Towards AI-Assisted Psychotherapy: Emotion-Guided Generative Interventions

Kilichbek Haydarov[1], Youssef Mohamed [1], Emilio Goldenhersch[2], Paul O'Callaghan[1], Li-jia Li[3], Mohamed Elhoseiny[1]

[1]King Abdullah University of Science and Technology,
[2]Laboratorio de Investigación en Neurociencia y Ciencias Sociales, Universidad de Flores,
[3]HealthUnity

Correspondence: kilichbek.haydarov@kaust.edu.sa

Abstract

Large language models (LLMs) hold promise for therapeutic interventions, yet most existing datasets rely solely on text, overlooking non-verbal emotional cues essential to real-world therapy. To address this, we introduce a multimodal dataset of 1,441 publicly sourced therapy session videos containing both dialogue and non-verbal signals such as facial expressions and vocal tone. Inspired by Hochschild's concept of emotional labor, we propose a computational formulation of *emotional dissonance*—the mismatch between facial and vocal emotion—and use it to guide emotionally aware prompting. Our experiments show that integrating multimodal cues, especially dissonance, improves the quality of generated interventions. We also find that LLM-based evaluators misalign with expert assessments in this domain, highlighting the need for human-centered evaluation. Project page link: `https://kilichbek.github.io/webpage/mental/`

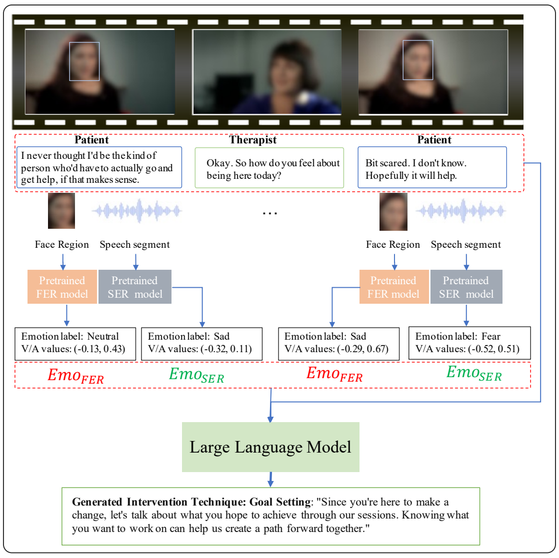


Figure 1: Intervention generation pipeline. An LLM uses dialogue context to generate interventions, while non-verbal cues (EmoSER from speech and EmoFER from video) enhance their plausibility and relevance. FER: Facial Emotion Recognition; SER: Speech Emotion Recognition. Images used with permission from (Johnson) and blurred due to privacy concerns.

1 Introduction

The integration of AI into mental health therapy represents a transformative opportunity in psychiatric care, enabling novel approaches to address the multifaceted needs of individuals with mental health conditions (Minerva and Giubilini, 2023). In recent years, advancements in Natural Language Processing (NLP) — and in particular the rapid development of LLMs — have revolutionized our ability to understand and generate natural language. These models exhibit remarkable in-context learning capabilities and strong zero-shot performance across diverse tasks, significantly expanding the scope of AI applications in mental health care (van Heerden et al., 2023; Ji et al., 2023; Hua et al., 2024). By efficiently retrieving and synthesizing information from sources such as electronic health records, mobile interactions, and social media, LLMs provide clinicians with valuable insights into patient behaviors and experiences, facilitating early intervention strategies and personalized treatment plans. Their conversational abilities further offer individuals an accessible and empathetic platform to express their emotional states and personal experiences.

Despite these promising developments, most existing research on LLMs in mental health has focused solely on textual interactions, thereby neglecting the rich non-verbal cues that are critical in real-world therapeutic settings. While prior studies have applied language models to detect suicidal ideation (Allen et al., 2019; Cao et al., 2020; Ji et al., 2022), diagnose depression (Morales et al., 2021; Cheng and Chen, 2022; Park and Moon, 2022; Ansari et al., 2022; Naseem et al., 2022), and deliver therapy via chatbots (Nicol et al., 2022), the

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32736–32755
November 4-9, 2025

generation of therapeutic intervention techniques that integrate both verbal and non-verbal signals remains underexplored.

Motivated by this gap, our work explores the potential of incorporating emotional signals from multiple modalities—such as facial expressions and speech—to enhance the effectiveness of LLM-generated interventions. To facilitate this investigation, we have assembled a diverse corpus of 1441 video recordings of mental health sessions sourced from publicly available platforms. These recordings capture not only the dialogue between therapists and patients but also the non-verbal cues that convey emotional context, offering a holistic view of the therapeutic interaction.

Central to our approach is the enrichment of LLM prompts with predicted emotional cues derived from both speech and facial expressions. This multimodal augmentation enables the models to generate responses that are more contextually aware and emotionally attuned. Moreover, we introduce a novel concept we term *emotional dissonance*, which quantifies the divergence between the affective tone embedded in the verbal content and that conveyed through non-verbal modalities. By capturing this nuanced discrepancy, our approach provides a deeper insight into complex emotional dynamics, thereby enriching the contextual input for LLMs and elevating the quality and relevance of the generated therapeutic interventions.

Our contributions can be summarized as follows:

- We collect and analyze a novel multimodal dataset comprising 1441 video recordings of mental health sessions, which serves as a foundation for exploring generative intervention techniques.
- We conduct a comprehensive evaluation of the zero-shot generative capabilities of both open-source and proprietary LLMs, focusing on their ability to generate contextually relevant therapeutic interventions.
- We investigate the benefits of augmenting dialogue contexts with non-verbal emotional cues and introduce the novel concept of *emotional dissonance* to further refine model performance.

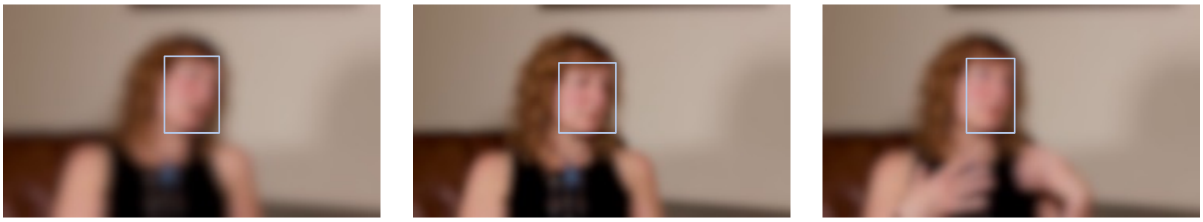
By integrating emotional insights from multiple modalities into LLM-based interventions, our study aims to advance the development of more empathetic and effective AI-assisted mental health care systems.

2 Related Work

LLMs in Mental Health: Research on leveraging Large Language Models (LLMs) for mental health applications is still emerging, with a limited yet growing body of work addressing tasks such as stress, depression, and suicide detection (Zhou et al., 2023; Bao et al., 2023; Hayati et al., 2022). Among the promising approaches are few-shot prompting—which supplies task demonstrations to models prior to execution—and chain-of-thought (CoT) prompting (Wei et al., 2022) that guides the reasoning process. Building on these methods, techniques such as chain-of-empathy prompting (Lee et al., 2023) integrate psychotherapeutic insights to encourage cognitive reasoning about human emotion, while diagnosis-of-thought prompting (Chen et al., 2023) employs multi-stage diagnostic processes to further enhance model performance in mental health contexts.

Another research direction fine-tunes general LLMs with domain-specific texts for mental health, e.g., MentaLLaMA (Yang et al., 2023, 2024), Mental-LLM (Yang et al., 2024) (social media data), and ChatCounselor (Liu et al., 2023a) (real client–psychologist interactions). Other works include (Moon and Bhattacharyya, 2024), which generates CBT-based responses, and (Singh et al., 2024), which applies multimodal LLMs for cognitive distortion detection.

Mental-Health Datasets: Despite a scarcity of comprehensive mental health datasets, several efforts are underway to curate data that provide critical insights into psychological well-being. The HOPE (Malhotra et al., 2022) dataset, for instance, offers counseling transcripts annotated with dialogue acts, while MEMO (Srivastava et al., 2022) extends HOPE by incorporating summaries aimed at dialogue summarization in therapeutic contexts. The ALONE dataset (Wijesiriwardene et al., 2020) contributes a multimodal perspective by capturing emojis and images from toxic social media interactions among high school students, shedding light on digital mental health challenges. Additional resources include datasets from counselchat.com (Bertagnolli, 2020), the PAIR dataset (Pérez-Rosas et al., 2022) which documents varied reflective listening skills, and Reddit-based



Speaker	Start Time	End Time	Utterance
Therapist	00:01:09	00:01:12	What would you rate your mood right now?
Patient	00:01:17	00:01:19	Okay, so a two
Therapist	00:01:20	00:01:23	And what's going on right now to make it a two?
Patient	00:01:29	00:01:41	I was able to get up today. Yesterday I couldn't get up. I got here
Patient	00:01:44	00:01:47	And I'm hoping this will help
Patient	00:01:51	00:02:03	I am just I feel I can't everything around me seems dark.

Figure 2: a representative example from our dataset, showing transcripts, speaker tags, timestamps, and bounding boxes. Image used with permission from Psychotherapy.net (psy) and blurred due to privacy concerns.

datasets (Gupta et al., 2022; Roy et al., 2022) that label posts related to suicidality and depression. However, many of these datasets are limited in size and predominantly unimodal, highlighting the need for richer multimodal resources.

Emotion Representation and Recognition: Emotion representation research typically follows two frameworks: the discrete categorical system (Ekman, 1992) and the continuous two-dimensional Valence-Arousal (VA) model (Russell, 1980; Gunes et al., 2011). The discrete approach classifies emotions into a fixed set of innate states (e.g., joy, sorrow), whereas the VA model conceptualizes emotions along continuous dimensions—valence (the pleasantness of the emotion) and arousal (its intensity) (Hamann, 2012). Extensive work has been conducted across various modalities, including Facial Expression Recognition (FER) (De Silva et al., 1997; Ko, 2018; Khaireddin and Chen, 2021; Hu et al., 2023), audio-based emotion recognition (Ooi et al., 2014), and text-based sentiment analysis (Alswaidan and Menai, 2020; Batbaatar et al., 2019). Recent studies have further explored multimodal approaches (Yoon et al., 2018; Cheng and Chen, 2022), contextual influences (Hazarika et al., 2018; Wang et al., 2020; Poria et al., 2019), and interpretability of emotional signals (Achlioptas et al., 2021; Mohamed et al., 2022; Haydarov et al., 2023).

Evaluation of LLMs: Evaluating LLMs for generating therapeutic interventions is challenging due to the inherently subjective and nuanced nature of therapy sessions, which demand assessments of relevance, coherence, and emotional sensitivity. Traditional metrics like BERTScore (Giorgi et al., 2023) and MAUVE (Pillutla et al., 2021) rely on comparing generated texts to fixed references—an approach that is often inadequate in contexts where a definitive reference is lacking. Recent proposals advocate for evaluation metrics rooted in human communication psychology (Giorgi et al., 2023) and tailored frameworks, such as the ChatCounselor evaluation method (Liu et al., 2023a) that leverages GPT-4 with domain-specific questions. These advancements underscore the need for more sophisticated, context-aware evaluation strategies in therapeutic dialogue generation.

3 Data Collection and Analysis

In this section, we describe our data collection process, compare our dataset to existing mental health resources, and introduce our approach to capturing *emotional dissonance* for a nuanced understanding of emotional dynamics.

# of Videos	1441
Duration (hours)	163.57
Avg/Median Duration (minutes)	6.8 / 2.9
Avg # of turns	2.8
Vocab Size	29005
# of Speakers	2

Table 1: Statistics of the proposed dataset.

Data Collection: A major challenge in assembling mental health datasets is the scarcity of pub-

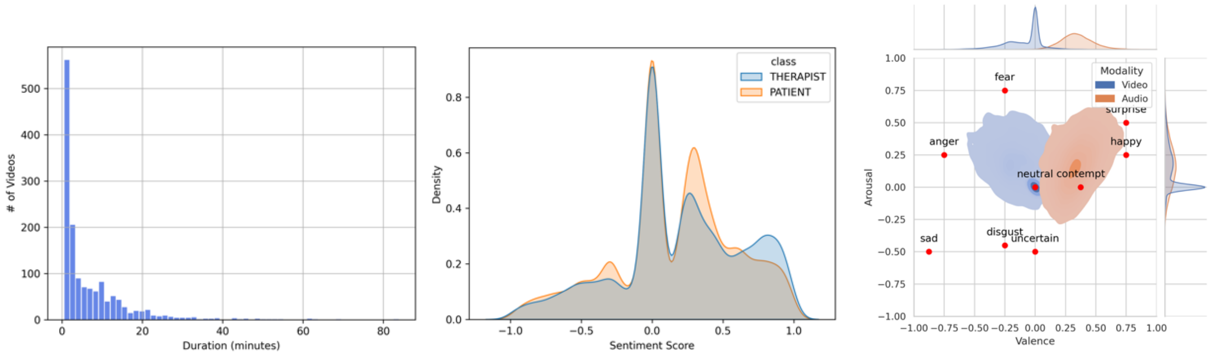


Figure 3: Data Analysis: **left:** Histogram of duration. **middle:** distribution of sentiment scores for therapist and patient utterances. **rigth:** Distribution of valence and arousal estimations from audio and video modalities in two-dimensional space.

Dataset	**Source**	**Size**	**Text**	**Audio**	**Video**
HOPE (Malhotra et al., 2022)	YouTube	212	✓	✗	✗
MEMO (Srivastava et al., 2022)	YouTube	212	✓	✗	✗
DAIC-WoZ (Gratch et al., 2014)	Recorded	612	✓	✓	✓
ALONE (Wijesiriwardene et al., 2020)	Twitter	688	✓	✗	✓
Counsel-CHAT (Bertagnolli, 2020)	Counselchat.com	2780	✓	✗	✗
PAIR (Pérez-Rosas et al., 2022)	Crowd-sourced	4683	✓	✗	✗
Ours	YouTube	1441	✓	✓	✓

Table 2: Comparison to other mental-health-related datasets. The presence of text, audio, and video modalities is marked with ✓, while absence is marked with ✗.

licly available counseling sessions due to privacy concerns. To overcome this, we crawled YouTube for high-quality role-played therapy sessions. Our data collection was guided by mental health experts who helped filter the videos to ensure they depict clinically relevant therapist-patient interactions. These sessions range from standard counseling to demonstrations of therapies such as Cognitive Behavioral Therapy (CBT) and Dialectical Behavior Therapy (DBT). In total, our curated collection consists of 1441 videos (See Appendix Appendix B). Figure 2 presents a representative example from our dataset, showcasing transcripts, speaker tags, timestamps, and face bounding boxes.

Dataset Statistics: Table 1 summarizes key statistics of our dataset, which encompasses 1441 videos with a total duration of 163.57 hours. The average and median video durations are 6.8 minutes and 2.9 minutes, respectively, with an average of 2.8 exchanges per session. The dataset is linguistically rich, containing 29,005 unique words, and every video features two speakers, reflecting the typical dyadic format of therapy sessions. Table 2 compares our dataset to other mental health-related datasets. While resources like HOPE (Malhotra et al., 2022) and MEMO (Srivastava et al., 2022) are limited to transcripts, and datasets such as DAIC-WoZ (Gratch et al., 2014) and ALONE (Wijesiriwardene et al., 2020) include additional modalities but are smaller in scale, our dataset uniquely offers a multimodal perspective by providing synchronized text, audio, and video.

Affective Analysis: To assess the emotional content within the sessions, we performed an affective analysis using both verbal and non-verbal cues. We first employed the Valence, Arousal, and Dominance (VAD) lexicon (Hutto and Gilbert, 2014) to evaluate sentiment in therapist and patient utterances. Our analysis indicates that patient utterances generally exhibit slightly negative sentiment, consistent with the nature of their concerns, while therapist utterances tend to be more positive. Furthermore, we extract valence-arousal pairs from both audio and video modalities. In Figure 3 (right), audio data often conveys higher levels of positive emotion, whereas video data tends to capture more pronounced negative emotional signals. This divergence underscores the importance of integrating multimodal cues to fully understand the emotional landscape during therapy sessions.

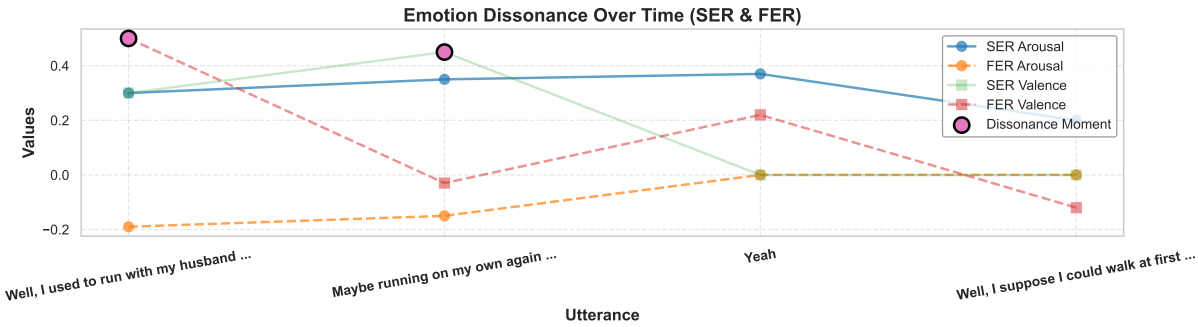


Figure 4: Illustration of an emotional dissonance moment, characterized by significant discrepancies in arousal levels between audio and video modalities. The segment corresponding to the first two utterances is highlighted to indicate potential emotional dissonance.

3.1 Emotional Dissonance

Inspired by the psychological concept of emotional labor (Hochschild, 1983)—which describes how individuals manage their emotional expressions to meet social or professional expectations — we propose a computational framework that adapts this concept. In our context, *emotional dissonance* refers to a measurable discrepancy between the emotions verbally expressed by a patient and those inferred from non-verbal cues such as facial expressions and vocal tone. While we do not directly access internal emotional states, this cross-modal mismatch provides a useful proxy for studying external emotional inconsistency in therapy dialogues. To our knowledge, this is the first work to model and inject such cross-modal dissonance cues into language model prompting for therapeutic interventions. Unlike emotional labor, which emphasizes the deliberate regulation of feelings, emotional dissonance captures instances where these channels diverge; for example, a significant difference in the estimated valence and arousal levels derived from speech (audio) and facial expressions (video) may indicate that the emotional intensity conveyed vocally does not align with that displayed visually, providing valuable insights into whether a patient is fully comfortable or might be subtly withholding their true emotional state. Overall, we present emotional dissonance as an exploratory tool to enhance our understanding of the complex interplay of emotional signals in therapeutic interactions. We formally define emotional dissonance as follows:

$$\text{Dissonance} = \mathbb{I}\left\{ \max\Big(|V_{\text{SER}} - V_{\text{FER}}|\,,\, |A_{\text{SER}} - A_{\text{FER}}|\Big) > T \right\} \quad (1)$$

Here, V_{SER} and V_{FER} denote the valence values extracted from audio and video segments respectively, and A_{SER} and A_{FER} represent the corresponding arousal values. The threshold T is empirically determined to capture significant discrepancies. Figure 4 illustrates a dissonance moment, where a substantial divergence between audio- and video-derived arousal levels is observed. Detecting such moments is crucial, as they may signal complex emotional states that warrant specialized intervention techniques, thereby enriching the context for LLM-based responses. We set T to 0.5 in our studies (see Appendix C for further details).

4 Method

Problem Formulation: Our goal is to enhance the capabilities of Large Language Models (LLMs) for generating contextually appropriate therapeutic interventions. Specifically, given an input context C (e.g., a segment of a patient-therapist interaction), the model should output both an intervention technique label T and a corresponding response R that effectively continues the therapeutic dialogue. This enables the model to zero-shot generate tailored interventions from nuanced conversational and emotional cues.

4.1 Prompt Design

We leverage the powerful in-context learning ability of LLMs (Brown et al., 2020) to generate therapeutic interventions without explicit fine-tuning. To this end, we systematically explore three distinct prompting strategies that progressively incorporate richer emotional information:

Base Prompting: In our baseline approach, the LLM is provided solely with the dialogue context and instructed to generate an intervention. This method relies exclusively on textual information, serving as a reference point for assessing the added value of emotional cues.

Emotion-Enhanced Prompting: Recognizing that emotional context is critical for effective therapy, we augment the prompt with emotion predictions. These predictions are provided in two formats. 1) Categorical labels (e.g., *happy*, *sad*, *angry*) that summarize the affective state; 2) Continuous Emotion Representations: fine-grained valence-arousal estimations that capture both the intensity and quality of emotions. This additional emotional information enables the LLM to generate responses that are more sensitive to the underlying affective dynamics of the interaction.

Dissonance-Based Prompting: To further refine the model's sensitivity to complex emotional signals, we introduce a strategy that highlights instances of *emotional dissonance*—a condition where there is a notable discrepancy between the emotional tone of the verbal content and that inferred from non-verbal cues. When such dissonance is detected, we explicitly prepend the relevant patient utterances with the phrase *"emotional dissonance"*. This deliberate cue is designed to draw the model's attention to potential mismatches between expressed and perceived emotions, thereby encouraging the generation of more nuanced and context-aware interventions (refer to Appendix A for more details).

Figure 1 shows the pipeline of our method.

5 Experiments and Results

Models: We evaluated four high-performance, widely accessible LLMs for generating therapeutic interventions: GPT-4o (OpenAI, 2023), LLaMa3-8b-instruct (AI@Meta, 2024), Mistral3-7b-instruct (Jiang et al., 2023), and Gemini. We also tested Mental-LLaMA-Chat-7B (Yang et al., 2024), but despite being instruction-tuned, it struggled with our structured prompting and often failed to produce coherent, context-sensitive interventions. This suggests that fine-tuning on unrelated classification tasks may hinder generalization to generative settings, so we excluded such models from our study.

Evaluation Framework: We develop a systematic evaluation framework based on three key criteria:

Applicability: measures both **relevance** and **practicality** of the model-generated responses in addressing specific therapeutic scenarios. Highly applicable outputs align well with the needs of individuals seeking mental health support.

Empathy: evaluates the extent to which the responses convey **understanding** and **emotional resonance**. In therapeutic settings, it is essential that responses not only address clinical concerns but also reflect an empathetic tone that resonates with the individual's emotional state.

Comprehensiveness: evaluates the **breadth** and **depth** of the response in addressing the patient's concern. A comprehensive response goes beyond surface-level acknowledgment to incorporate multiple facets of the issue or offer a more complete therapeutic perspective.

Expert Evaluation: Five psychologists evaluated responses generated for 100 videos using a 1-to-5 Likert scale (1 = poor, 5 = exceptional). We utilized the same 3 aforementioned criteria.

5.1 Results and Observations

To compare prompting techniques to the base prompt, we computed the average improvement in empathy ratings for each model. Improvement was defined as the difference between the mean rating obtained using an alternative technique (continuous, discrete, or dissonance) and that obtained using the base prompt. All models showed improvements in empathy, with GPT-4o gaining up to 0.138 points and Gemini improving by 0.054 (continuous) and 0.125 (dissonance). Similar trends were observed for LLaMa3 and Mistral (Figure 5).

A one-way analysis of variance (ANOVA) on empathy ratings revealed a significant main effect of condition, $F(3, 6204) = 72.18,\ p < .001$. Subsequent Tukey's HSD post-hoc comparisons showed that the base condition differed significantly from all alternative prompting techniques: for the continuous condition, the mean difference (Δ) was 0.391 (95% CI: [0.300, 0.482], $p < .001$); for the discrete condition, $\Delta = 0.437$ (95% CI: [0.346, 0.528], $p < .001$); and for the dissonance condition, $\Delta = 0.445$ (95% CI: [0.354, 0.537], $p < .001$). No significant differences were found among the alternative techniques themselves. Figure 7 presents qualitative results, with additional findings in the Appendix. In contrast, no improvements were observed for the comprehensiveness and applicability criteria. A possible explanation is that applicability ratings may be influenced by the evaluators' therapeutic orientations—for instance, experts with a humanistic orientation might assess applicability differently from those with a cognitive-based approach. Additionally, the comprehensiveness criterion, reflecting the breadth of

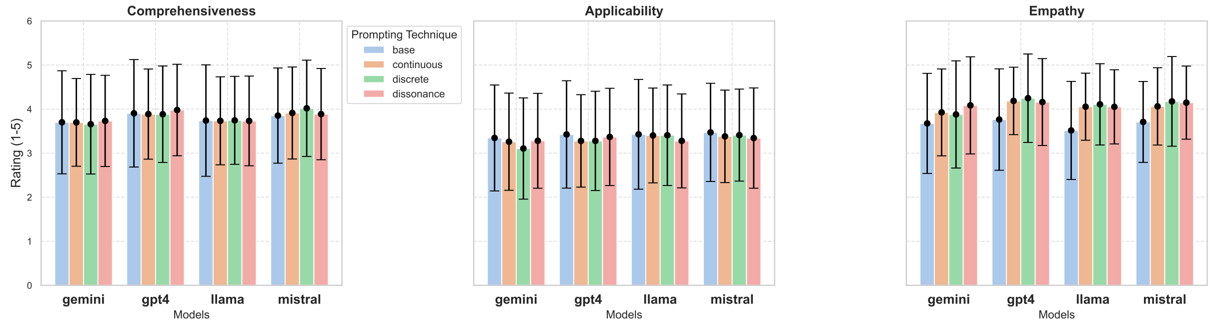


Figure 5: Results of the expert evaluation based on three criteria. The emotionally enriched prompts significantly improve performance, particularly in Empathy axis.

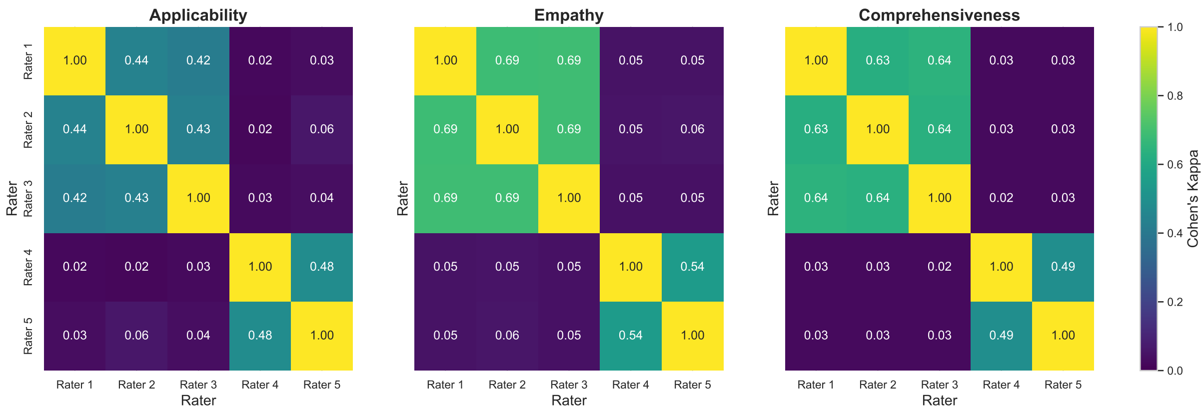


Figure 6: Pairwise Cohen's κ analysis of inter-rater agreement. Raters 1–3 show substantial agreement for empathy and comprehensiveness ($\kappa > 0.64$), while their agreement with Raters 4 and 5 is very low ($\kappa < 0.10$), suggesting that differences in expertise and therapeutic orientation influence evaluation outcomes.

issues addressed by the intervention, appears inherently less sensitive to prompting variations; our results indicate that all models perform similarly in covering the overall problem domain.

How consistent are the evaluators in their assessments across different criteria? To evaluate inter-rater reliability, we conducted pairwise Cohen's kappa κ analysis across the three evaluation criteria. We observed substantial agreement for both empathy and comprehensiveness ($\kappa > 0.64$), and moderate agreement for applicability ($\kappa < 0.60$), suggesting that applicability may involve more subjective interpretation based on clinical perspective (see Figure 6). To further understand this variance, we grouped raters based on their professional backgrounds. Group 1 (Raters 1, 2, and 3) had an average of 5 years of experience and held Master's degrees, while Group 2 (Raters 4 and 5) had over 15 years of experience, PhD degrees, and a Cognitive Behavioral Therapy (CBT) orientation. Interestingly, Group 2 consistently assigned more conservative scores (3.20±0.97) compared to Group 1 (3.88 ± 1.11). This divergence reflects different evaluative styles shaped by training and orientation rather than disagreement on model quality. These findings offer valuable insight into how expert interpretation can vary across backgrounds, highlighting the importance of incorporating diverse clinical perspectives in future evaluation protocols.

LLM-Based Evaluation. To explore whether LLM-based evaluators can serve as reliable proxies for expert judgment in therapeutic response evaluation, we tasked GPT-4o with rating generated outputs on a 1-to-5 scale across three criteria, following the G-Eval framework (Liu et al., 2023b). We then computed Pearson, Spearman, and Kendall-Tau correlations between GPT-4o's scores and those provided by licensed mental health professionals. Across all criteria, correlations were consistently low, indicating a substantial misalignment between automated and expert evaluation (see Table 3). Notably, GPT-4o tended to assign higher scores than human experts, suggesting a bias toward overly generous assessment. This finding contrasts with prior work in open-domain tasks such as summarization and QA, where LLM-based

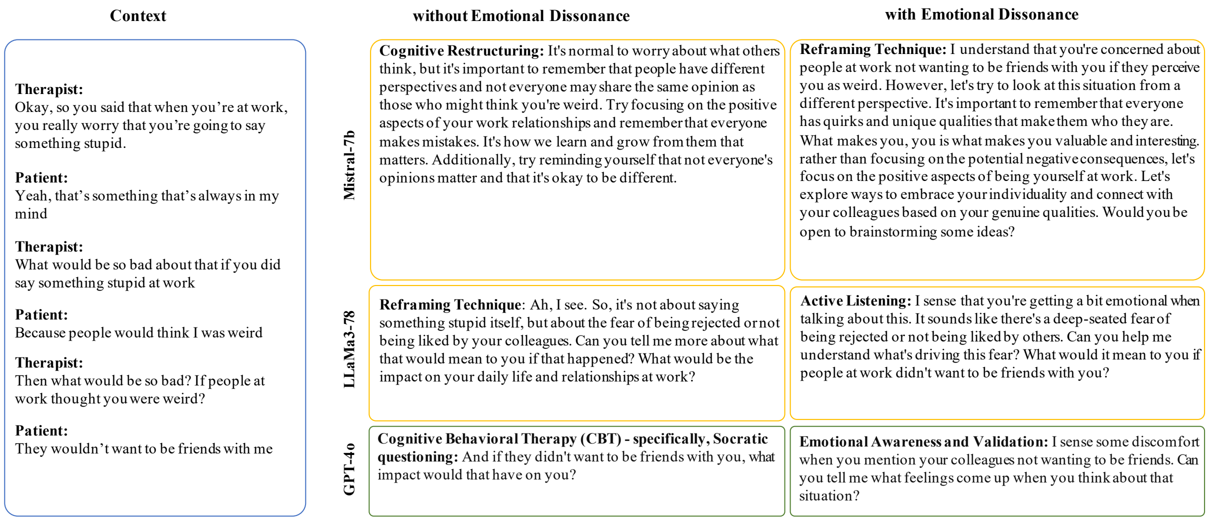


Figure 7: Qualitative results showcasing generations from various models for a given context.

evaluation has shown strong alignment with human ratings (Liu et al., 2023b). Our results highlight a key limitation: while automated evaluators offer scalability, they may be ill-suited for emotionally nuanced, safety-critical domains like mental health. Hence, we advocate against blindly applying LLM-based evaluation in mental health related tasks.

Criteria	r	ρ	τ
Applicability	0.11	0.12	0.10
Empathy	0.16	0.17	0.14
Comprehensiveness	0.19	0.15	0.12

Table 3: Correlation between expert and GPT-4o ratings. Reported are Pearson (r), Spearman (ρ), and Kendall-Tau (τ) coefficients across evaluation criteria.

6 Discussion and Conclusion

Our findings shed light on how emotional cues can significantly enhance the empathy of LLM-generated therapeutic responses. This supports the broader hypothesis that integrating cross-modal affective information can enrich LLM behavior in high-stakes interpersonal domains like psychotherapy. Despite this, improvements were largely confined to the empathy dimension. Applicability and comprehensiveness showed negligible change across prompting strategies. This divergence highlights a key insight: *empathy is more sensitive to affective signals*, whereas applicability and comprehensiveness may rely more heavily on contextual relevance, therapeutic framing, or task-specific reasoning beyond emotional content.

Our inter-rater analysis further emphasizes the importance of evaluator identity. The striking contrast between different groups not only reduced inter-group agreement but also revealed systematic biases in rating severity. This finding underscores a broader challenge in the evaluation of therapeutic dialogue generation: *"ground truth" is not monolithic*. Expert judgment is inherently filtered through years of training, orientation, and epistemology. This suggests that future evaluations could benefit from structured rater calibration or stratified analysis to control for such factors.

Finally, our exploration of LLM-based evaluation reveals an important caution: while convenient, these automatic evaluations diverged significantly from human expert judgment, especially in the applicability criterion. This reaffirms that *automated evaluation is not yet mature enough to replace human assessment in sensitive domains like therapy*. Avenues for improvement may include domain-adapted evaluators, alignment via preference modeling, or hybrid expert-in-the-loop setups.

In this study, we introduced a novel multimodal dataset to enhance mental health interventions by integrating non-verbal cues into LLMs. Our experiments show that emotionally enriched prompts significantly improve empathy in generated responses, suggesting that incorporating diverse emotional signals yields more contextually relevant interventions. Addressing the remaining challenges is essential to ensure AI systems capture the complexity of human emotion while upholding ethical standards. This work paves the way for more personalized, effective, and compassionate AI-assisted therapy.

7 Limitations, Ethical, and Privacy Considerations

In this section, we discuss key limitations of our study and address ethical and privacy concerns related to the integration of AI in mental health applications.

7.1 Study Limitations

Our study has some limitations:

- **LLM Hallucinations:** We observed that LLMs sometimes generate hallucinated responses. Future work could explore Retrieval Augmented Generation (RAG) techniques — potentially leveraging patients' dialogue history — to mitigate this issue. While this might be promising direction to explore, using scientific medical papers as a datastore has been studied in (Alonso et al., 2024), and observed the existence of the context led to non-significant little gains while harming the performance with finetuned models.

- **Reliance on Expert Judgment:** Our evaluation is heavily based on expert assessments. Automatic evaluation methods have not yet aligned well with expert judgment, underscoring the need for more robust, automated metrics.

Although our dataset consists of mock therapy sessions designed to emulate real-world cases, they may not fully capture the complexities and nuances of actual therapy. However, a survey of mental health experts indicated that 90% of the videos were deemed to resemble real-world therapeutic scenarios, lending support to the validity of our dataset as a meaningful proxy for clinical interactions.

7.2 Data Privacy and Consent

Our dataset consists of publicly available YouTube videos of therapy-related conversations. We do not redistribute or modify the videos; rather, we provide only the URLs, ensuring compliance with YouTube's licensing policies and respecting content ownership. Nonetheless, we recognize that public sharing does not necessarily imply consent for use in AI-driven research, particularly in sensitive domains like psychotherapy.

7.3 Dataset Demographics and Bias

The reliance on mock therapy sessions raises concerns about representativeness. Since the videos are primarily in English and feature Western therapeutic frameworks, our findings may not generalize across different cultures, languages, or mental health paradigms. Additionally, the predominance of dyadic interactions (therapist and client) does not capture the complexity of group, family, or crisis interventions. We acknowledge potential cultural biases in interpreting non-verbal cues. Future studies should aim to expand datasets to include diverse therapeutic approaches and multilingual interactions.

Nevertheless, the use of simulated therapy sessions is well-supported in the literature. Prior work in psychotherapy training (Melluish et al., 2007; Sholomskas et al., 2005) has used standardized role-played sessions to model therapeutic interactions. These approaches are especially useful in domains where real clinical data is difficult to access for ethical or privacy reasons. We position our dataset as a practical and ethical starting point for advancing research in AI-assisted psychotherapy, and we plan to incorporate real-world, clinician-reviewed interactions in future iterations of this work.

7.4 Risks of AI in Mental Health Applications

While LLMs can generate empathetic therapeutic interventions, we believe they should *not replace* trained professionals. Key risks include:

- **Over-Reliance on AI:** Users might misinterpret AI-generated responses as clinical advice, potentially leading to harmful self-diagnosis or self-treatment.

- **Misinformation and Hallucinations:** LLMs may produce plausible yet incorrect or harmful advice, especially in high-stakes mental health contexts.

- **Emotional Impact:** Misinterpretation of emotional cues could result in interventions that exacerbate distress rather than alleviate it.

To mitigate these risks, our system is intended solely for research purposes, with safeguards such as human-in-the-loop evaluations. Future work should also focus on enhancing explainability, bias detection, and adversarial robustness.

7.5 Transparency and Responsible AI Use

We are committed to responsible AI practices, which include:

- **Dataset Transparency:** Detailed documentation of the dataset's scope, limitations, and processing steps is provided.
- **Reproducibility:** Although dataset access is limited by YouTube licensing, we describe our data filtering and annotation pipeline to facilitate replication.
- **Human Oversight:** Our evaluations rely on expert human ratings, and we advocate for hybrid AI-human approaches in therapeutic applications.

In summary, while our work demonstrates the potential benefits of integrating multimodal cues into AI-assisted psychotherapy, ethical considerations and inherent limitations—such as those related to data, LLM behavior, and evaluation methods—must remain central to our approach.

References

psychotherapy.net. `https://www.psychotherapy.net/`. Used with permission. Accessed: September 19, 2025.

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579.

AI@Meta. 2024. Llama 3 model card.

Kristen Allen, Shrey Bagroy, Alex Davis, and Tamar Krishnamurti. 2019. Convsent at clpsych 2019 task a: using post-level sentiment features for suicide risk prediction on reddit. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 182–187.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62:2937–2987.

Luna Ansari, Shaoxiong Ji, Qian Chen, and Erik Cambria. 2022. Ensemble hybrid learning methods for automated depression detection. *IEEE transactions on computational social systems*, 10(1):211–219.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Eliseo Bao, Anxo Pérez, and Javier Parapar. 2023. Explainable depression symptom detection in social media. *arXiv e-prints*, pages arXiv–2310.

Erdenebileg Batbaatar, Meijing Li, and Keun Ho Ryu. 2019. Semantic-emotion neural network for emotion recognition from text. *IEEE access*, 7:111866–111878.

Nicolas Bertagnolli. 2020. Counsel chat: Bootstrapping high-quality therapy data.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lei Cao, Huijun Zhang, and Ling Feng. 2020. Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE Transactions on Multimedia*, 24:87–102.

Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. *arXiv preprint arXiv:2310.07146*.

Ju Chun Cheng and Arbee LP Chen. 2022. Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems*, 59(2):319–339.

Liyanage C De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat.*, volume 1, pages 397–401. IEEE.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Salvatore Giorgi, Shreya Havaldar, Farhan Ahmed, Zuhaib Akhtar, Shalaka Vaidya, Gary Pan, Lyle H Ungar, H Andrew Schwartz, and Joao Sedoc. 2023. Human-centered metrics for dialog system evaluation. *arXiv preprint arXiv:2305.14757*.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik.

Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. 2011. Emotion representation, analysis and synthesis in continuous space: A survey. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 827–834. IEEE.

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts. *arXiv preprint arXiv:2205.13884*.

Stephan Hamann. 2012. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences*, 16(9):458–466.

Mohamad Farid Mohd Hayati, Mohd Adli Md Ali, and Ahmad Nabil Md Rosli. 2022. Depression detection on malay dialects using gpt-3. In *2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 360–364. IEEE.

Kilichbek Haydarov, Xiaoqian Shen, Avinash Madasu, Mahmoud Salem, Jia Li, Gamaleldin Elsayed, and Mohamed Elhoseiny. 2023. Affective visual dialog: A large-scale benchmark for emotional reasoning based on visually grounded conversations. *arXiv preprint arXiv:2308.16349*.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

Arlie Russell Hochschild. 1983. The managed heart: Commercialization of human feeling. *University of California*.

Bin Hu, Yongfeng Tao, and Minqiang Yang. 2023. Detecting depression based on facial cues elicited by emotional stimuli in video. *Computers in Biology and Medicine*, 165:107457.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, and Andrew Beam. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2022. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, 34(13):10309–10319.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, and Erik Cambria. 2023. Rethinking large language models in mental health applications. *arXiv preprint arXiv:2311.11267*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Bin Jiang, Qiang Ren, Fei Dai, Jian Xiong, Jie Yang, and Guan Gui. 2020. Multi-task cascaded convolutional neural networks for real-time dynamic face recognition method. In *Communications, Signal Processing, and Systems: Proceedings of the 2018 CSPS Volume III: Systems 7th*, pages 59–66. Springer.

Judith Johnson. judithjohnsonphd.com. https://judithjohnsonphd.com/. Used with permission. Accessed: September 19, 2025.

Yousif Khaireddin and Zhuofa Chen. 2021. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint arXiv:2105.03588*.

Byoung Chul Ko. 2018. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023a. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

R. Lotfian and C. Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 735–745.

Steve Melluish, Jon Crossley, and Alison Tweed. 2007. An evaluation of the use of simulated patient role-plays in the teaching and assessment of clinical consultation skills in clinical psychologists' training. *Psychology Learning & Teaching*, 6(2):104–113.

Francesca Minerva and Alberto Giubilini. 2023. Is ai the future of mental healthcare? *Topoi*, pages 1–9.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Ward Church, and Mohamed Elhoseiny. 2022. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. *arXiv preprint arXiv:2211.10780*.

Palash Moon and Pushpak Bhattacharyya. 2024. We care: Multimodal depression detection and knowledge infused mental health therapeutic response generation. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 296–310, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Michelle Morales, Prajjalita Dey, and Kriti Kohli. 2021. Team 9: a comparison of simple vs. complex models for suicide risk assessment. In *Proceedings of the seventh workshop on computational linguistics and clinical psychology: Improving access*, pages 99–102.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, pages 2563–2572.

Ginger Nicol, Ruoyun Wang, Sharon Graham, Sherry Dodd, and Jane Garbutt. 2022. Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the covid-19 pandemic: feasibility and acceptability study. *JMIR Formative Research*, 6(11):e40242.

Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. 2014. A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858–5869.

OpenAI. 2023. Gpt-4. Accessed: 2025-01-05.

Junhee Park and Nammee Moon. 2022. Design and implementation of attention depression detection model based on multimodal analysis. *Sustainability*, 14(6):3569.

Verónica Pérez-Rosas, Kenneth Resnicow, Rada Mihalcea, et al. 2022. Pair: Prompt-aware margin ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Yejin Choi, and Zaïd Harchaoui. 2021. Mauve: Human-machine divergence curves for evaluating open-ended text generation. *CoRR*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Kaushik Roy, Manas Gaur, Qi Zhang, and Amit Sheth. 2022. Process knowledge-infused learning for suicidality assessment on social media. *arXiv preprint arXiv:2204.12560*.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Diane E Sholomskas, Gia Syracuse-Siewert, Bruce J Rounsaville, Samuel A Ball, Kathryn F Nuro, and Kathleen M Carroll. 2005. We don't train in vain: a dissemination trial of three strategies of training clinicians in cognitive-behavioral therapy. *Journal of consulting and clinical psychology*, 73(1):106.

Gopendra Singh, Sai Vemulapalli, Mauajama Firdaus, and Asif Ekbal. 2024. Deciphering cognitive distortions in patient-doctor mental health conversations: A multimodal llm-based detection and reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22546–22570.

Aseem Srivastava, Tharun Suresh, Sarah P Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3920–3930.

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50.

Alastair C van Heerden, Julia R Pozuelo, and Brandon A Kohrt. 2023. Global mental health services and the impact of artificial intelligence–powered large language models. *JAMA psychiatry*, 80(7):662–664.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. 2020. Alone: A dataset for toxic behavior among adolescents on twitter. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 427–439. Springer.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118. IEEE.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Weipeng Zhou, Laura C Prater, Evan V Goldstein, Stephen J Mooney, et al. 2023. Identifying rare circumstances preceding female firearm suicides: validating a large language model approach. *JMIR mental health*, 10(1):e49359.

A Prompt Types

In this section, we describe the prompts employed in our experiments to assess the impact of incorporating emotional cues. We designed these prompts into four categories. Below is an example of the **Simple Prompts**, which contain only the verbal dialogue without additional emotional information.

> *Imagine you are a mental health expert interacting with a patient. Based on the dialogue provided, please apply the most suitable intervention technique and specify its name. Then, continue the dialogue with a concise response.*
>
> ***Dialogue:*** [Context]
>
> ***Therapist:*** …
> ***Patient:*** …
> ***Therapist:*** …
> ***Patient:*** …
>
> *Your response must be in the form: [Technique] [Your Response]*

Discrete Emotion Prompts: In these prompts, the patient's emotional state is predicted and provided as categorical labels (e.g., "happy", "sad", "angry"), offering a straightforward representation of affect.

Continuous Emotion Prompts: In contrast, these prompts represent the patient's emotional state with continuous values that quantify the intensity of emotions along two dimensions (e.g., valence and arousal). This approach captures the finer nuances of emotional expression.

> *Imagine you are a mental health expert interacting with a patient. Based on the dialogue provided, please apply the most appropriate intervention technique and specify its name. Continue the dialogue with a concise response. For each patient utterance, emotion predictions from both speech and facial expressions are provided.*
>
> ***Dialogue:*** [Context]
>
> ***Therapist:*** …
> ***Patient (emotion labels):*** …
> ***Therapist:*** …
> ***Patient (emotion labels):*** …
>
> *Your response must be in the form: [Technique] [Your Response]*

Emotional Dissonance Prompts: These prompts were designed to capture discrepancies between verbal content and non-verbal cues—a phenomenon we refer to as "emotional dissonance." For instance, when a patient verbally conveys positive sentiments while their facial expressions or vocal tone suggest negativity, this mismatch is explicitly highlighted in the prompt. By incorporating these conflicting emotional signals, we aim to investigate whether the models can generate more contextually accurate and nuanced interventions.

> *Imagine you are a mental health expert interacting with a patient. Based on the dialogue provided, please apply the most suitable intervention technique and specify its name. You have also noticed emotional dissonance—when the patient's vocal tone does not match their facial expression. Continue the dialogue with a concise response.*
>
> ***Dialogue:*** [Context]
>
> ***Therapist:*** …
> ***Patient:*** …
> ***Therapist:*** …
> ***Patient (emotional dissonance):*** …
>
> *Your response must be in the form: [Technique] [Your Response]*

B Dataset Collection

We compiled our dataset by systematically scraping YouTube for mental health therapy sessions. The videos span a diverse range of therapeutic modalities, including Cognitive Behavioral Therapy (CBT), Dialectical Behavior Therapy (DBT), and general counseling sessions. In total, our initial collection comprised 1562 videos. We used search queries like "CBT session roleplay," "counseling demonstration," "DBT therapist-client example". All of the videos in our dataset are role-played. **Data Inclusion/Exclusion:** During the curation process, we identified and removed duplicate videos and excluded content with poor audio quality that rendered portions of the dialogue incomprehensible. Additionally, we omitted group therapy sessions to focus exclusively on one-on-one interactions between a therapist and a patient. Ultimately, 121 videos were deemed unsuitable, resulting in a final curated dataset of 1441 videos.

For each video, we extract transcripts and assign speaker tags. Initially, GPT-4 is employed to analyze utterances and label them as either *therapist* or *patient*, followed by a manual review for accuracy.To ensure reliability, we conducted a manual validation of a random sample (10%) of the labeled sessions. In this review, we observed over 100% accuracy in role attribution. Additionally, we convert all videos to 30 FPS and utilize the Multi-Task Cascaded Convolutional Neural Networks (MTCNN) (Jiang et al., 2020) for face detection. The detected face bounding boxes are then tracked using the Kalman-filter-based SORT algorithm (Bewley et al., 2016), and tracks are annotated with speaker labels.

Video Licenses: The majority of the videos in our dataset are governed by YouTube's standard license, which restricts usage primarily to viewing on YouTube and prohibits redistribution or modification. This ensures that content creators retain control over how their videos are displayed and used. In contrast, only 2 videos are licensed under Creative Commons (CC BY), which permits broader usage—including redistribution, modification, and commercial use—provided that appropriate credit is given to the original creator. Accordingly, we provide only URL links for all videos, ensuring compliance with licensing terms and preserving the integrity of content attribution.

Figure 8: Geographical distribution of recruited mental health experts across Europe. Annotators were sourced from various European countries to ensure a diverse and representative evaluation panel.

C Implementation Details

Valence-Arousal Estimation from Audio:
We employ a pretrained model by Wagner et al. (Wagner et al., 2023) to estimate valence, arousal, and dominance from raw audio signals. This model is based on the Wav2Vec2-Large-Robust architecture (Baevski et al., 2020), which was fine-tuned on the MSP-Podcast (v1.7) dataset (Lotfian and Busso, 2019). Prior to fine-tuning, the model was pruned from 24 to 12 transformer layers to enhance computational efficiency.
Valence-Arousal Estimation from Video:
For video-based emotion estimation, we utilize EmoNet (Toisoul et al., 2021), a model that jointly predicts both categorical and continuous emotions. We extract valence-arousal values for each frame within a segment based on audio-aligned timestamps and aggregate these estimates by averaging, thereby obtaining a robust representation of the emotional state over the segment.
Threshold for identifying Emotional Dissonance
We selected the threshold for emotional dissonance (0.5) based on a visual inspection of the absolute differences in valence and arousal between the audio and video modalities (see Figure 10). This observational analysis indicated that a threshold of 0.5 effectively separates segments with notable discrepancies from those with minor differences. While this choice is primarily heuristic, it provides a reasonable starting point for identifying significant emotional dissonance and can be further refined in future studies. While heuristic, this choice was data-driven, transparently reported, and positioned as a starting point for future refinement, not a fixed clinical standard.

D Human Studies

Evaluation Procedure: We engaged a panel of mental health experts (with proficiency in English) to evaluate the therapeutic interventions generated by our models based on three criteria — Applicability, Empathy, and Comprehensiveness — each rated on a 1-to-5 scale. Evaluations were conducted via a custom user interface (see Figure 9), where experts assessed interventions in the context of provided video segments. This approach enabled an objective assessment of the models' ability to generate meaningful and contextually appropriate therapeutic responses.
Expert Recruitment and Background:
To ensure diversity among evaluators, we administered a brief recruitment questionnaire to collect background information. Based on self-reported data (see Figure 11), we found that:

- Two experts hold Bachelor's degrees, one holds a Master's degree in psychology, and two hold PhD degrees.
- Experts with PhDs have over 10 years of clinical experience, while the remaining experts average approximately 4 years.
- Three experts have an orientation toward Cognitive Behavioral Therapy (CBT).
- Four experts strongly agree that "the therapeutic relationship and empathy are the most critical factors in achieving successful therapy outcomes."

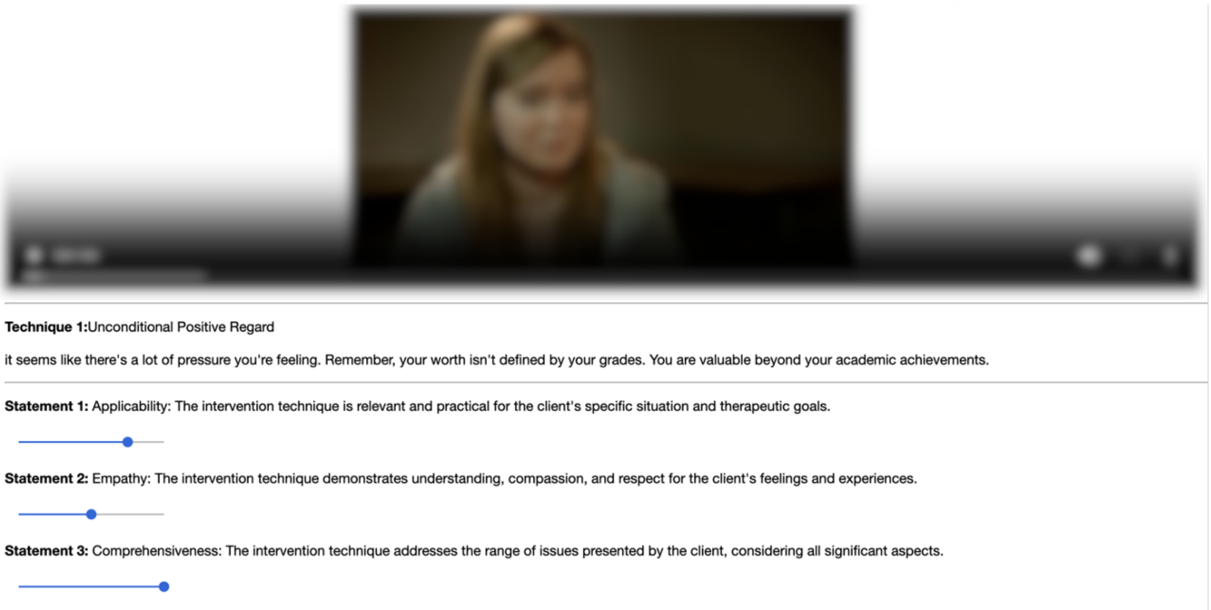


Figure 9: User interface for expert evaluations. This interface presents the video context along with the generated intervention technique and response, enabling mental health experts to provide structured ratings on a Likert scale from 1 to 5 across the defined evaluation criteria. Images used with permission from (Johnson).

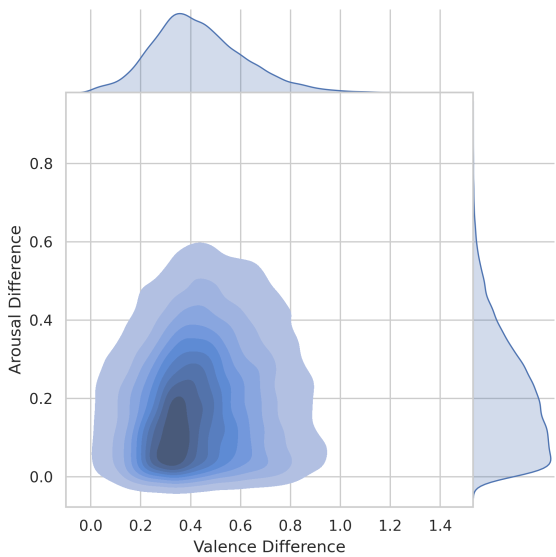


Figure 10: Absolute Valence-Arousal Difference plot between audio and speech modalities

- Three experts reported extensive formal training or certification in CBT approaches, while one indicated a moderate level.

- More than half of the experts emphasized the importance of incorporating non-verbal cues (e.g., facial expressions, body language) into clinical assessments, and acknowledged the critical role of non-verbal communication in understanding a client's emotional state.

Mock-Therapy vs. Real-World Scenarios: Although our dataset comprises mock therapy sessions, we sought to validate their realism by surveying mental health experts. In our survey, experts were asked whether the scenarios depicted in the videos closely resembled real-world therapeutic cases. Notably, 90% of respondents agreed that the videos capture key aspects of genuine therapy, thereby supporting our assumption that our mock-therapy dataset serves as a meaningful proxy for real-world interactions.

Compensation:
Experts were compensated $7 per video, ensuring fair remuneration for their time and professional expertise.

E Compute Consumption

All experiments were executed on a single NVIDIA V100 GPU, which offered an optimal balance of performance and efficiency. Table 4 details the GPU hours consumed for each component of our pipeline. Notably, bounding box extraction required 20 GPU hours, valence-arousal estimation used 25 GPU hours for video and 10 GPU hours for audio processing, and generating responses from open-source models (LLaMa3-8b and Mistral-7b) required an additional 2 GPU hours.

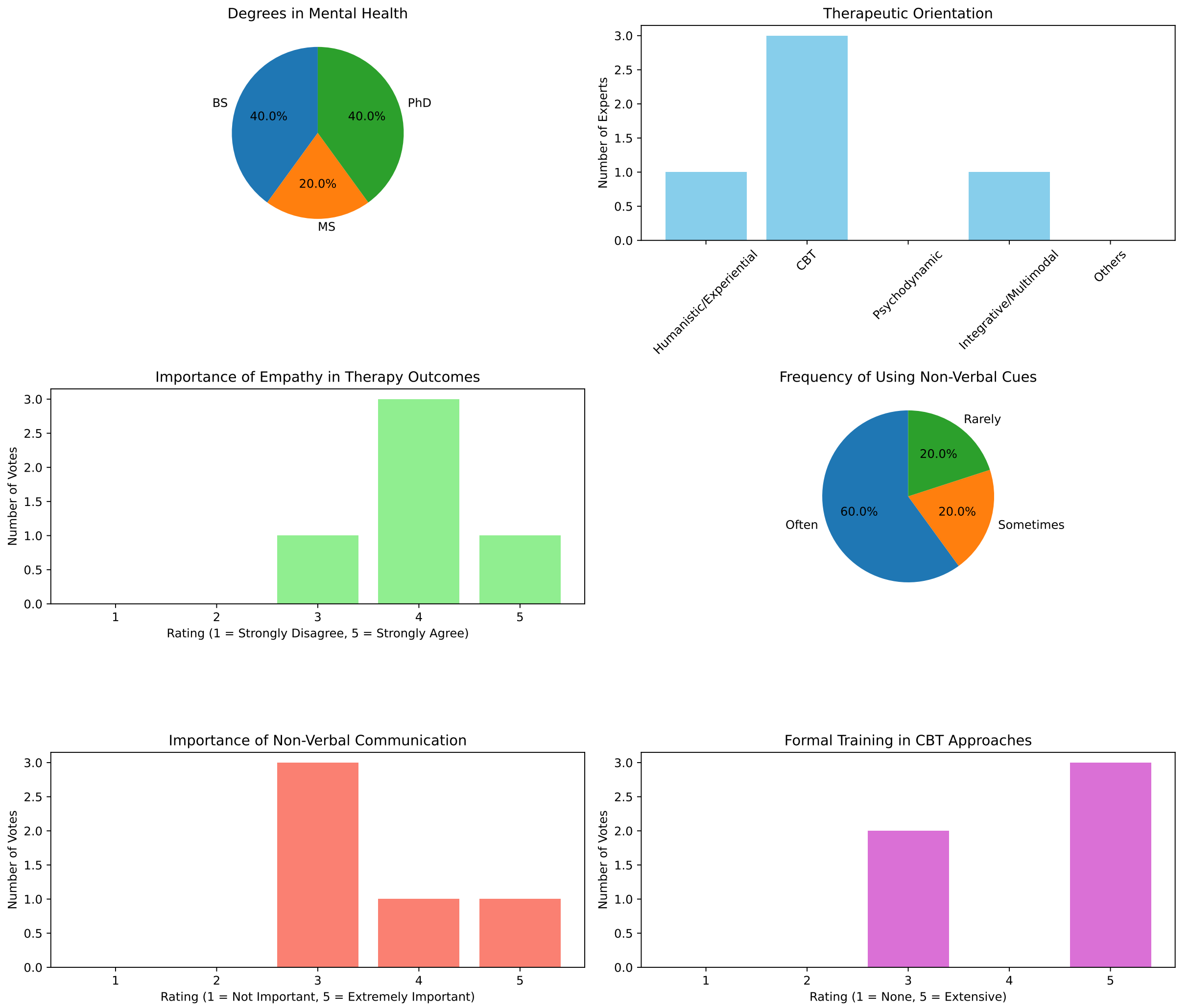


Figure 11: Summary of expert background and evaluation metrics: (a) Degrees in Mental Health, (b) Therapeutic Orientation, (c) Importance of Empathy in Therapy Outcomes, (d) Frequency of Using Non-Verbal Cues, (e) Importance of Non-Verbal Communication, and (f) Formal Training in CBT Approaches.

Table 4: GPU Hours Consumption for Experiments

Experiment	**GPU Hours**
Bounding Box Extraction	20
Valence-Arousal Estimation (video)	25
Valence-Arousal Estimation (audio)	10
Response Extraction (Open-Source Models)	2

F Failure Cases

Despite the overall promising performance of LLMs, we observed several failure cases where the generated responses did not effectively continue the dialogue. Instead of providing a proper conversational continuation, the model occasionally produced explanatory suggestions or clarifications. For instance, some responses were structured as follows:

- *"Clarification: This encourages the patient to elaborate on their feelings and provides valuable insight into their emotional state."*
- *"Clarification: This response seeks to gently clarify the patient's statements, particularly the juxtaposition of 'phone' with deeply personal needs."*

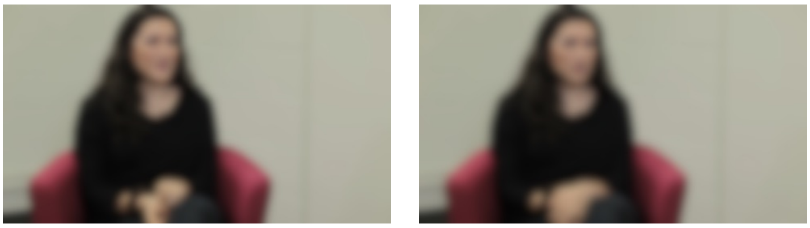
These examples illustrate that, in certain instances, the model defaults to offering an explanation rather than generating an interactive continuation of the therapeutic dialogue. Such failures highlight the need for further refinement or model fine-tuning to ensure that responses remain contextually relevant and maintain the flow of the conversation.

G Use of AI Assistant in Paper Writing

Throughout the writing process, we leveraged ChatGPT as an AI writing assistant. Its contributions in refining language and enhancing the clarity of our arguments significantly improved the overall coherence and organization of the manuscript, allowing us to focus on the substantive aspects of our research.

H Additional Qualitative Examples

Figures 12, 13 provide further qualitative examples generated by our method. We also experimented with video understanding models based on LLMs, such as Video-LLaMa (Zhang et al., 2023) and Video-LLaVa (Lin et al., 2023) . However, these models exhibited limitations in handling generative intervention tasks and long-context videos. They often failed to adhere to provided instructions and primarily produced descriptive outputs, as illustrated in Figure 12. This shortcoming is attributed to their lack of fine-tuning for domain-specific tasks and their optimization for shorter input durations (up to 1 minute).



Speaker	Start Time	End Time	Utterance
A	00:00:29	00:00:37	You. Okay, so you said that when you're at work, you really worry that you're going to say something stupid.
B	00:00:38	00:00:41	Yeah, that's something that's always on my mind.
A	00:00:41	00:00:45	What would be so bad about that if you did say something stupid at.
B	00:00:45	00:00:49	Work because people would think I was weird.
A	00:00:49	00:00:53	Then what would be so bad? If people at work thought you were weird?
B	00:00:54	00:00:56	They wouldn't want to be friends with me.

GPT-4: **Technique: Cognitive Behavioral Therapy (CBT) - specifically, Socratic questioning.**
And if they didn't want to be friends with you, what impact would that have on you?

LLaMa3: **Technique: Cognitive-behavioral therapy (CBT)**
Okay, so it sounds like you're really struggling with these negative beliefs about yourself. Can you tell me more about where those beliefs come from?

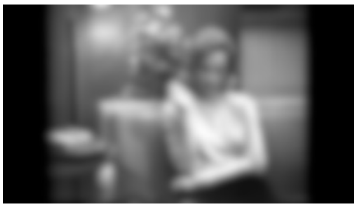
GPT-4-Disson: **Technique: Emotional Awareness and Validation.**
I sense some discomfort when you mention your colleagues not wanting to be friends. Can you tell me what feelings come up when you think about that situation?

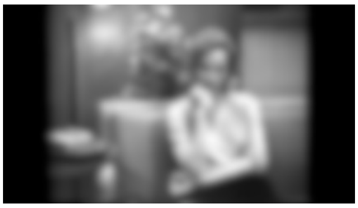
LLaMa3-Disson: **Technique: Cognitive Restructuring**
I understand that you're worried about what your coworkers might think of you, but is it really likely that they would judge you so harshly? And even if they did, would that really be the end of the world? It's important to remember that everyone has their own quirks and imperfections, and it's okay to make mistakes. It's how we learn and grow.

Video-LLaMa: The video appears to be a professional video with good lighting and quality. The technique used for this video could be the Stutter Step technique or possibly even the Anderson Counseling technique.

Video-LLaVa: Based on the video, the best intervention technique for the patient would be to provide them with a comfortable and supportive environment, such as a red chair, to help them feel at ease.

Figure 12: Example 1. Images used with permission from (Johnson)





Speaker	Start Time	End Time	Utterance
A	00:00:00	00:00:00	You.
B	00:00:02	00:00:11	Now you'll have to talk before I can talk because I just get scared when it gets quiet.
A	00:00:13	00:00:39	Well, we started out with how you happened to come into the hospital and we've been talking about your parents. What were the particular happenings that's well.
B	00:00:40	00:01:22	When I was little, we lived with my grandmother a lot. His mother. And she and Mother would go uptown to the movie or something. And they'd always leave me with him because he was either home drunk or with a hangover from being drunk. And then that is when he would do what he did to me. And I can remember going to the door and crying for them not to leave me. And they would every time you I can remember one time when I had a girl to spend the night with me, and he did. And my mother was there a lot of times.
A	00:01:39	00:01:44	You think he's sick? I guess you've already said that.
B	00:01:46	00:01:57	In fact, I can look at him now. And he has really I think he did what I did. He just regressed or something. He's got the mind of a child.

GPT-4o: **Technique: Trauma-focused Cognitive Behavioral Therapy (TF-CBT)**
I'm sorry to hear what you've been through. It must have been very difficult. It's important to understand that what happened was not your fault. You were a child and you deserve to be cared for and loved. Let's work on understanding how these significant experiences have shaped your emotions and behavior, and as we progress, we'll focus on how to cope and heal.

LLaMa3: **Technique: Reflective listening**
"It sounds like you're saying that your father abused you when you were a child, and that it's had a profound impact on your life. You're worried that he might hurt someone else, and you're struggling to come to terms with what happened to you. Is that right?”

GPT-4o-Disson: **Technique: Psychoeducation**
It sounds like you've been through a lot of difficult experiences with your grandfather. It’s understandable that these memories are painful for you. Let's explore how these experiences have affected you and how we can work through them together.

LLaMa3-Disson: **Technique: Emotion Dissonance:**
"I notice that you seem to be struggling with conflicting emotions when you talk about your father’s abuse. It's understandable that you would feel upset and scared, but it's also clear that you feel a sense of empathy for your father and his own struggles. It's important to recognize that these feelings are valid and can coexist, even if they may seem contradictory. Can you tell me more about what you’re feeling and how you're making sense of these complex emotions?"

Figure 13: Example 2.