark the modality: $_{LA}$ = native live audio, $_A$ = synthesised audio, $_T$ = text.

Emotion	Human	G3.1 $_{LA}$	12B $_A$	12B $_T$	E2B $_A$	E2B $_T$	Q3.6 $_T$	Q3.5 $_T$
Major Depressive Disorder (MDD)								
<i>Patient</i>								
Joy	3.09	1.59	3.44	1.83	-1.15	-1.59	0.98	0.64
Trust	5.41	1.84	1.77	0.84	2.87	1.72	2.95	1.30
Fear	-2.40	1.32	0.68	0.75	-0.20	0.32	-1.09	1.67
Surprise	1.60	2.14	4.47	2.59	0.96	0.98	3.73	4.39
Sadness	0.30	3.85	5.12	4.75	2.99	3.45	3.03	5.15
Disgust	-3.93	-1.88	-1.28	-1.12	-2.02	-2.33	-3.07	0.83
Anger	-2.50	-0.82	-0.25	0.62	0.32	0.25	-0.90	2.40
Anticipation	5.77	3.62	4.41	3.83	2.72	2.66	5.24	3.44
<i>Therapist</i>								
Joy	2.34	0.48	1.13	0.06	-1.40	-0.80	-1.71	-0.90
Trust	6.98	3.37	3.30	3.78	4.46	4.66	4.80	4.15
Fear	-1.57	0.17	0.01	-0.15	-0.64	-1.25	-1.48	2.07
Surprise	1.73	1.26	1.90	0.74	1.17	1.23	2.37	3.54
Sadness	-0.31	3.39	3.17	3.03	2.55	2.18	1.73	3.62
Disgust	-5.19	-2.59	-3.52	-2.05	-3.62	-3.92	-4.66	-1.99
Anger	-2.84	-0.88	-1.55	-0.98	-0.81	-1.54	-3.53	-0.06
Anticipation	5.76	4.21	3.16	3.47	4.39	5.59	4.79	4.56
Generalized Anxiety Disorder (GAD)								
<i>Patient</i>								
Joy	3.09	0.62	-2.30	-1.87	-2.05	-2.04	-2.96	-1.72
Trust	5.41	2.37	2.37	2.46	3.05	1.07	3.00	2.73
Fear	-2.40	1.56	3.54	2.09	2.29	2.46	0.73	3.09
Surprise	1.60	2.67	3.18	4.05	3.03	2.77	3.74	4.51
Sadness	0.30	2.30	2.85	2.97	1.89	3.51	1.73	2.94
Disgust	-3.93	-2.75	-3.16	-3.64	-2.80	-3.23	-3.34	-1.40
Anger	-2.50	-1.05	-0.48	0.96	0.05	0.19	-0.94	0.46

continued on next page

Table S1 (continued)

Emotion	Human	G3.1 _{LA}	12B _A	12B _T	E2B _A	E2B _T	Q3.6 _T	Q3.5 _T
Anticipation	5.77	4.89	4.38	4.74	4.58	4.09	4.87	4.54
<i>Therapist</i>								
Joy	2.34	0.32	-1.90	-1.69	-2.15	-2.06	-1.66	-2.11
Trust	6.98	3.97	3.05	4.96	4.12	4.39	5.24	5.42
Fear	-1.57	0.07	3.07	1.38	1.65	-0.44	-1.10	2.09
Surprise	1.73	1.55	3.92	2.59	2.71	2.97	1.87	2.96
Sadness	-0.31	1.45	2.09	1.25	1.26	1.29	-0.23	1.33
Disgust	-5.19	-4.51	-4.10	-4.71	-4.86	-4.57	-5.88	-3.76
Anger	-2.84	-2.24	-1.24	-1.43	-1.27	-1.61	-4.44	-1.01
Anticipation	5.76	4.72	4.69	5.09	6.49	6.18	7.33	4.82
Borderline Personality Disorder (BPD)								
<i>Patient</i>								
Joy	3.09	-1.18	0.06	-0.37	-2.31	-2.06	-2.20	-2.19
Trust	5.41	2.03	1.35	0.81	-0.08	-0.93	2.11	2.76
Fear	-2.40	2.37	3.55	3.35	2.03	2.60	3.30	3.16
Surprise	1.60	2.41	3.95	3.46	1.83	2.54	3.16	4.01
Sadness	0.30	4.12	4.71	3.57	4.66	4.92	3.81	3.54
Disgust	-3.93	-0.28	0.89	-0.45	-0.07	0.08	-0.83	1.15
Anger	-2.50	2.15	1.60	2.23	3.17	2.85	1.61	2.52
Anticipation	5.77	2.13	3.96	2.98	2.37	3.05	3.71	2.66
<i>Therapist</i>								
Joy	2.34	-1.42	-0.89	-1.80	-2.11	-1.17	-3.01	-2.56
Trust	6.98	3.73	2.16	1.79	2.56	2.36	4.69	3.31
Fear	-1.57	1.56	4.08	4.24	1.04	1.45	1.55	3.26
Surprise	1.73	1.28	2.28	2.77	3.00	2.53	3.93	4.00
Sadness	-0.31	3.66	4.38	4.15	4.04	2.80	3.29	3.75
Disgust	-5.19	-2.09	-1.80	-2.26	-3.41	-3.00	-3.72	-1.02
Anger	-2.84	1.28	0.14	1.13	0.96	1.15	-0.89	2.21
Anticipation	5.76	3.27	3.44	3.07	4.70	5.34	5.37	3.09