Assessing Mental Health Counseling Interventions with LLMs

Nurun Naher
University of Central Florida
Orlando, FL

Joshua Langfus and Linda Pfiffner
University of California, San Francisco
San Francisco, CA

Gita Sukthankar
University of Central Florida
Orlando, FL

***Abstract*—Large Language Models (LLMs) are becoming increasingly valuable for psychologists analyzing the large volume of data that is collected from human subjects transcripts, talk-based therapy inventions, and other types of counseling sessions. This paper presents a case study on using LLMs to assess the quality of counseling interventions. The aim of this system was to evaluate the treatment fidelity of school clinicians delivering a group-based behavioral parent training intervention for ADHD (Attention Deficit Hyperactivity Disorder). We examine the following research questions: 1) how consistent is the LLM in its assessment if queried repeatedly? 2) how well do LLMs agree with expert human coders? 3) do different LLMs agree with one other? Our results show that Gemini slightly outperforms GPT-4o at assessing treatment fidelity and produced qualitatively superior explanations for its decisions. Based on our experiments, we propose a set of best practices for using LLMs for evaluating counseling interventions.**

I. Introduction

Large Language Models (LLMs) are increasingly being considered for applications in mental health assessment, intervention delivery, and behavioral support. Across review studies and empirical investigations, researchers have explored LLMs' potential to support diagnosis, simulate therapy, summarize patient narratives, and evaluate clinical fidelity [1]–[3].

Scoping reviews have found that LLMs are being deployed for generative tasks like therapy simulation, reflective listening, and clinical decision support in psychotherapy and psychiatry [4]–[6]. For example, Guo et al. [7] report growing interest in LLM-based mental health tools, especially those integrating structured outputs like SOAP notes or risk screening summaries.

Despite their promise, the use of LLMs in clinical contexts is met with caution. Studies such as [8] and [9] have raised critical concerns around ethical risks, including hallucinated diagnoses, reproducibility challenges, and limitations in emotional nuance. Lawerence et al. [10] further note that LLM-generated outputs can carry unintended biases that might adversely affect vulnerable populations if deployed without sufficient oversight.

From an implementation standpoint, researchers have highlighted practical limitations such as prompt sensitivity, lack of transparency, and concerns about privacy and trust [7], [11]. These issues have led to calls for human-in-the-loop deployment strategies that maintain clinical accountability while leveraging LLMs' capabilities.

Meanwhile, proposals like the one from [12] emphasize the need for responsible evaluation frameworks, suggesting that mental health-specific benchmarks and hybrid workflows could guide safe integration into behavioral healthcare. LLMs offer exciting opportunities for augmenting mental health assessment and behavioral intervention, particularly in low-resource and high-volume environments. However, their effectiveness and safety are highly context-dependent, and their real-world deployment must be accompanied by domain-specific constraints, expert validation, and ethical safeguards.

This paper presents a case study on the application of LLMs to assess the quality of counseling interventions. Unlike free-form human conversations, clinical interventions typically follow a more structured script. In psychosocial and behavioral interventions, assessing treatment fidelity is crucial for determining the "dose" and quality of the intervention delivered. Accurate fidelity assessment informs both treatment validity and implementation feasibility. However, human-conducted fidelity assessments are widely reported as time-consuming, costly, and difficult to scale. Raters—typically doctoral- or master's-level clinicians—must be extensively trained and calibrated to evaluate not only content adherence but also delivery quality. This level of expertise is expensive and often scarce, especially in under-resourced contexts [13], [14].

Manual fidelity assessment is especially labor-intensive. For instance, implementing and evaluating programs like the Collaborative Life Skills (CLS) intervention [15] entails reviewing over 25 hours of recorded training sessions per site. Scaling such fidelity reviews across multiple schools, clinics, or sites creates substantial logistical and financial barriers. Similar resource constraints have been observed in other domains such as medical education and behavioral health [16], [17].

Empirical studies consistently highlight three major challenges in human fidelity assessment: high costs, reliance on specialized human expertise, and limited scalability. Costs arise not only from expert compensation but also from training, calibration, and rater consensus meetings [13], [18]. Expertise constraints are further exacerbated by inter-rater variability and subjective interpretation, which undermine reliability [19], [20]. Furthermore, traditional manual approaches are not feasible for large-scale training programs or high-throughput environments like online education or national clinical trials [21].

These persistent barriers hinder the sustainability and reach

of effective interventions. Consequently, recent research has explored automation—especially via large language models (LLMs)—as a scalable, cost-effective alternative. Multiple studies report that LLMs such as GPT-4 can achieve high alignment with expert ratings (Cohen's kappa up to 0.88), with accuracy as high as 95.7% and F1 scores up to 0.92 [16], [17], [22]. Hybrid human-AI systems have further demonstrated human-AI agreement levels of up to 96%, while also reducing labor costs by up to 57% [18], [23].

These findings indicate that automation using LLMs not only addresses current challenges in fidelity assessment but also enables a transition toward more efficient, consistent, and scalable evaluation frameworks. While nuanced human judgment remains important in certain contexts, emerging evidence supports the use of LLMs to augment and, in some cases, replace traditional fidelity assessment workflows-especially where resources are limited or throughput is high.

We developed an LLM-powered system to evaluate the treatment fidelity of school clinicians delivering a group-based behavioral parent training intervention for ADHD. This paper examines the following research questions:

1) how consistent is the LLM in its assessment if queried repeatedly?
2) how well do LLMs agree with expert human coders?
3) do different LLMs agree with one other?

II. Related Work

Fidelity assessment—evaluating how well practitioner behavior adheres to established protocols—is a central component of behavioral training. Traditionally, this process relies on human experts who rate trainee responses using structured rubrics. However, human coding is time-intensive, costly, and subject to inter-rater variability. As a result, researchers are exploring whether large language models (LLMs) can help automate fidelity assessment in behavioral interventions such as cognitive behavioral therapy (CBT).

Emerging research shows that LLMs, when guided by explicit scoring rubrics, can approximate human performance in specific fidelity-related tasks. For example, Kurland et al. [24] demonstrated that GPT-4 was able to assess story retelling in aphasia rehabilitation with high alignment to human ratings, a task analogous to concept recognition in CBT fidelity coding. Similarly, Flemotomos et al. [25] used contextualized language representations to automatically assess CBT session quality, showing promising agreement with expert coders.

In broader educational and communication contexts, Pilny et al. [26] assessed LLMs' ability to conduct content analysis, concluding that machine-based scores closely tracked human judgment under structured conditions. This is supported by findings from [27], who systematically compared LLMs with human raters across multiple domains and highlighted cases where LLMs can reliably replace manual scoring, particularly for discrete and well-bounded tasks. In another research, Tai et al. [28] examined the application of LLMs to aid in deductive coding within qualitative research methodologies. Their study demonstrated the potential of these models to support traditional coding processes by providing consistent, reliable outputs. This approach is significant because it helps mitigate human bias, a common issue in qualitative data analysis, while also enhancing the efficiency and scalability of coding practices.

Despite these promising results, limitations persist. Iftikhar et al. [19] found that while LLMs could simulate therapeutic interactions, psychologists rated human peers as better at managing emotional nuance in CBT tasks. Likewise, [29] reported that LLMs perform well on linguistically concise prompts but fall short when assessing behaviorally rich or ambiguous content.

Accuracy is also highly context-dependent. For instance, Mahmoudi et al. [30] found that LLMs achieved over 90% accuracy in explicit data extraction tasks but performed poorly when interpreting subjective behavioral components. In domains such as harmful content detection, LLMs like GPT-4 showed strong precision and recall, especially when integrated into human-in-the-loop workflows [31].

More recent proposals advocate for computational frameworks specifically tailored to LLM-driven behavioral assessment. For example, [32] outlined a structured system to evaluate the therapeutic quality of LLM dialogue agents, calling for domain-specific validation benchmarks.

In summary, LLMs show substantial promise in automating fidelity assessment when criteria are explicit and task design is well-controlled. Their role is particularly valuable in scaling up training environments, reducing coding burdens, and supporting formative feedback. However, they remain less effective in open-ended, emotionally complex interactions without careful human oversight. This paper compares two of the largest LLMs (OpenAI's GPT-4o and Google's Gemini 2.5) and presents a set of best practices, drawn from our experience, on using an LLM for fidelity assessment.

III. Method

A. Data Collection

School mental health providers (SMHPs) implemented an eight-week multi-component curriculum targeting attention and behavior problems as part of the Collaborative Life Skills (CLS) program. CLS is an evidence-based treatment for ADHD in school-age children that includes parent-groups, student groups, and classroom daily report cards. By leveraging data from an existing clinical trial, we compared the LLM ratings and observations to human clinicians' ratings of the same content in each session.

Research staff met with SMHPs every week to teach and review the intervention content. SMHPs then led a 60-minute parent group and a 60-minute student group each week with 6 to 8 students they identified might benefit from the program and who met study inclusion criteria. All meetings were recorded and transcribed. To be eligible, students' parents and teachers had to endorse at least six or more symptoms of inattention or hyperactivity with some impairment both at home and at school. All children assented to participate and parents completed informed consent documents approved by

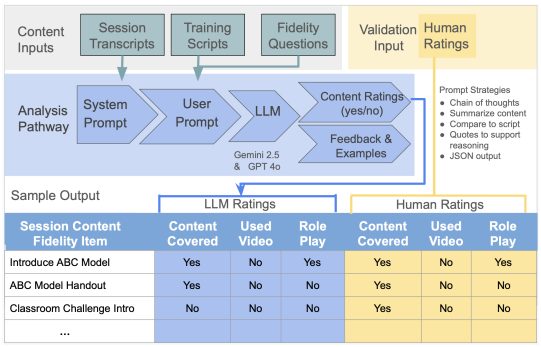

Fig. 1. System architecture: This diagram illustrates how session transcripts, training scripts, and fidelity questions are combined to generate user prompts for the LLM, which is used to rate content. The resulting JSON output from the LLM is then paired with human ratings for further analysis.

the institutional review board at University of California, San Diego.

During all training and intervention sessions, masters- and doctoral-level clinicians observed and rated treatment fidelity using a Qualtrics assessment form. Raters noted whether key content areas were covered, and for the consultations, whether a teaching video or role-play exercise was used. For the current analyses, only a subset of the data from the SMHP training meetings was analyzed. To preserve the privacy of the participants, all transcripts were preprocessed to remove personally identifiable information (PII).

B. LLM-Based Assessment Pipeline

Figure 1 shows our system architecture. To reduce ambiguity and make post-processing easier, the LLM was prompted to respond in json format. First we gave a high-level system prompt:

> You are a psychology researcher who is an expert in behavioral parent training techniques. You are reviewing transcripts of recorded meetings between a team of researchers (called "trainers") and school social worker (called "leaders").
> The trainers are training the leaders to deliver a behavioral parent training intervention to parents at their schools. This intervention involves parents learning new skills and techniques for parenting. The trainers are using a script to teach the leaders how to deliver these skills to the parents. They are following the content in the script, however they can also use their own words to convey the concepts.
> The text below following the word TRANSCRIPT is the transcript of the meeting.
> The user prompt contains a label for the content the leaders were supposed to cover (e.g., "Program Introductions") and the script that the trainers were supposed to follow to cover that content.

We prototyped five user prompts to cover the following content types from the CLS program:

1) Home Activities Check-in
2) Define Consequences that Influence Behavior
3) Handout: Rewards and Negative Consequences
4) Introduce the Home Challenge
5) Types of Home Challenges

Content quality was judged using following questions which were also posed to human raters on Qualtrics:

1) Did the trainers cover the content?
2) Did the trainers themselves demonstrate how to present the outline to the parents?
3) Did the trainers use a reference video to teach this material?
4) Did the leaders engage in a role play exercise to practice the material with the trainers?

An example LLM prompt is shown in the appendix.

C. Maximizing Response Consistency

Due to the combination of sampling, parallelism, and floating-point rounding differences, it is often impossible to make modern LLMs completely deterministic. In LLM API calls, temperature is a critical parameter that controls the randomness and creativity of the generated text. By setting the temperature to zero, it forces the model to act as a greedy decoder that chooses the word with the highest predicted probability. Some APIs also provide a seed parameter to initialize the pseudo-random number generator to a consistent value. For our fidelity evaluations, we observed a slight response drift even after controlling these parameters. Therefore, we performed multi-pass consensus over three sets of ratings (described in Section IV)

D. LLM Model Selection

This paper compares the performance of two LLMs, GPT-4o and Gemini 2.5, at assessing counseling fidelity. Since our task requires including long transcripts in the prompt, we only evaluated models with larger context windows. GPT-4o is the latest multimodal language model developed by OpenAI, capable of processing and generating text, images, and audio with high speed and accuracy. While the exact parameter count remains undisclosed, GPT-4o represents a significant leap in performance and efficiency compared to previous versions like GPT-4 Turbo. It supports long-context processing (128,000 tokens) and demonstrates advanced reasoning and understanding, making it particularly well-suited for large-scale textual analysis tasks such as summarization, sentiment analysis, and semantic extraction across diverse data sources.

Similarly, Gemini 2.5 is a cutting-edge multimodal AI model developed by Google DeepMind, designed to understand and generate text, code, images, and other complex data types. As the latest iteration in the Gemini series, it builds on previous versions with improved reasoning, memory, and multimodal capabilities. Gemini 2.5 is adept at handling long-form content with its context window of 1 million tokens. It performs well at complex tasks such as document analysis, code generation, and knowledge extraction.

IV. Evaluation

To assess the reliability, alignment, and added value of large language models (LLMs) in fidelity coding, we adopted a structured, multi-level evaluation approach. Our strategy moved beyond simple binary agreement metrics to more nuanced consensus scoring and comparison with human ratings.

A. Multi-Pass LLM Consensus

Each LLM (GPT-4o and Gemini 2.5) was queried three separate times per fidelity question to account for inherent variability in model responses, even at low temperature settings. To identify each LLM response, we will refer them as R1, R2 and R3 respectively. Each completion included binary ratings ("yes"/"no") along with reasoning and transcript-based quotes. We computed:

- **Intra-Model Agreement:** Percent agreement and Cohen's Kappa were used to assess internal consistency across the three runs per LLM. Gemini 2.5 showed high within-model reliability with 96.67% agreement and $\kappa = 0.85$. GPT-4o also demonstrated strong consistency, with 93.33% agreement and $\kappa = 0.79$.
- **Majority Vote Consensus:** For each model, the final AI rating per item was determined by majority vote across its three completions (2/3 or 3/3 agreement). This consensus rating was used in subsequent human-AI comparisons.

B. Human-AI Consensus Evaluation

To benchmark LLM performance, we compared AI consensus ratings to fidelity judgments made by trained human coders (treated as the gold standard). Evaluation metrics included:

- **Raw Agreement Rates:** Binary alignment between human rating and LLM majority vote.
- **Cohen's Kappa:** Used where ordinal confidence scores were available.

C. Cross-Model Comparison

To examine how different models interpret the same content, we compared the GPT-4o and Gemini 2.5 consensus outputs across all fidelity items. Inter-model agreement was evaluated using raw percent agreement and Cohen's Kappa. Differences in scoring thresholds and reasoning patterns were qualitatively analyzed. Discrepancies often reflected subtle interpretive or emphasis differences between the models.

D. Qualitative Review of Reasoning

Beyond numerical agreement, we qualitatively assessed the LLMs' reasoning and quote selection for:

- Relevance to fidelity criteria
- Appropriateness of justifications
- Common reasoning errors (e.g., overgeneralization, superficial quote matches)

This layer of analysis helped evaluate the interpretability and transparency of AI ratings, key factors for clinical deployment.

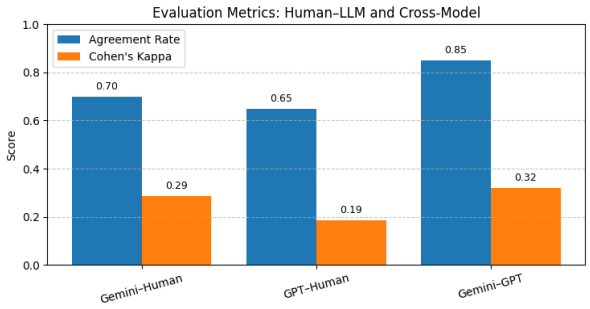

Fig. 2. Evaluation metrics comparing agreement between human raters and two LLMs (Gemini 2.5 and GPT-4o) using agreement rate and Cohen's Kappa. The chart also includes cross-model comparison (Gemini vs. GPT-4o).

V. Results

We evaluated the alignment between LLM-generated fidelity ratings and those provided by trained human coders. For each item, a consensus LLM rating was computed via majority vote across three independent completions. Human ratings served as the gold standard for comparison.

A. Agreement Between Human and LLM Ratings

Figure 2 summarizes the agreement metrics between human and LLM and cross LLMs consensus ratings. The Gemini model achieved a higher match rate and reliability with human raters, indicating more consistent alignment with human-defined fidelity criteria. In contrast, the GPT model, while still showing above-chance agreement, exhibited a lower Cohen's Kappa, suggesting potential differences in decision thresholds or interpretation strategies compared to human coders.

B. Agreement Between Different LLM Ratings

To evaluate the consistency between the two LLMs used in this study, we conducted a direct comparison of fidelity ratings generated by GPT-4o and Gemini 2.5. Each model produced a majority-vote rating for each transcript item based on three independent completions. The goal of this analysis was to examine the degree to which the models aligned in their application of fidelity criteria across a shared dataset. Figure 2 shows the results as Gemini + GPT. While the raw agreement rate indicates that the models provided the same binary rating on the majority of items, the Cohen's Kappa value suggests only modest agreement beyond chance. This discrepancy highlights the importance of using adjusted reliability metrics when evaluating automated scoring systems, particularly in domains where agreement by chance may be high due to imbalanced label distributions.

The modest Kappa coefficient may reflect underlying differences in model behavior, including sensitivity to instructional phrasing, emphasis on different features of the transcript, or interpretive flexibility in ambiguous cases. While both models followed the same prompt structure, qualitative review of disagreement cases (discussed in the next section) revealed that they occasionally prioritized different elements of the session when making their determinations.

C. Nature of Disagreement

Across both models, disagreements with human ratings were observed in approximately 30–35% of items. Qualitative review of these cases revealed several recurring patterns:

1) LLMs occasionally inferred content coverage from partial or indirect references: To illustrate how interpretive variability manifests in practice, we present a representative example of an LLM rating divergence for the content item *"Define Consequences that Influence Behavior"* from Human rating and Gemini 2.5. R1 rated ***no***, whereas R2 and R3 and human rated ***yes***. Below is R1's justification for rating ***no***:

> "Trainers introduced the topic of defining consequences and covered that it involves revisiting the ABC model and focusing on positive strategies first. However, crucial introductory elements from the script, such as immediately linking the ABC model to the 'Home Challenge' and the detailed 'ASK PARENTS' segment about prioritizing positive strategies, were not covered as prescribed for this specific section. These elements were either deferred or addressed in a less detailed manner later in the transcript, altering the scripted flow for this content block."

In contrast, R2 rated the same item as ***yes*** and offered the following rationale:

> "The trainers covered most key points from the script. They instructed leaders on revisiting the ABC model, defining positive and negative consequences, and the program's emphasis on starting with positive strategies before negative ones. This was achieved through direct instruction, a reference video, and preparing leaders for role-play. For instance, TRAINER1 mentioned revisiting the ABC model and teaching about rewards and negative consequences. The reference video also highlighted focusing on positive consequences first. However, the specific script point 'ASK PARENTS' Any ideas why we start with positive strategies?' along with its detailed prompts was not explicitly instructed to be delivered by the leaders to the parents in that interactive format; TRAINER1's instruction was more about the content flow."

This example underscores how even structured prompts and zero-temperature settings may not eliminate variability when fidelity judgments involve interpretive nuance or prioritization of script structure. While R1 focused on the absence of scripted delivery elements and sequence fidelity, R2 emphasized general coverage of the core ideas.

2) Some LLM ratings favored leniency in marking fidelity as present, possibly due to overgeneralization from minimal evidence: Several false-positive cases revealed a pattern in which the LLM inferred fidelity based on superficial mention of a topic, even when deeper instructional components or required formats (e.g., question prompts, interactive discussions) were absent. For instance, a reference to "consequences" in a casual aside or brief mention of a concept sometimes led models to rate the item as "yes" despite missing core script elements such as modeling, parent engagement, or leader preparation. This suggests an LLM tendency to generalize from minimal semantic cues without strict alignment to the scripted delivery expectations.

3) Variability in quote selection and interpretive depth may have contributed to divergent decisions: Even when models agreed on the presence of key topics, they often cited different sections of the transcript to support their ratings. In some cases, selected quotes lacked the instructional or interactive dimensions expected by human raters, while in others, LLMs over-interpreted loosely related content. This variability in quote selection may reflect differences in how each model interprets task framing, the salience of evidence, or alignment with script-based standards. It highlights the need for better prompt engineering or rating calibration when using LLMs for fine-grained fidelity analysis.

VI. Lessons Learned

1) Augmenting the prompt with as much additional context as possible improves assessment performance. Thus, LLM context window length may matter more than model size.
2) LLM consistency can be improved by specifying the seed value for the pseudo-random number generator and setting the temperature to zero, but the output will not be totally deterministic due to hardware issues. Majority vote consensus can be used to be produce more stable values.
3) Due to overlap in training data and similarities in generative architectures, LLMs may agree with one another more than they agree with human raters.
4) Outputting assessments in json format reduces ambiguity in the model response and make post-processing easier.

VII. Conclusion and Future Work

Our results indicate that while LLMs hold promise for semi-automated fidelity coding, their outputs should be interpreted with caution, especially in edge cases. The observed discrepancies underscore the importance of combining automated scoring with human oversight, particularly for fidelity judgments that require nuanced understanding of context and intent. However, we believe that LLM-based assessment of counseling interventions is so time and cost effective that it should be regularly employed to ensure that the human clinicians are delivering effective counseling interventions. Unlike many other forms of communication, these counseling interventions are relatively structured and thus easier for an LLM to score based on a single training example. Future work should examine on other types of scoring and vote aggregation methods, such as round robin tournaments, the use of multiple LLM models, or generative self aggregation (GSA) [33].

REFERENCES

[1] H. M. Pandey, "Harnessing Large Language Models for Mental Health: Opportunities, Challenges, and Ethical Considerations," *arXiv.org*, 2025. [Online]. Available: https://arxiv.org/abs/2501.10370

[2] Y. Jin, J. Liu, P. Li, B. Wang, Y. Yan, H. Zhang, C. Ni, J. Wang, Y. Li, Y. Bu, and Y. Wang, "The Applications of Large Language Models in Mental Health: Scoping Review," *Journal of Medical Internet Research*, vol. 27, p. e69284, may 5 2025. [Online]. Available: http://dx.doi.org/10.2196/69284

[3] Y. Hua, F. Liu, K. Yang, Z. Li, H. Na, Y.-h. Sheu, P. Zhou, L. V. Moran, S. Ananiadou, A. Beam, and J. Torous, "Large Language Models in Mental Health Care: a Scoping Review," *arXiv.org*, 2024. [Online]. Available: https://arxiv.org/abs/2401.02984

[4] Y. Hua, H. Na, Z. Li, F. Liu, X. Fang, D. Clifton, and J. Torous, "Applying and Evaluating Large Language Models in Mental Health Care: A Scoping Review of Human-Assessed Generative Tasks," *arXiv.org*, 2024. [Online]. Available: https://arxiv.org/abs/2408.11288

[5] H. Na, Y. Hua, Z. Wang, T. Shen, B. Yu, L. Wang, W. Wang, J. Torous, and L. Chen, "A Survey of Large Language Models in Psychotherapy: Current Landscape and Future Directions," *arXiv.org*, 2025. [Online]. Available: https://arxiv.org/abs/2502.11095

[6] Y. Hua, H. Na, Z. Li, F. Liu, X. Fang, D. Clifton, and J. Torous, "A scoping review of large language models for generative tasks in mental health care," *npj Digital Medicine*, vol. 8, no. 1, apr 30 2025. [Online]. Available: http://dx.doi.org/10.1038/s41746-025-01611-4

[7] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large Language Models for Mental Health Applications: Systematic Review (Preprint)," *arXiv.org*, feb 18 2024. [Online]. Available: http://dx.doi.org/10.2196/preprints.57400

[8] N. C. Chung, G. Dyer, and L. Brocki, "Challenges of Large Language Models for Mental Health Counseling," *arXiv.org*, 2023. [Online]. Available: https://arxiv.org/abs/2311.13857

[9] S. Ji, T. Zhang, K. Yang, S. Ananiadou, and E. Cambria, "Rethinking Large Language Models in Mental Health Applications," *arXiv.org*, 2023. [Online]. Available: https://arxiv.org/abs/2311.11267

[10] H. R. Lawrence, R. A. Schneider, S. B. Rubin, M. J. Matarić, D. J. McDuff, and M. Jones Bell, "The Opportunities and Risks of Large Language Models in Mental Health," *JMIR Mental Health*, vol. 11, pp. e59 479–e59 479, jul 29 2024. [Online]. Available: http://dx.doi.org/10.2196/59479

[11] S. Volkmer, A. Meyer-Lindenberg, and E. Schwarz, "Large language models in psychiatry: Opportunities and challenges," *Psychiatry Research*, vol. 339, p. 116026, 9 2024. [Online]. Available: http://dx.doi.org/10.1016/j.psychres.2024.116026

[12] E. C. Stade, S. W. Stirman, L. H. Ungar, C. L. Boland, H. A. Schwartz, D. B. Yaden, J. Sedoc, R. J. DeRubeis, R. Willer, and J. C. Eichstaedt, "Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation," *npj Mental Health Research*, vol. 3, no. 1, apr 2 2024. [Online]. Available: http://dx.doi.org/10.1038/s44184-024-00056-z

[13] A. Ahmadi, M. Noetel, M. Schellekens, P. Parker, D. Antczak, M. Beauchamp, T. Dicke, C. Diezmann, A. Maeder, N. Ntoumanis, A. Yeung, and C. Lonsdale, "A Systematic Review of Machine Learning for Assessment and Feedback of Treatment Fidelity," *Psychosocial Intervention*, vol. 30, no. 3, pp. 139–153, 7 2021. [Online]. Available: http://dx.doi.org/10.5093/pi2021a4

[14] S. L. Kopelovich, R. M. Brian, M. Tanana, R. Slevin, B. Pace, S. K. Stewart, V. Shepard, D. Ben-Zeev, S. A. Baldwin, C. S. Soma, S. Stanco, and Z. Imel, "Development and validation of a cognitive behavioral therapy for psychosis online training with automated feedback." *Psychotherapy*, vol. 62, no. 1, pp. 1–11, 3 2025. [Online]. Available: http://dx.doi.org/10.1037/pst0000548

[15] L. Pfiffner, M. Rooney, L. Haack, M. Villodas, K. Delucchi, and K. McBurnett, "A randomized controlled trial of a school-implemented school-home intervention for attention-deficit/hyperactivity disorder symptoms and impairment," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 55, no. 9, pp. 762–770, 2016.

[16] J. Geathers, Y. Hicke, C. Chan, N. Rajashekar, J. Sewell, S. Cornes, R. F. Kizilcec, and D. Shung, "Benchmarking Generative AI for Scoring Medical Student Interviews in Objective Structured Clinical Examinations (OSCEs)," *arXiv.org*, 2025. [Online]. Available: https://arxiv.org/abs/2501.13957

[17] A. H. Shakur, M. J. Holcomb, D. Hein, S. Kang, T. O. Dalton, K. K. Campbell, D. J. Scott, and A. R. Jamieson, "Large Language Models for Medical OSCE Assessment: A Novel Approach to Transcript Analysis," *arXiv.org*, 2024. [Online]. Available: https://arxiv.org/abs/2410.12858

[18] G. Lam, Y. Shammoon, A. Coulson, F. Lalloo, A. Maini, A. Amin, C. Brown, and A. H. Sam, "Utility of large language models for creating clinical assessment items," *Medical Teacher*, vol. 47, no. 5, pp. 878–882, aug 26 2024. [Online]. Available: http://dx.doi.org/10.1080/0142159X.2024.2382860

[19] Z. Iftikhar, S. Ransom, A. Xiao, and J. Huang, "Therapy as an NLP Task: Psychologists' Comparison of LLMs and Human Peers in CBT," *arXiv.org*, 2024. [Online]. Available: https://arxiv.org/abs/2409.02244

[20] M. Hardy, ""All that Glitters": Approaches to Evaluations with Unreliable Model and Human Annotations," *arXiv.org*, 2024. [Online]. Available: https://arxiv.org/abs/2411.15634

[21] H. Osterhoudt, C. E. Schneider, H. A. Mohammad, M. Shih, A. E. Harper, L. Zhou, E. R. Skidmore, and Y. Wang, "Automated Fidelity Assessment for Strategy Training in Inpatient Rehabilitation using Natural Language Processing," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2022. [Online]. Available: https://arxiv.org/abs/2209.06727

[22] D. R. Thomas, J. Lin, S. Bhushan, R. Abboud, E. Gatz, S. Gupta, and K. R. Koedinger, "Learning and AI Evaluation of Tutors Responding to Students Engaging in Negative Self-Talk," in *Proceedings of the Eleventh ACM Conference on Learning @ Scale*. ACM, jul 9 2024, pp. 481–485. [Online]. Available: http://dx.doi.org/10.1145/3657604.3664700

[23] L. Ryan, H. O. Ifdita, P. P. Juan, S. S. Raj, B. E., and Y. Diyi, "Can LLM-Simulated Practice and Feedback Upskill Human Counselors? A Randomized Study with 90+ Novice Counselors," *arXiv*, 2025.

[24] J. Kurland, V. Varadharaju, A. Liu, P. Stokes, A. Gupta, M. Hudspeth, and B. O'Connor, "Large Language Models' Ability to Assess Main Concepts in Story Retelling: A Proof-of-Concept Comparison of Human Versus Machine Ratings," *American Journal of Speech-Language Pathology*, pp. 1–11, mar 31 2025. [Online]. Available: http://dx.doi.org/10.1044/2025_AJSLP-24-00400

[25] N. Flemotomos, V. R. Martinez, Z. Chen, T. A. Creed, D. C. Atkins, and S. Narayanan, "Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations," *PLOS ONE*, vol. 16, no. 10, p. e0258639, oct 22 2021. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0258639

[26] A. Pilny, K. McAninch, A. Slone, and K. Moore, "From manual to machine: assessing the efficacy of large language models in content analysis," *Communication Research Reports*, vol. 41, no. 2, pp. 61–70, mar 12 2024. [Online]. Available: http://dx.doi.org/10.1080/08824096.2024.2327547

[27] C.-H. Chiang and H.-y. Lee, "Can Large Language Models Be an Alternative to Human Evaluations?" *Annual Meeting of the Association for Computational Linguistics*, 2023. [Online]. Available: https://arxiv.org/abs/2305.01937

[28] R. H. Tai, L. R. Bentley, X. Xia, J. M. Sitt, S. C. Fankhauser, A. M. Chicas-Mosier, and B. G. Monteith, "An examination of the use of large language models to aid analysis of textual data," *International Journal of Qualitative Methods*, vol. 23, p. 16094069241231168, 2024.

[29] N. Milano, M. Ponticorvo, and D. Marocco, "Comparing Human Expertise and Large Language Models Embeddings in Content Validity Assessment of Personality Tests," *arXiv.org*, 2025. [Online]. Available: https://arxiv.org/abs/2503.12080

[30] H. Mahmoudi, D. Chang, H. Lee, N. Ghaffarzadegan, and M. S. Jalali, "A Critical Assessment of Large Language Models for Systematic Reviews: Utilizing ChatGPT for Complex Data Extraction," *SSRN Electronic Journal*, 2024. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4797024

[31] K. Thomas, P. G. Kelley, D. Tao, S. Meiklejohn, O. Vallis, S. Tan, B. Bratanič, F. T. Ferreira, V. K. Eranti, and E. Bursztein, "Supporting Human Raters with the Detection of Harmful Content using Large Language Models," *arXiv.org*, 2024. [Online]. Available: https://arxiv.org/abs/2406.12800

[32] Y. Y. Chiu, A. Sharma, I. W. Lin, and T. Althoff, "A Computational Framework for Behavioral Assessment of LLM Therapists," *arXiv.org*, 2024. [Online]. Available: https://arxiv.org/abs/2401.00820

[33] Z. Li, X. Feng, Y. Cai, Z. Zhang, T. Liu, C. Liang, W. Chen, H. Wang, and T. Zhao, "Llms can generate a better answer by aggregating their own responses," 2025. [Online]. Available: https://arxiv.org/abs/2503.04104

VIII. Appendix

A. System Prompt

```
You are a psychology researcher who is an expert in behavioral parent training techniques.
You are reviewing transcripts of recorded meetings between a team of researchers (called "trainers")
and school social worker (called "leaders").
The trainers are training the leaders to deliver a behavioral parent training intervention to parents
at their schools.
This intervention involves parents learning new skills and techniques for parenting.
The trainers are using a script to teach the leaders how to deliver these skills to the parents.
They are following the content in the script, however they can also use their own words to convey
the concepts.
The text below following the word TRANSCRIPT is the transcript of the meeting.
The user prompt contains a label for the content the leaders were supposed to
cover (e.g., "Program Introductions" and the script that the trainers were supposed to follow to
cover that content.

**Your final response MUST be a JSON object.**
**The JSON object should have the following structure and keys:**
```json
{
  "summary_of_script_key_points": "Summarize the key points from the script here, paying attention
  to who was speaking.",
  "content_coverage": "Provide details on whether each key point was covered",
  "answers_to_questions": {
    "Q1": {
      "label": "did trainers cover content?",
      "rating": "yes/no",
      "reasoning": "state reason of your rating here",
      "quote_from_script": "providing direct quotes from the transcript",

    },

    "Q2":{
      "label": "did trainers demonstrate outline presentation?"
      "rating": "yes/no",
      "reasoning": "state reason of your rating here",
      "quote_from_script": "providing direct quotes from the transcript",

    },
    "Q3": {
      "label": "did trainers use reference video?",
      "rating": "yes/no",
      "reasoning": "state reason of your rating here",
      "quote_from_script": "providing direct quotes from the transcript",

    },
    "Q4": {
      "label": "did leaders engage in role play?",
      "rating": "yes/no",
      "reasoning": "state reason of your rating here",
      "quote_from_script": "providing direct quotes from the transcript",

    },
  },

}

For "reasoning" key, "Explain your reasoning for the answers, providing direct quotes from
the transcript. If any answer is 'no', provide examples of missed content."

To answer these question, do the following steps:

1.Summarize the key points in the script in the user prompt. Pay attention to who was speaking.
2.For each key point, check to see if the content was covered in the transcript
3.Answer each question above based on the comparison. If most of the content was covered, answer "yes." I
```
```

```
Before giving your final rating, explain your reasoning for giving the rating.
Provide direct quotes from transcript to support your rating.
Do not provide quotes from the script in the user prompt.
If answering "no" to any of the questions, provide examples of content that was missed.
Be concise with your response.

TRANSCRIPT
```

B. User Prompts: Home Activities Check-in

```
# Content Label
Home Activities Check-in

# Script
TELL PARENTS
* We're going to start group by checking in
about how it went at home this past week.
* We'll talk about the skills we learned and
the Classroom Challenge.
* I'll be taking notes about what your kids
have done well at home this week.
* I give them this feedback during student
groups. They love hearing positive things
their parents mention!
* Please pull up your child's Progress
Tracking Sheet & the first handout.

TELL PARENTS
* Who would like to start? Please share one
behavior you praised this week.
(When one parent shares, engage the
other parents.)
* Examples:
- Did anyone else have a similar
experience this week?
- Does anyone have a suggestion for
this problem?
(Highlight teamwork & tie to CLS Values.)
* Example:
- It's great to see on the Tracking Sheet
that you shared the rewards they
earned last week. That's a great
example of our team values, \working
together" and \sharing information".

TELL PARENTS
* Continue praising this week.
* Remember that getting praise is
especially important for kids with
attention and behavior issues because
they often get more corrections.
* And research shows that across cultures
and backgrounds, praise is helpful to
improve parent-child relationships and kids'
self-esteem even when they're teenagers.
* Remember to look out for behaviors you
sometimes but want to see more of.
* Praise should happen right away and be
specific.
```

IX. Acknowledgments

This research was supported by NIH grant P50MH126231 and the IN STEP Children's' Mental Health Research Center funded by the National Institute of Health [5P50MH126231]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.