

Mostly Grounded, Occasionally Risky: Expert Evaluation of LLM-Generated Supervisory Feedback in a Psychotherapy Training Simulator

Adrian Montesano
Open University of Catalonia
amontesano@uoc.edu

Justin K. Bloomberg
Open University of Catalonia
jbloomborg@uoc.edu

Marc Pérez-Buriel
Universitat de Girona
marc.perez@udg.edu

Abstract

Automated feedback is increasingly cited as a key advantage of AI-based psychotherapy training, yet the clinical groundedness of LLM-generated supervisory feedback remains unevaluated. We present an expert evaluation of supervisory feedback generated by PRACTICE, an LLM-powered open-ended psychotherapy training simulator, across 21 feedback instances from four novice trainees. Two clinical psychology experts independently coded 167 feedback propositions as Justified, Unjustified, or Unsure. Inter-rater reliability was near-perfect (raw agreement = 98.2%; $\kappa = 0.902$). Of the 167 propositions, 149 (89.2%) were rated Justified; however, 52.4% of feedback instances contained at least one non-justified proposition, and qualitative analysis identified three recurring failure types: unsupported clinical inference, referential grounding error, and pragmatic failure. In clinical training contexts, even low error rates carry ethical weight: unjustified feedback risks reinforcing inappropriate clinical behaviors in trainees that can be transferred to real practice. These findings provide an initial empirical basis for the responsible deployment of LLM-generated feedback in clinical training and call for traceable, expert-auditable feedback architectures.

1 Introduction

Training novice psychotherapists requires repeated, deliberate practice in conditions that are safe, structured, and clinically realistic. Traditionally, this has relied on role-plays with peers or supervisors, methods that are resource-intensive, difficult to standardize, and constrained by access to supervisors' time. The emergence of large language model (LLM)-based conversational agents has opened a new avenue for scalable, on-demand simulation of clinical interactions.

A growing body of work demonstrates the feasibility and perceived value of this approach. LLM-

driven virtual clients have been shown to be perceived as realistic by trainees, to offer a psychologically safe environment for skill rehearsal, and to support gains in counseling self-efficacy (Beeson et al., 2025; Jeong et al., 2025; Tian et al., 2025). Various platforms have demonstrated feasibility across different training populations, with trainees reporting high usability, learning gains, and value relative to traditional role-play (Cabrera Lozoya et al., 2025; Stige et al., under review). Automated feedback is consistently cited as a key advantage of AI-based training over traditional role-play, and trainees who engage with feedback features tend to report greater satisfaction and perceived clinical impact (Kopelovich et al., 2025b).

Progress has also been made on the technical validation of automated feedback. Kopelovich et al. (2025a) developed and validated a machine learning system capable of scoring discrete CBT competencies in response to standardized client vignettes, achieving strong human-ML agreement (mean ICC = .77; $r_s = .74$). Zilcha-Mano (2025) argues that this work marks a new era for clinical training and envisions its extension to increasingly complex training contexts. Importantly, however, this line of work targets model-specific skills, behaviorally discrete CBT techniques such as psychoeducation, behavioral activation, and reality testing, evaluated through structured, constrained performance tasks.

What remains unaddressed is whether LLM-generated feedback can be clinically grounded in more open-ended conditions, where trainees engage in free dialogue, clinical skills are transtheoretical rather than model-specific, and feedback must be anchored in the emergent dynamics of a real interaction rather than a constrained performance task. PRACTICE (Psychotherapy Realistic AI Conversation for Training In Clinical Expertise) was designed to operate in precisely this space. Rather than model-specific techniques, it targets transtheoretical interpersonal skills grounded in

common factors, the Helping Skills and the Facilitative Interpersonal Skills (FIS) framework, the relational capacities that consistently differentiate therapist effectiveness across theoretical orientations (Anderson et al., 2016; Flückiger et al., 2018; Del Re et al., 2021; Wampold and Owen, 2021). These skills unfold through free, open-ended dialogue with a dynamically responsive virtual patient, and feedback is generated on demand via retrieval-augmented generation (RAG) grounded in clinical training manuals.

Existing work has examined whether trainees perceive feedback as useful (Stige et al., under review) or face-valid (Kopelovich et al., 2025b), but perceived usefulness is not the same as clinical justification. This distinction carries ethical weight: in therapist training, feedback that is not grounded in the actual interaction does not merely fail to help, it actively shapes the clinical habits of practitioners who will eventually work with people in genuine distress.

In this paper, we present the first expert evaluations of clinical groundedness in LLM-generated supervisory feedback produced under open-ended, conversational conditions. We describe PRACTICE and its RAG-based feedback architecture, report expert clinical judgments on 167 automatically generated feedback propositions, and characterize recurring patterns of justification and failure. These findings provide an initial empirical basis for evaluating LLM-generated feedback in clinical training and inform its responsible use.

2 System Description

PRACTICE is a web-based psychotherapy training platform in which trainees conduct open-ended simulated clinical sessions with an LLM-powered virtual patient. The platform uses the OpenAI API governed by a behavioral control architecture designed to counteract the natural tendency of LLMs toward premature helpfulness. Four relational parameters, openness, emotional intensity, collaboration, and readiness for change, are updated turn-by-turn based on the trainee's interventions, producing a dynamically responsive patient whose relational stance evolves as a function of the quality of the trainee's clinical engagement. A gating mechanism prevents the virtual patient from moving toward concrete problem-solving until the trainee has established sufficient relational ground, operationalized as a threshold value of readiness for change.

Three clinical cases are available, covering frustration and confrontation, acute anxiety and emotional dysregulation, and ambivalence in the context of substance use.

Feedback is generated on demand: trainees may request it at any point by entering a dedicated command, which triggers a retrieval-augmented generation (RAG) pipeline, a technique in which the model's output is conditioned on passages retrieved from a predefined document collection rather than relying solely on pretrained knowledge, grounded in the Facilitative Interpersonal Skills (FIS) coding manual (Anderson et al., 2009) and a therapeutic skills training manual developed for the platform. Each feedback response comprises three structured components: validation of observed strengths, identification of one area for improvement with behavioral examples drawn from the session transcript, and an encouraging closing statement (See Appendix A, for an example). At the time of data collection, feedback was generated using gpt-5 via the OpenAI API. The RAG pipeline used LangChain's default retrieval configuration with top- $k = 4$, chunk size of 1,000 characters, and a chunk overlap of 200 characters.

3 Methods

Participants were four psychology trainees (3 women, 1 man; age range 31–42, $M = 37.5$ years) enrolled in a postgraduate clinical training program. All reported fewer than two years of specific psychotherapy training. Trainees completed between 3 and 6 simulated clinical sessions each ($M = 5.25$), selecting freely among the three available clinical cases, with one participant requesting feedback twice within a single session, yielding 21 feedback instances across 20 unique sessions, conducted in Spanish. At the end of the training, trainees rated the overall quality of the feedback received using a brief measure adapted from established frameworks in clinical education (Alsañafi et al., 2023; Johnson et al., 2021).

Each feedback instance was segmented into atomic propositions, yielding 167 codeable units in total. Propositions were defined as minimal units of meaning corresponding to a single clinically interpretable claim (e.g., a skill, behavioral example, or evaluative statement). Segmentation was conducted jointly by both raters prior to independent coding, ensuring a shared unit of analysis. Two clinical psychology experts independently coded

each proposition as Justified (grounded in behavior observable in the session transcript), Unjustified (contradicted by or unsupported by the transcript), or Unsure (insufficient evidence to adjudicate), using a coding system grounded in the Facilitative Interpersonal Skills framework (Anderson et al., 2009). See Appendix A, for an example of an annotated feedback instance. Both raters were male psychologists and psychotherapists, with 9 and 18 years of clinical experience respectively. Raters worked from a standardized protocol including the full session transcript in the original Spanish, the original feedback text, and the propositions, and were blind to each other’s judgments. The feedback instance presented in Appendix A has been translated into English by the authors for accessibility; proposition boundaries and expert verdicts are identical to those applied to the original Spanish text. Inter-rater reliability was assessed using Cohen’s κ (Landis and Koch, 1977). Propositions rated as non-Justified under full rater agreement were subjected to a secondary qualitative analysis to identify systematic patterns of feedback failure.

4 Results

Of the 167 propositions, 149 (89.2%) were rated Justified by both experts, 9 (5.4%) as Unjustified, and 6 (3.6%) as Unsure, yielding 15 (9.0%) non-justified propositions in total (See Figure 1). Inter-rater agreement was high (98.2%; $\kappa = 0.902$, 95% CI [0.79, 1.00]), with disagreements limited to borderline cases between Justified and Unsure. Each participant received at least one non-justified proposition from PRACTICE, with individual rates ranging from 3.2% (P007, 1/31) to 12.7% (P046, 7/55), with P011 at 4/37 (10.8%) and P012 at 3/44 (6.8%). Each feedback instance comprised, on average, 8.0 propositions ($SD = 1.9$) and 277.5 words ($SD = 67.0$). On average, trainees received 0.71 non-justified propositions per feedback instance ($SD = 0.90$). Across a typical five-session training trajectory, this translates to approximately 3–4 non-justified propositions out of ~40 received.

At the feedback-instance level, 11 of 21 feedback instances (52.4%) contained at least one non-justified proposition (See Figure 1). Among these, the mean number of non-justified propositions was 1.36 (range: 1–3) and in 9 of the 11 cases, only a single proposition was non-justified. The remaining 10 instances (47.6%) were fully grounded. Non-justified propositions were disproportionately con-

centrated in the Strengths component of feedback (9/58, 15.5%), compared to Area for Improvement (3/75, 4.0%) and Actionable Suggestions (3/34, 8.8%). Given the small sample size and the nested structure of the data, no inferential test was applied to this comparison.

Qualitative analysis of the 15 non-justified propositions identified three failure types (See Table 1). *Unsupported clinical inference* ($n = 8$) occurred when the described behavior existed in the transcript but was rendered through unverifiable inferences about patient-internal states, mislabeling of the clinical skill involved, or an implied behavioral absence without observable referent. *Referential grounding error* ($n = 3$) occurred when the system drew on real session material rendered inaccurately through inexact or partially incorrect quotation, logically disconnected elements within a single proposition, virtual patient behaviors that were misattributed to the trainee, or behaviors that did not occur in the session as described. *Pragmatic failure* ($n = 4$) occurred when a proposition was found to be semantically ambiguous or unintelligible, impeding reasonable understanding. Notwithstanding these findings, trainees rated the overall quality of the feedback highly ($M = 9.88/10$, range 9.50–10.00).

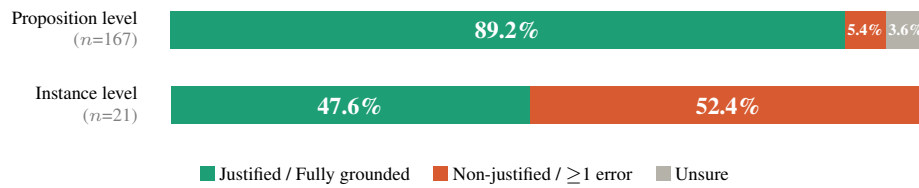
Table 1: Illustrative examples of failures

Failure type	Example proposition
<i>Unsupported clinical inference</i> ($n = 8$)	“You validated her experience (‘We have all the time in the world’)[. . .]” Coder note: <i>That quote does not describe validation.</i>
<i>Referential grounding error</i> ($n = 3$)	“[. . .] you were able to name more precisely what Claudia expressed (<i>fear of being judged, ‘a thousand open eyelashes’</i>) to increase rapport.” Coder note: <i>The quotes given as examples are from Claudia, not the therapist.</i>
<i>Pragmatic failure</i> ($n = 4$)	“For anxiety: co-project a 2-step plan anchored to cues (‘if I notice acceleration: I grab a bottle, breathe 1 cycle, bridge phrase: let me organize the thought’).” Coder note: <i>The confusing language makes the utility of this suggestion difficult to assess.</i>

5 Discussion

The central finding of this study is that PRACTICE generates clinically justified feedback at a high rate: 89.2% of propositions were rated as grounded in observable trainee behavior by both expert raters.

Figure 1: Proportion of justified and non-justified units at two levels of analysis.



This result is encouraging for the responsible deployment of LLM-based feedback in open-ended clinical training, and extends the evidence base beyond trainee-reported face validity (Kopelovich et al., 2025b; Stige et al., under review) to independent expert judgment.

The picture becomes more nuanced at the feedback-instance level. Across 21 feedback instances, 52.4% contained at least one unjustified or uncertain proposition. This does not mean that half of all feedback is clinically problematic, as the majority of propositions within each instance remained grounded. Rather, the proposition-level and instance-level analyses capture different aspects of system reliability: proposition-level rates estimate the overall density of grounded versus non-grounded claims, whereas instance-level rates estimate the likelihood that trainees encounter at least one problematic proposition within an otherwise largely grounded feedback response. In a context of deliberate skill acquisition, even a small number of such inaccuracies may be consequential, as trainees may treat automated feedback as authoritative and lack mechanisms to identify or correct individual errors within otherwise plausible feedback. For example, a trainee may receive positive reinforcement for having “validated” a patient’s experience when the cited intervention does not actually constitute validation. If accepted uncritically, such feedback could, over repeated interactions, reinforce inaccurate understandings of core therapeutic skills and contribute to clinically suboptimal interaction patterns.

This interpretive risk is compounded by a well-documented blind spot in clinical self-assessment. Therapists and trainees systematically overestimate the quality of their own performance and are unable to detect failures in their clinical work without external corrective feedback (Hannan et al., 2005; Hatfield et al., 2010; Walfish et al., 2012). The near-ceiling trainee ratings of feedback quality observed in the present study ($M = 9.88/10$) are con-

sistent with this pattern, suggesting that trainees may not reliably discriminate between justified and non-justified propositions on the basis of perceived quality alone, though this inference was not directly tested in the present design. If clinicians cannot reliably detect failures in their own therapeutic behavior, it is unlikely that novice trainees can detect failures in automated feedback about that behavior. Tracey et al. (2014) further argue that the absence of expertise development in psychotherapy is directly attributable to therapists’ lack of access to quality outcome information, a structural problem that unsupervised LLM-based feedback may reproduce rather than solve. Deliberate practice requires accurate, granular, and correctable feedback as a precondition for skill refinement (Chow et al., 2015); feedback that is partially grounded in unsupported clinical inferences or pragmatically opaque behavioral referents undermines this process.

These risks must be contextualized against what is known about human clinical supervision. To our knowledge, no study has systematically measured the rate of ungrounded claims in human supervisory feedback provided under equivalent conditions, that is, from transcript review without direct session observation. More fundamentally, the goal of clinical supervision is not propositional accuracy but the simultaneous pursuit of formative, normative, and restorative functions: facilitating skill development, ensuring ethical practice, and supporting the supervisee’s professional growth (Bernard and Goodyear, 2019; Milne, 2007). Evaluated against this standard, the feedback generated by PRACTICE, despite its documented imperfections, appears capable of serving a legitimate scaffolding function, opening interpretive possibilities and constructing actionable paths for clinical improvement.

It is worth noting that the operationalization of clinical groundedness adopted here does not reduce to purely descriptive correspondence with the transcript. The coding scheme was applied by clinical experts with 9 and 18 years of psychotherapy

experience, whose judgments necessarily incorporated clinically informed interpretation constrained by observable interactional evidence. Accordingly, propositions rated as Justified could include clinically warranted inferences, not merely literal behavioral descriptions. In this sense, the criterion captures something closer to expert-validated inferential groundedness than to strict transcript mirroring. At the same time, transcript-based evaluation remains inherently limited in its access to psychotherapy processes that may unfold beyond explicit verbal content, particularly those involving relational depth, emotional engagement, or transference dynamics. Thus, even feedback judged as fully grounded may capture only part of the clinically relevant processes involved in therapeutic development. Future work should therefore examine not only whether feedback is grounded, but also whether it is timely, actionable, and clinically generative. Notwithstanding these conceptual caveats, what the 9% non-justification rate does demand is sustained effort to minimize errors and, critically, to prevent the reinforcement of clinically inappropriate behavior: an outcome that did not occur in the present sample, but that falls within the range of risk the current architecture carries. The disproportionate concentration of non-justified propositions in the Strengths component (15.5% vs. 4.0% in Area for Improvement) is plausibly linked to a structurally predictable pressure: a generative model with a template-driven obligation to identify strengths may over-calibrate to linguistic plausibility when transcript evidence is weak, absent or undetected.

The three failure types identified here can be situated within the NLP literature on faithfulness failures in retrieval-augmented generation (RAG). Following [Huang et al. \(2025\)](#), faithfulness-related hallucinations occur when generated content diverges from the source material provided to the model, in this case, the session transcript. Two of the failure categories clearly reflect this pattern: *Unsupported clinical inference* arises when observable transcript content is rendered through unverifiable interpretive inferences or mislabeling of clinical processes, while *Referential grounding error* corresponds to attribution and grounding failures ([Ji et al., 2023](#)), in which behaviors are misattributed to the trainee or cited inaccurately. By contrast, *Pragmatic failure* reflects a pragmatic generation failure in which the proposition becomes difficult to interpret or operationalize despite remaining loosely related to

the session content. Importantly, these errors differ from broader factuality hallucinations, as the system generally has access to the relevant source material through the RAG pipeline but fails to remain fully faithful, interpretable, or clinically grounded in its use of that material. This framing positions the present dataset as a potential resource for developing domain-specific evaluation methods for grounded clinical text generation.

These findings point to a broader architectural challenge for the field. The failure modes documented here are not unique to PRACTICE, they are inherent to any system in which LLM-generated clinical claims remain opaque and unverifiable at the point of delivery. Addressing this requires moving beyond generation toward grounded verification: trained classifiers capable of assessing whether individual feedback propositions are anchored in observable session content, operating over shared, expert-annotated corpora of psychotherapy transcripts. This agenda aligns with calls within the CLPsych community for fine-tuned, domain-specific models that outperform general-purpose LLMs on clinically sensitive annotation tasks ([Tseriotou et al., 2025](#)), and with proposals for the collaborative curation of labeled psychotherapy datasets as a prerequisite for responsible clinical AI ([Imel et al., 2024](#)). The dataset introduced in this study, 167 expert-annotated feedback propositions linked to full session transcripts, represents a relevant step toward such a resource, and notably does so in Spanish, a language severely underrepresented in clinical NLP corpora despite serving hundreds of millions of speakers worldwide.

Limitations

This study has several limitations. First, the sample is small, four participants across 21 feedback instances, which severely constrains statistical power and generalization across training contexts, theoretical orientations, and trainee experience levels. A fully-powered validation study with a larger and more diverse sample is needed before conclusions can be extended beyond this pilot. Furthermore, agreement estimates should be interpreted in light of the constrained coding scheme (three categories), which reduced variability in how units were defined.

Second, this evaluation addresses clinical groundedness and perceived feedback quality, but does not assess pedagogical effectiveness: whether

clinically justified feedback translates into actual skill development remains an open question. A feedback proposition may be grounded in observable behavior yet poorly timed, overly generic, or difficult to operationalize in practice.

Third, no human supervisor baseline was included. Beyond the question of error rates, a comparison between human and LLM-generated feedback on the same session transcripts would constitute a richer research question: to what extent do human supervisors and automated systems converge or diverge in what they identify as strengths and areas for improvement, and which type of feedback trainees find more actionable? Such a design would require expert supervisors to generate written feedback from the same transcripts under equivalent conditions, with resulting propositions subjected to the same coding protocol. This represents a meaningful direction for future work on the clinical utility of automated feedback.

Fourth, PRACTICE operates exclusively in text, whereas psychotherapy is fundamentally a verbal and paraverbal practice. Prosody, silence, tone, and non-verbal attunement, core dimensions of clinical skill, are absent from both the simulation and the feedback. This limits both the ecological validity of the training context and the range of skills the system can meaningfully evaluate.

Fifth, the feedback evaluation was conducted at the proposition level but did not assess which specific clinical skills the system is better or worse at detecting and evaluating. Systematic variation in feedback accuracy across skill domains remains unexamined and constitutes an important direction for future work.

Sixth, the RAG pipeline underlying PRACTICE's feedback mechanism lacks full interpretability: it is not possible to determine precisely which portions of the session transcript drove each generated proposition. This opacity is a general limitation of LLM-based generation in clinical contexts (Schmidt et al., 2025; Yosef et al., 2024), and constrains our ability to explain or correct specific feedback failures.

Ethical Considerations

This study involves secondary analysis of training session data produced by psychology trainees in the context of an institutional educational program. All data were collected under informed consent protocols approved by the relevant institutional review

board. Trainees participated voluntarily, with full awareness that their session transcripts would be used for research purposes. Participant identifiers have been anonymized throughout. Raters were clinical psychology experts who worked from a standardized evaluation protocol and were blind to each other's judgments; no participant data beyond session transcripts and pre-atomized feedback propositions were made available to them.

The system evaluated here trains clinical skills through interaction with LLM-powered virtual patients. While no real patients are involved, the competencies developed and the clinical habits reinforced by feedback will ultimately be applied to people in genuine distress. Feedback perceived as authoritative, even when not clinically justified, may shape clinical practice in ways that are difficult to detect or correct. The low rate of unjustified propositions observed here (9%) is encouraging, but should not be taken as grounds for unsupervised deployment.

PRACTICE incorporates a disclaimer notifying trainees that feedback is generated entirely by an AI system for training purposes, should not be interpreted as clinical supervision or faculty evaluation, and may contain inaccuracies inherent to automated models. Users are encouraged to contrast feedback with supervised training and authorized professional sources. This represents a meaningful mitigation measure, though it does not substitute for the architectural safeguards recommended in the Discussion. We recommend that AI-generated feedback in clinical training contexts be used as a complement to, not a substitute for, expert clinical supervision, and that ongoing expert auditing be integrated into deployment pipelines.

References

- A. Alshafi, D. L. X. Ling, M. Newell, and T. Kropmans. 2023. [A systematic review of effective quality feedback measurement tools used in clinical skills assessment](#). *MedEdPublish*, 12(11).
- T. Anderson, M. E. J. Crowley, L. Himawan, J. K. Holmberg, and B. D. Uhlin. 2016. [Therapist facilitative interpersonal skills and training status: A randomized clinical trial on alliance and outcome](#). *Psychotherapy Research*, 26(5):511–529.
- T. Anderson, B. M. Ogles, C. L. Patterson, M. J. Lambert, and D. A. Vermeersch. 2009. *Facilitative interpersonal skills coding manual*. Ohio University.
- E. T. Beeson, E. Counselman-Carpenter, T. A. Field,

- and M. Ghoston. 2025. [Artificial intelligence role-play for counselor skill development: A qualitative investigation](#). *Counselor Education and Supervision*.
- J. M. Bernard and R. K. Goodyear. 2019. *Fundamentals of Clinical Supervision*, 6 edition. Pearson.
- D. Cabrera Lozoya, E. Hernandez Lua, J. A. Barajas Perches, M. Conway, and S. D'Alfonso. 2025. Synthetic empathy: Generating and evaluating artificial psychotherapy dialogues to detect empathy in counseling sessions. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 152–163. Association for Computational Linguistics.
- D. L. Chow, S. D. Miller, J. A. Seidel, R. T. Kane, J. A. Thornton, and W. P. Andrews. 2015. The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy*, 52(3):337–345.
- A. C. Del Re, C. Flückiger, A. O. Horvath, and B. E. Wampold. 2021. Examining therapist effects in the alliance-outcome relationship: A multilevel meta-analysis. *Journal of Consulting and Clinical Psychology*, 89(5):371–378.
- C. Flückiger, A. C. Del Re, B. E. Wampold, and A. O. Horvath. 2018. The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, 55(4):316–340.
- C. Hannan, M. J. Lambert, C. Harmon, S. L. Nielsen, D. W. Smart, K. Shimokawa, and S. W. Sutton. 2005. A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, 61(2):155–163.
- D. Hatfield, L. McCullough, S. H. B. Frantz, and K. Krieger. 2010. Do we know when our clients get worse? an investigation of therapists' ability to detect negative client change. *Clinical Psychology & Psychotherapy*, 17(1):25–32.
- L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Z. E. Imel, M. Steyvers, and D. C. Atkins. 2024. Computational psychotherapy research: Current status and future directions. *Behaviour Research and Therapy*, 172:104430.
- H. Jeong, S. Lee, and B. K. Sohn. 2025. Exploring ai-based role-play simulation for mental health counseling training. *Psychiatric Services*.
- Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- C. E. Johnson, J. L. Keating, M. Leech, P. Congdon, F. Kent, M. K. Farlie, and E. K. Molloy. 2021. Development of the feedback quality instrument: A guide for health professional educators in fostering learner-centred discussions. *BMC Medical Education*, 21(1):382.
- S. L. Kopelovich, K. Doud, M. Ofer, and B. Stiles. 2025a. Development and validation of a machine learning system for rating CBT for psychosis competencies. *Psychotherapy*, 62(1):1–11.
- S. L. Kopelovich, K. Doud, M. Ofer, and B. Stiles. 2025b. Trainee and supervisor perspectives on ai-generated feedback in CBT for psychosis training. *Psychotherapy*, 62(1):12–21.
- J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- D. Milne. 2007. An empirical definition of clinical supervision. *British Journal of Clinical Psychology*, 46(4):437–447.
- F. Schmidt, K. Hammerfald, H. H. Jahren, and V. Vlassov. 2025. CFiCS: Graph-based classification of common factors and microcounseling skills. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 106–115. Association for Computational Linguistics.
- S. H. Stige, A. Montesano, and K. Aafjes-van Doorn. under review. Use of generative AI as a path to strengthen clinical training: First steps and emerging challenges for active learning. *Active Learning in Higher Education*. Under review.
- Z. Tian, L. Yang, and W. Liang. 2025. Ai-simulated clients in counseling training: Effects on self-efficacy and skill development. *Training and Education in Professional Psychology*.
- T. J. G. Tracey, B. E. Wampold, J. W. Lichtenberg, and R. K. Goodyear. 2014. Expertise in psychotherapy: An elusive goal? *American Psychologist*, 69(3):218–229.
- T. Tseriotou, J. Chim, A. Klein, A. Shamir, G. Dvir, I. Ali, C. Kennedy, G. S. Kohli, A. Hills, A. Zirikly, D. Atzil-Slonim, and M. Liakata. 2025. Overview of the CLPsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 178–196. Association for Computational Linguistics.
- S. Walfish, B. McAlister, P. O'Donnell, and M. J. Lambert. 2012. An investigation of self-assessment bias in mental health providers. *Psychological Reports*, 110(2):639–644.

B. E. Wampold and J. Owen. 2021. Therapist effects: History, methods, magnitude, and characteristics of effective therapists. In M. Barkham, W. Lutz, and L. G. Castonguay, editors, *Bergin and Garfield's handbook of psychotherapy and behavior change*, 7th edition, pages 297–326. Wiley.

S. Yosef, M. Zisquit, B. Cohen, A. Brunstein Klomek, K. Bar, and D. Friedman. 2024. Assessing motivational interviewing sessions with ai-generated patient simulations. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 1–11. Association for Computational Linguistics.

S. Zilcha-Mano. 2025. [Commentary: Toward a new era in psychotherapy training](#). *Psychotherapy*, 62(1):22–27.

A Appendix A. Example Feedback Instance with Expert Annotations

The following feedback instance (Participant P011, Session 3) illustrates the structure of PRACTICE-generated feedback and the proposition-level annotation scheme applied by expert raters. Each proposition is labeled with its feedback component and expert verdict.

Strengths

1. “*You demonstrated consistent and specific validation of Pedro’s distress (‘you feel singled out...’; ‘I’m not here to judge you’). This reduced his defensiveness and increased his openness.*” [Unsure]
2. “*You also conducted a good functional exploration of his substance use (disconnecting from stress → drinking → conflict at home).*” [Justified]
3. “*...and connected with personal resources (walking with music, cooking paella).*” [Justified]
4. “*You concluded with a small, concrete behavioral plan and a backup plan, reinforcing agency and realistic hope.*” [Justified]

Area for Improvement

5. “*More explicit alliance repair when Pedro feels judged: you could name the relational tension at the moment (‘I notice this sounds like judgment to you; if it happens here, I’d prefer we address it openly’). This strengthens the alliance and prevents ruptures.*” [Justified]

6. “*Avoid quickly resorting to exercise as a solution. Instead, focus more on their language and motivation (e.g., ‘turn down the volume in their head’) and co-create equivalent micro-alternatives for disconnecting. Frame the invitation as an experiment, not an obligation.*” [Justified]

7. “*Delve into ambivalence and autonomy: explore the pros and cons of drinking and alternatives with open-ended questions (‘what does it give you and what does it take away, when is it easier for you to stop?’). Reflect on Pedro’s own reasons for change and ask for his permission before making suggestions.*” [Justified]

8. “*Close with agency marker and follow-up: agree on how you will measure success (e.g., ‘I walk = success’) and what you will do if an obstacle arises. Encourage participants to bring specific data to the next session.*” [Justified]

Note. PRACTICE operates in Spanish; all training sessions and generated feedback are conducted in that language. The feedback instance presented in this appendix has been translated into English by the authors for accessibility. Proposition boundaries and expert verdicts are identical to those applied to the original Spanish text. Propositions are labeled as Justified, Unjustified, or Unsure based on full agreement between both expert raters. Proposition 1 was rated Unsure by both raters: the behavioral referents cited (the quoted phrases) are present in the transcript, but the causal claim, that they “reduced defensiveness and increased openness”, was deemed insufficiently grounded in observable evidence. This instance contains 1 of 15 non-justified propositions in the dataset, concentrated in the Strengths component, consistent with the pattern reported in the Results section.