Feel the Difference? A Comparative Analysis of Emotional Arcs in Real and LLM-Generated CBT Sessions

Xiaoyi Wang[1], Jiwei Zhang[1], Guangtao Zhang[2*], Honglei Guo[3*]
[1]Department of Computer Science, Shantou University, Shantou, Guangdong, China
[2]Department of Automation, Tsinghua University, Beijing, China
[3]BNRIST, Tsinghua University, Beijing, China

Abstract

Synthetic therapy dialogues generated by large language models (LLMs) are increasingly used in mental health NLP to simulate counseling scenarios, train models, and supplement limited real-world data. However, it remains unclear whether these synthetic conversations capture the nuanced emotional dynamics of real therapy. In this work, we introduce RealCBT, a dataset of authentic cognitive behavioral therapy (CBT) dialogues, and conduct the first comparative analysis of emotional arcs between real and LLM-generated CBT sessions. We adapt the Utterance Emotion Dynamics framework to analyze fine-grained affective trajectories across valence, arousal, and dominance dimensions. Our analysis spans both full dialogues and individual speaker roles (counselor and client), using real sessions from the RealCBT dataset and synthetic dialogues from the CACTUS dataset. We find that while synthetic dialogues are fluent and structurally coherent, they diverge from real conversations in key emotional properties: real sessions exhibit greater emotional variability, more emotion-laden language, and more authentic patterns of reactivity and regulation. Moreover, emotional arc similarity remains low across all pairings, with especially weak alignment between real and synthetic speakers. These findings underscore the limitations of current LLM-generated therapy data and highlight the importance of emotional fidelity in mental health applications. To support future research, our dataset RealCBT is released at https://gitlab.com/xiaoyi.wang/realcbt-dataset.

1 Introduction

Mental disorders pose a major global health challenge, affecting nearly one in eight individuals worldwide (WHO). As demand for mental health services continues to rise, a significant barrier remains: the shortage of trained counselors. To help bridge this gap, large language models (LLMs) such as ChatGPT and Gemini are increasingly explored as conversational agents capable of simulating therapeutic interactions and supporting clients in need. The effectiveness of these systems, however, hinges heavily on the quality and diversity of training data—particularly, counseling dialogues grounded in real-world counseling practice.

Unfortunately, access to real counseling dialogues remains extremely limited due to stringent privacy, ethical, and legal constraints. This scarcity of authentic data has led to the widespread adoption of synthetic therapy dialogues generated by LLMs. These synthetic interactions are now commonly used in mental health NLP applications for training, evaluation, and scenario simulation. For example, datasets like CACTUS (Lee et al., 2024) incorporate psychological principles from Cognitive Behavioral Therapy (CBT) to structure model-generated conversations. While these synthetic dialogues are often fluent and structured around known therapeutic techniques, it remains unclear whether they capture the nuanced emotional dynamics that characterize real counseling sessions.

Emotion plays a central role in both diagnosis and treatment. In CBT, client emotions reveal cognitive distortions, inform intervention strategies, and signal therapeutic progress. A meaningful session often involves shifts in emotional state, such as the alleviation of distress or the emergence of insight. These emotional trajectories, or *emotional arcs*, offer a powerful lens through which to assess the depth, authenticity, and therapeutic quality of a dialogue. It is thus important to understand the extent to which synthetic therapy conversations mirror the emotional dynamics observed in real sessions.

Prior work has demonstrated that emotional arc analysis can yield rich insights into narrative structure across various domains, including novels (Vishnubhotla et al., 2024; Ohman et al., 2024;

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19999–20017
November 4-9, 2025

Teodorescu and Mohammad, 2023; Mohammad, 2012), social media posts (Vishnubhotla and Mohammad, 2022), and movie scripts (Hipson and Mohammad, 2021). However, emotional arcs remain largely unexplored in the context of psychotherapy, and no prior study has directly compared real and synthetic counseling dialogues from an emotional dynamics perspective.

To address this gap, we curate RealCBT, a dataset of authentic cognitive behavioral therapy dialogues transcribed from public video-sharing platforms (e.g., YouTube and Vimeo). We then conduct a comparative emotion trajectory analysis between these real-world sessions and synthetic dialogues from the CACTUS dataset (Lee et al., 2024). We examine emotional arcs from two perspectives: (1) *The Emotion Dynamics of Real vs. Synthetic Therapy Dialogues*: How does the emotion change in real and LLM-generated CBT dialogues: for entire dialogues, for counselors, and for clients? (2) *Emotion Arc Alignment Across Speaker Pairs*: To what extent do emotional trajectories align within and across speaker types (specifically among real–real, synthetic–synthetic, and real–synthetic pairings) for both counselors and clients?

To answer these questions, we conduct the first in-depth comparison of emotional dynamics between real and LLM-generated CBT dialogues. We use the well-established Utterance Emotion Dynamics (UED) framework (Teodorescu and Mohammad, 2023; Hipson and Mohammad, 2021), adapted specifically for CBT dialogues, to measure emotion trajectories over time within counselor–client interactions. Our goal is both to compare real and LLM-generated CBT dialogues and to establish an empirical benchmark for assessing whether synthetic sessions capture the nuanced emotional dynamics of real therapy. This benchmark provides an important foundation for future work exploring more contextualized and neural emotion models.

Our analysis reveals that although synthetic dialogues are structurally fluent, they diverge from real sessions across several key emotional dimensions. Real therapy sessions display greater emotional variability, more emotion-laden language, and more authentic patterns of emotional reactivity and regulation. Emotional arc correlations are low across all pair types, though synthetic–synthetic pairs show slightly higher alignment than real–real or real–synthetic pairs.

These findings suggest that while current LLM-generated dialogues may capture surface-level therapeutic cues, they fall short in reproducing the nuanced, co-constructed emotional flow that characterizes real-world therapy, especially in the client role.

Our contributions are threefold:

- **RealCBT dataset**: We curate and release RealCBT, one of the first publicly available datasets of authentic CBT dialogues.
- **Empirical benchmark**: We establish an empirical benchmark for comparing emotional dynamics in real vs. synthetic therapy dialogues, using the UED framework adapted for CBT.
- **Comparative insights**: We identify key divergences between LLM-generated and real CBT sessions, highlighting where current models fail to capture the variability, regulation, and alignment of authentic therapeutic emotion trajectories.

2 Related Work

2.1 Cognitive Behavior Therapy

Cognitive Behavioral Therapy (CBT) is a widely adopted, evidence-based form of psychotherapy used to treat a variety of psychological disorders, including depression, anxiety, and addiction (Beck et al., 2011). The core of CBT aims to help clients identify and challenge maladaptive thought patterns, and subsequently restructure these cognitive distortions to promote more realistic and adaptive thinking, ultimately facilitating emotional and behavioral change (Carroll and Kiluk, 2017; Longmore and Worrell, 2007). Given its structured format and demonstrated effectiveness, CBT has increasingly been adapted into virtual agents to support individuals with limited access to in-person therapy.

2.2 Synthetic Mental Health Counseling Dialogue Datasets

Training CBT-based conversational models requires high-quality therapy dialogue datasets. However, such data are difficult to obtain due to privacy, ethical, and legal constraints. To address this limitation, recent studies have explored the use of LLMs to generate synthetic therapy dialogues that can supplement scarce real-world data. Earlier work (Sharma et al., 2023; Maddela et al.,

2023; Sun et al., 2021) has focused on generating single-turn counseling dialogues based on CBT or other psychological strategies. More recent research has shifted toward multi-turn generation that better simulates the interactive nature of real therapy sessions (Lee et al., 2024; Xiao et al., 2024; Qiu et al., 2024). Among these, CACTUS (CBT-augmented Counseling Chat Corpus) stands out as a publicly available multi-turn dataset designed to capture the flow and depth of CBT-style conversations (Lee et al., 2024). It uses LLMs to simulate counselor–client interactions guided by CBT theory and therapeutic intent. Given its theoretical grounding and multi-turn structure, we adopt CACTUS as the synthetic benchmark dataset for comparison with real-world CBT dialogues in this study. Although CACTUS has been evaluated favorably on key metrics such as helpfulness and empathy, recent work (Lee et al., 2025) raises concerns that LLM-generated therapy dialogues may lack the richness and diversity of emotional expression observed in real interactions. This underscores the need to better understand how synthetic dialogues compare to real-world CBT conversations, particularly in terms of emotional dynamics, to guide the design of more realistic and effective counseling models.

2.3 Lexicon-based Emotion Dynamics Computation

Emotion dynamics is a psychological framework for measuring how an individual's emotional state evolves over time (Vishnubhotla et al., 2024; Kuppens and Verduyn, 2017; Hollenstein, 2015). In NLP, lexicon-based approaches are among the most widely used techniques for modeling emotion trajectories in spoken or written language. Prominent resources include LIWC (Tausczik and Pennebaker, 2010), WordNet-Affect (Bobicev et al., 2010), SentiWordNet (Baccianella et al., 2010), VADER (Hutto and Gilbert, 2014), and the NRC Emotion Lexicons (Mohammad, 2018). These lexicons provide emotion-related scores at the word level, enabling the quantification of emotional trends across text.

Lexicon-based methods have been validated in various contexts and shown to achieve high reliability in capturing emotional dynamics (Ohman et al., 2024; Teodorescu and Mohammad, 2023; Öhman, 2021). In this study, we adopt a lexicon-based approach, specifically using the NRC Valence, Arousal, and Dominance (VAD) Lexicon (Mohammad, 2025, 2018) in conjunction with the Utterance Emotion Dynamics (UED) framework (Hipson and Mohammad, 2021). This combination allows us to compute the emotional valence, arousal, and dominance at the utterance level, and to model how these emotional states change over the course of each counseling session.

Previous research has successfully applied similar methods to track emotional arcs in literary novels (Vishnubhotla et al., 2024; Teodorescu and Mohammad, 2023; Mohammad, 2012), social media narratives (Vishnubhotla and Mohammad, 2022), and movie scripts (Hipson and Mohammad, 2021). We extend this line of work by applying NRC and UED metrics to therapeutic dialogues, allowing us to compare emotional progression in real and synthetic CBT conversations.

3 RealCBT: Real-world CBT Dialogue Dataset

In this section, we describe the process of curating the **RealCBT** dataset.

3.1 Data Collection

Due to privacy, ethical, and legal constraints, publicly available CBT dialogue datasets are extremely limited. To enable comparison with LLM-generated data, we collected real-world CBT Dialogues by identifying video clips from public video-sharing platforms such as YouTube and Vimeo. We restricted our search to videos explicitly labeled as CBT-based counseling sessions. For a detailed description of the selection criteria, please see Appendix A.1. In total, we collected 76 video-based CBT dialogues that met these criteria. Detailed summary statistics of **RealCBT** is provided in Table 3 in Appendix A.2.

3.2 Preprocessing and Transcription

All videos were first converted into a standard MP4 format and then preprocessed to remove non-conversational, user-generated content such as introductory titles, animations, and narration. We used the Transkriptor[1] service to generate initial transcripts for each video. To ensure transcript quality and temporal alignment, each transcript was manually reviewed and corrected to accurately reflect the spoken dialogue in the video. Additionally, we reviewed metadata associated with each video (e.g., the number of views, likes, and information

[1] https://transkriptor.com/

about the channel owner) to verify that the content originated from credible sources. These indicators collectively suggest that the selected videos represent professionally conducted CBT sessions.

3.3 Metadata Annotation

Similar to CACTUS, we annotated each dialogue with the following attributes from video metadata: (1) client problem, (2) client gender, and (3) overall client attitude toward the session. We employed three state-of-the-art language models (i.e., ChatGPT-4o Mini, Grok-v3, and Gemini 2.0 Flash) to independently infer these tags for each dialogue. A majority voting scheme was then used to determine the final label for each attribute. To evaluate the accuracy of this automated annotation approach, we manually reviewed a randomly selected subset comprising 30% of the dialogues. The automated predictions were consistent with the human annotations in this subset, yielding an estimated labeling accuracy of 100%. Detailed distribution of the metadata of **RealCBT** is provided in Table 4 in Appendix A.2.

4 Emotion Dynamics Computation

We adapt emotion dynamics methods from prior work by Mohammad and colleagues (Vishnubhotla et al., 2024; Teodorescu and Mohammad, 2023; Hipson and Mohammad, 2021), who demonstrated robust approaches for capturing emotional arcs in narratives and movies. Specifically, we adopt the Utterance Emotion Dynamics (UED) framework to compute emotion metrics from sequences of utterances, treating each role (counselor or client) as a separate character trajectory.

Given that CBT sessions are considerably shorter than novels or long-form narratives, we adjust the emotion arc computation accordingly. Following prior work (Hipson and Mohammad, 2021; Vishnubhotla et al., 2024), we apply a rolling window of 10 words, advancing one word at a time, to compute fine-grained emotion values across each dialogue. At each windowed step, we estimate the emotional state using version 2 of the NRC Valence, Arousal, and Dominance (VAD) Lexicon (Mohammad, 2025), which provides word-level emotion scores for the three affective dimensions: valence (V), arousal (A), and dominance (D). Each V/A/D score ranges from –1 to 1, where 1 indicates a strong positive association and –1 indicates a strong negative association with the corresponding affective dimension.

From the resulting VAD time series, we derive a set of UED metrics (Hipson and Mohammad, 2021) that capture the shape and character of emotion progression in therapy dialogues: (1) *Emotion Mean*: The average V/A/D score across a dialogue or speaker's arc; (2) *Emotion Variability*: The standard deviation of V/A/D scores, representing emotional fluctuation or richness; (3) *Average Displacement Length*: Measures the average number of emotion words spoken more or less than usual. Longer displacements suggest stronger or more persistent emotional shifts; (4) *Emotion Rise Rate*: Captures how quickly and intensely a speaker shifts into an emotional state (emotional reactivity or sensitivity); (5)*Emotion Recovery Rate*: Indicates how quickly a speaker returns to their baseline emotional state after a shift (emotion regulation).

Based on the UED metrics and NRC VAD Lexicon, we compute three types of emotional trajectories per CBT session: (1) the overall dialogue-level arc, (2) the counselor's arc, and (3) the client's arc. These arcs are computed separately for both real and synthetic CBT sessions, allowing us to compare matched pairs across datasets (i.e., real vs. synthetic dialogues, real vs. synthetic counselors, and real vs. synthetic clients).

5 Experimental Setup

We conduct our study using two datasets. CACTUS consists of 31,564 dialogues with broadly balanced distributions across multiple attributes, making it well suited for large-scale statistical analysis. In contrast, RealCBT comprises 76 dialogues collected from real counseling sessions. Although smaller in scale, RealCBT captures authentic therapeutic interactions and naturally reflects demographic and attitudinal patterns found in practice.

To ensure a fair and meaningful comparison between the two datasets, we adopt a problem-focused sampling strategy. Specifically, we focus on the three most frequent client concerns in RealCBT, including anxiety and fear (25 dialogues), self-esteem and confidence issues (19), and relationship-related concerns (14). Together, these account for 76.3% of the dataset (58 out of 76 dialogues). For detailed summary statistics and distributional information, see Table 5 and Table 6 in Appendix A.2. We then sample synthetic dialogues from CACTUS to match this distribution.

While CACTUS is broadly balanced, RealCBT

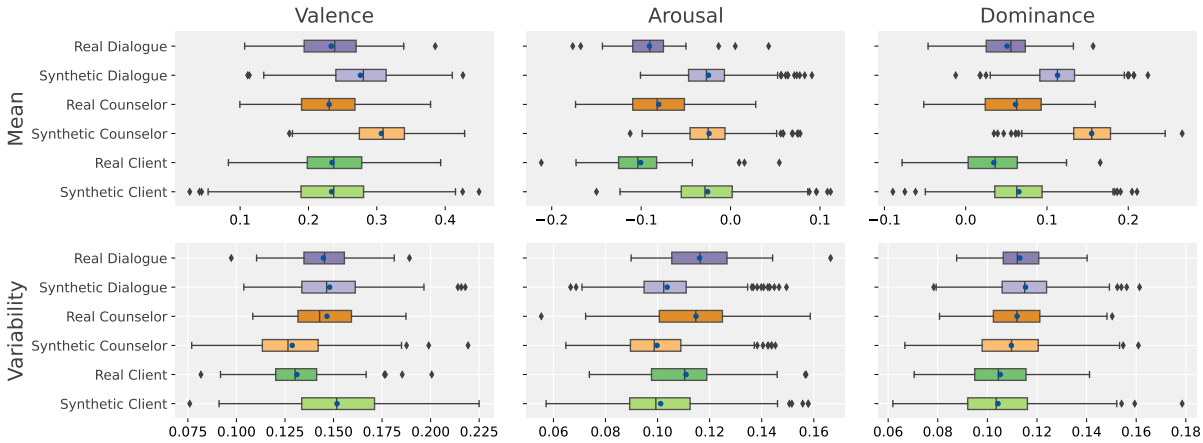

Figure 1: Boxplots showing the distributions of the mean and variability for each of the three affective dimensions across three comparisons: Real vs. Synthetic Dialogues, Real vs. Synthetic Counselors, and Real vs. Synthetic Clients.

reflects natural demographic and attitudinal skews (e.g., a higher proportion of female clients and generally positive orientations toward therapy). Such contextual characteristics make strict subgroup balancing less stable, further motivating our problem-focused design.

To enhance the robustness of our analysis, we repeat the sampling process 10 times without replacement, generating 10 non-overlapping subsets of synthetic dialogues. All quantitative results are averaged across these runs to account for sampling variability.

6 Experimental Results

6.1 The Emotion Dynamics of Real vs. Synthetic Therapy Dialogues

We analyze and compare the following UED metrics across valence, arousal, and dominance: Emotion Mean, Variability, Displacement Length, Rise and Recovery Rates. Detailed definitions and computation methods for each metric are provided in Section 4. To assess statistical significance between real and synthetic groups, we apply the Mann–Whitney U test ($\alpha = 0.05$) to compare real sessions and each of 10 independently sampled synthetic datasets. For robustness, we report median p-values and corresponding effect sizes (r). Figure 1 presents boxplots illustrating the distributions of Emotion Mean and Variability for each role (i.e., dialogue, counselor, and client) in both real and synthetic sessions. Blue dots indicate the aggregated Emotion Mean and Variability for each affective dimension, summarizing overall emotional trends. Detailed summary statistics for Emotion Mean, Emotion Variability, and all UED metrics are provided in Table 7 - 10 in Appendix A.3.

6.1.1 Emotion Mean

As shown in Figure 1, synthetic speakers (whole dialogue, counselor, and client) generally sound more emotionally elevated or intense than real human speakers. An exception is observed in valence, where real and synthetic clients show similar scores. The pairwise comparisons support this observation. For entire dialogues (as a whole speaker), we found significant differences in arousal in 90% of cases (median p-value < 0.001, $r = 0.24$). For counselors, we observed significant differences in valence in all comparisons (median p-value < 0.001, $r = 0.36$). For clients, we found significant differences in arousal in 70.0% of comparisons (median p-value < 0.05, $r = 0.23$) and in valence in 90% (median p-value < 0.05, $r = 0.26$).

6.1.2 Emotion Variability

For entire CBT dialogues, real sessions exhibited significantly greater variability in arousal than synthetic ones, indicating more fluctuation or richness in real speakers. The Mann–Whitney U tests showed significant differences, with a median p-value < 0.0001 and $r = 0.84$, indicating a strong effect. In contrast, valence and dominance variability were similar between real and synthetic dialogues, with only partial significant differences (i.e., 5/10 in dominance and 7/10 in valence).

For counselors, real speakers also showed significantly greater variability in both arousal and valence compared to synthetic counselors. For arousal, all comparisons were significant (median p-value < 0.0001, $r = 0.41$). For valence, all comparisons were significant (median p-value $<$

0.00001, r = 0.42). For dominance, real counselors tended to exhibit greater variability, though the differences were not consistently significant across the comparisons.

For clients, synthetic sessions tended to show greater variability in valence, while real clients showed slightly more variability in arousal. However, these differences were only partially significant (i.e., 6/10 for arousal and 1/10 for valence). Dominance variability was similar across both groups, with no significant differences observed.

6.1.3 Displacement Lengths

We found that speakers in real dialogues talk significantly more emotion words than the ones in synthetic dialogues in both arousal and dominance. In particular, for arousal, all comparisons showed significant differences, with a median p-value $<$ 0.0001 and r = 0.35. For dominance, 70% of comparisons were significant, with a median p-value $<$ 0.05 and r = 0.23. In contrast, both real and synthetic speakers uttered similar amount of emotion words in dominance with minimal difference—only 1 out of 10 comparisons reaching significance.

Real counselors tend to speak more emotion words than synthetic counselors. 70% of comparisons showed significant differences in both arousal and valence. The median p-values were 0.02 (arousal) and 0.007 (valence), with average effect sizes of 0.23 in both cases. In contrast, dominance minimal showed significant differences in only 20% of comparisons, indicating similar amount of emotion words uttered by both real and synthetic counselors.

Real clients utter more emotion words than synthetic ones in the dimensions of arousal and dominance. Arousal showed consistent differences across all comparisons, with a median p-value $<$ 0.001 and r = 0.335. Dominance showed partial significant differences in 70% of comparisons, with a median p-value $<$ 0.05 and r = 0.24. In contrast, valence showed no significant differences across any comparisons.

6.1.4 Emotion Rise Rate

We observed that synthetic speakers tend to respond more quickly and intensely to emotional situations than real speakers in dominance (synthetic 0.033 vs. real 0.031) and valence (synthetic 0.026 vs. real 0.024). In particular, dominance showed relative strong significant differences in 70% of comparisons with the median p-value $<$ 0.05 and r = 0.2). Similarly, valence demonstrate partial significant differences (50% of comparisons, median p-value $<$ 0.05 and r = 0.19). However, real speakers seem to be slightly higher arousal scores than synthetic ones with no significant differences.

Real and synthetic counselors have very similar scores in valence (0.033 vs. 0.031), arousal (0.025, 0.026), and dominance (0.0247, 0.0254), indicating that they react to emotional situations with similar intensity and speed. There were no consistence significant differences observed.

Synthetic clients seem to be more actively respond to emotional situations than real ones in the dimensions of valence (0.035, 0.032) and dominance (0.026, 0.024). Both dimensions showed significant differences in 60% of comparisons with the median p-value valence $<$ 0.05 and r = 0.197 and dominance $<$ 0.05 and r = 0.19. Both real and synthetic clients reacted similarly in arousal (0.027, 0.026) with no significant differences.

6.1.5 Emotion Recovery Rate

In general, synthetic and real speakers showed similar ability to return to their baseline emotional states after a shift in valence (0.033 vs. 0.031) and arousal (0.025 vs. 0.026), with no significant differences observed. However, dominance showed partial differences: 60% of comparisons were significant, with a median p-value $<$ 0.05 and r = 0.22, suggesting that synthetic speakers may exert slightly more emotional control in this dimension.

For counselors, emotion regulation ability was comparable across all three affective dimensions: valence (real: 0.0324 vs. synthetic: 0.0304), arousal (0.0261 vs. 0.0260), and dominance (0.0250 vs. 0.0248). No significant differences were observed.

In contrast, synthetic clients exhibited stronger emotion regulation than real clients in both valence and dominance. Valence showed significant differences in all comparisons (median p-value $<$ 0.0001 and r = 0.4), while dominance showed partial significance in 50% of comparisons (median p-value $<$ 0.05 and r = 0.17). No significant differences were found in arousal, suggesting similar regulation patterns in that dimension.

6.1.6 Emotional Arc Case Study

To complement the quantitative results, we conduct a qualitative case study to illustrate how emotional dynamics manifest in real and synthetic CBT dialogues. Our analysis spans all three dimensions

Arousal Level	Real Counselor Snippet	Synthetic Counselor Snippet
High	"It's completely understandable to feel overwhelmed! You've taken a huge step today—this is a big deal, and you're doing really well." *Explanation*: Demonstrates emotional reappraisal and support, key to emotion regulation.	"Wow! That's amazing progress! You should be really proud of yourself—this is wonderful!" *Explanation*: General praise, not a deliberate attempt to regulate the client's emotional state.
Medium	"That's good to hear. Can you walk me through what happened when you started to feel anxious?" *Explanation*: Shows calibrated empathy by gently probing the client's state with a tailored follow-up.	"That's interesting. Let's explore that feeling more." *Explanation*: It's vague and not clearly grounded in what the client just said.
Low	"Okay. Let's keep going." *Explanation*: Even when emotional intensity is low, the counselor maintains engagement, which contributes to dyadic co-regulation.	"Thanks for sharing that." *Explanation*: Like a conversational endpoint, with no guidance or co-regulation of emotional flow.

Table 1: Illustrative Counselor Utterances Comparison at Varying Arousal Levels

(valence, arousal, and dominance) for both counselors and clients, with detailed examples included in Appendix A.4 Table 11–15. Here, we highlight counselor arousal (Emotion Mean) as a representative case: Table 1 presents excerpts showing how different levels of arousal appear in practice.

We selected three real and three synthetic sessions representing the highest, median, and lowest mean arousal scores among counselor utterances. Table 1 provides representative examples from each, along with interpretive commentary informed by key psychotherapy theories (e.g., emotion regulation (Gross, 1998), calibrated empathy (Elliott et al., 2018, 2013), affective co-regulation (Butler and Randall, 2013), and therapeutic alliance theory (Bordin, 1979)).

These examples demonstrate how different levels of affective intensity appear in real versus LLM-generated therapy conversations. For instance, while both real and synthetic counselors express high arousal through supportive language, synthetic speakers tend to rely on generic praise ("That's amazing progress"), which lacks co-regulatory intent and risks feeling generic or unearned. In contrast, real counselors often express enthusiasm through emotionally supportive reappraisals grounded in the client's situation (e.g., "You've taken a huge step today…"). This reflects emotion regulation strategies that promote client resilience. At medium arousal, real counselors demonstrate calibrated empathy, engaging clients with reflective questions that validate while encouraging elaboration (e.g., "Can you walk me through what happened…?"). Synthetic counselors, however, tend to offer vague prompts (e.g., "Let's explore that feeling"), which may miss therapeutic opportunities to deepen client reflection. In low-arousal moments, real counselors maintain interactional grounding by guiding the session forward, preserving the structure of therapeutic engagement. Synthetic counterparts often deliver flat acknowledgments like "Thanks for sharing," which, while polite, offer little in terms of affective co-regulation or forward momentum. This qualitative analysis reinforces our quantitative findings: although LLMs can simulate affectively expressive language, their responses may lack the situational nuance, relational grounding, and interactive adaptability found in real therapeutic interactions.

6.1.7 Discussion

Synthetic sessions are generally fluent and well-structured, yet the emotional expressions of speakers still differ in subtle but important ways from those observed in real therapy interactions. In the following, we discuss several important distinctions in the emotional dynamics of real and synthetic CBT dialogues based on our analysis.

First, synthetic dialogues tend to exhibit higher overall emotion means, especially in arousal and valence. This phenomenon suggests that synthetic speakers may adopt a more emotionally elevated or expressive tone. This could be a result of LLMs overemphasizing affective expression to appear engaged or empathetic. In contrast, real counselors and clients demonstrated more restrained emotional expression, aligning with the grounded and calibrated communication style expected in professional therapy.

Second, emotion variability serves as an indica-

tor of emotional richness in dialogue. We observed consistently higher variability in real sessions, particularly in arousal and valence. This suggests that real speakers fluctuate more in their emotional tone throughout the conversation, reflecting more authentic, responsive, and context-sensitive affective expression. In contrast, synthetic dialogues tend to be more emotionally uniform, which may result from LLMs generating emotionally consistent responses without deeply modeling the evolving emotional context.

Notably, this difference in variability remains even though the synthetic dataset is ten times larger than the real dataset (580 vs. 58 dialogues). Despite the increased diversity that might be expected from a larger sample, the real dataset still exhibits greater emotional variability. This suggests that current LLMs are struggling to replicate emotion dynamics in real therapy sessions.

Third, displacement length analysis further supports this observation: real speakers used emotion-laden language more dynamically, especially in arousal and dominance dimensions. This suggests that real speakers vary their emotional emphasis in different parts of the session, whereas synthetic speakers may be more emotionally formulaic or flat in structure.

Fourth, in terms of emotional reactivity, synthetic speakers, particularly clients, exhibited faster and more intense shifts in emotion in dominance and valence. While this may appear to reflect emotional engagement, it might also signal exaggerated or less nuanced affective responses, possibly resulting from LLMs optimizing for expressive content over authentic regulation.

Relatedly, emotion recovery rate, reflecting emotion regulation, was largely comparable across real and synthetic sessions, though synthetic clients demonstrated stronger emotion regulation in valence and dominance. This pattern might again reflect stylized generation patterns that smooth or suppress emotional volatility, possibly at the cost of naturalistic variation.

Taken together, these findings suggest that synthetic dialogues approximate but do not fully replicate the emotional dynamics of real therapy conversations. Synthetic counselors show relatively close alignment with real counselors across most metrics, indicating that LLMs are reasonably good at simulating professional emotional tone. In contrast, synthetic clients diverge more notably, particularly in reactivity and regulation, which may affect the authenticity of the simulated therapeutic process.

To explain the observed discrepancies in emotional arcs between real and LLM-generated therapy dialogues, we identify several possible underlying factors. First, current LLMs may not be trained on emotionally grounded or therapeutically contextualized data (Chung et al., 2023). General-purpose corpora lack the subtle affective trajectories and turn-by-turn emotion regulation seen in real counseling. Second, common prompting strategies often elicit exaggerated or uniform emotional expressions that fail to reflect the adaptive, client-contingent nature of real therapist responses (Nudo et al., 2025). Third, LLMs typically lack mechanisms for maintaining contextual coherence over multiple turns, limiting their ability to simulate affective progression across a session (Shu et al., 2025). Finally, real therapy involves ongoing co-regulation—therapists dynamically modulate their tone, intensity, and response based on client signals (Butler and Randall, 2013). This interactive nuance is missing in current LLM outputs, which treat utterances as isolated generations.

To mitigate these limitations, future work should explore fine-tuning LLMs on real counseling data to better capture authentic emotional arcs. Prompting strategies can also be refined to encourage adaptive rather than exaggerated affect. Moreover, integrating affect-aware generation and explicit co-regulatory modeling could allow LLMs to simulate more human-like emotional responsiveness and interactional flow. These directions offer promising steps toward enhancing the emotional fidelity of synthetic therapy data, which is critical for trustworthy use of LLMs in mental health applications.

6.2 Emotional Arc Similarity across Real and Synthetic Speakers

In this section, we investigate how closely the emotional trajectories of LLM-generated speakers align with those of real individuals.

6.2.1 Correlation Analysis

Building on the method by Vishnubhotla et al. (Vishnubhotla et al., 2024), we temporally align emotion arcs and compute Spearman correlations between pairs of sessions. To capture different patterns of alignment, we include three pair types for each speaker role: Real–Real, Syn–Syn, and Real–Syn. This allows us to compare emotional arc similarity within real sessions, within synthetic sessions, and across real and synthetic sessions.

Higher correlation values reflect stronger similarity in emotional progression. We report these comparisons separately for clients and counselors to highlight role-specific alignment dynamics (Table 2).

Client	**Real-Syn**	**Real-Real**	**Syn-Syn**
Valence Mean	0.014	0.0153	0.2151
Arousal Mean	0.020	0.0054	0.0225
Dominance Mean	0.002	-0.0047	0.0874
Counselor	**Real-Syn**	**Real-Real**	**Syn-Syn**
Valence Mean	0.044	0.0043	0.1365
Arousal Mean	-0.011	0.0077	0.0213
Dominance Mean	0.058	0.0278	0.1418

Table 2: Correlation Analysis for Arc Alignment

For clients, emotional arc similarity is weak across all pairings. Syn–Syn pairs show the highest values for valence (0.2151) and dominance (0.0874), but even these are modest, and arousal alignment remains negligible across all pair types. Both Real–Real and Real–Syn correlations are near zero or negative, indicating little consistency either within real sessions or between real and synthetic clients. These findings suggest that synthetic clients exhibit only limited internal consistency in their affective dynamics and fail to approximate the more complex, contingent emotion arcs of real clients.

For counselors, Syn–Syn correlations again yield the highest values, for valence (0.1365) and dominance (0.1418), but these are still weak. Real–Real and Real–Syn alignments are even lower, especially for valence and arousal. Notably, dominance alignment is slightly higher in Real–Syn pairs (0.058) than in Real–Real (0.0278), though the difference is small. Overall, these results indicate that LLM-generated counselors produce emotionally variable output, but without structured or relationally grounded patterns that resemble real therapeutic trajectories. Please see the three representative examples of emotion arc correlation between real and synthetic clients in Appendix A.5.

6.2.2 Discussion

Our quantitative analysis of emotional arc similarity reveals that alignment is consistently low across all session pairings, including Real–Real, Syn–Syn, and Real–Syn combinations. This outcome is particularly illuminating in several ways.

First, the low correlation between Real–Real pairs (e.g., 0.0153 for valence among clients and 0.0043 for counselors) reflects the natural heterogeneity of real-world therapy, where emotional trajectories are shaped by unique client issues, counselor styles, and the evolving goals of each session. In psychotherapy, affective expression is not scripted or uniform—rather, it is personalized, adaptive, and deeply contextual. Thus, low Real–Real correlation values serve as a baseline, not a shortcoming, and emphasize the richness and diversity of authentic human interactions.

By contrast, LLM-generated sessions show limited internal emotional consistency (e.g., Syn–Syn valence correlations of 0.2151 for clients and 0.1365 for counselors). While these scores are higher than those of Real–Real or Real–Syn pairs, they remain modest overall. This suggests that even when generating multiple synthetic sessions, current LLMs do not produce strongly consistent emotional arcs across speakers, pointing to a lack of coherent affective modeling.

Most notably, Real–Syn correlations are near zero across all emotion dimensions (e.g., valence = 0.014 for clients, 0.044 for counselors), indicating that synthetic speakers do not successfully emulate the affective trajectories of real speakers. This underscores a fundamental challenge: although LLMs can generate grammatically fluent and affectively expressive utterances, they struggle to reproduce the dynamic, situated, and role-sensitive emotional flow found in real therapy.

7 Conclusion

In this work, we conduct the first systematic comparison of emotional dynamics between real and LLM-generated CBT dialogues. Using the UED framework, we analyzed emotion mean and variability, reactivity, regulation, and arc similarity across counselors and clients. Our findings show that while synthetic dialogues are fluent and structured, they diverge from real therapy in key emotional dimensions—particularly in variability and emotional arc alignment. These results highlight the limitations of current LLMs in replicating the nuanced emotional flow of real therapeutic interactions, especially on the client side. This work introduces the RealCBT dataset alongside an in-depth analysis for evaluating affective realism in synthetic therapy, paving the way for future research on more emotionally grounded and psychologically credible dialogue generation.

8 Limitations

Our study has several limitations. First, RealCBT is limited in size and exhibits subgroup imbalance, which affects generalizability. These constraints stem from significant privacy, ethical, and legal challenges in collecting and sharing real-world CBT data. Unlike domains where large-scale conversational datasets are readily available, therapy transcripts are highly sensitive, and even small-scale public datasets are rare. Second, our analysis focuses exclusively on CBT sessions and comparisons with a single synthetic dataset (CACTUS). While CBT is widely practiced and structurally well-defined, making it a suitable starting point, our findings may not extend to other therapy modalities, mental health domains, or cultural and linguistic contexts. Nevertheless, RealCBT remains one of the few publicly available, authentic CBT datasets. Its inclusion enables a rare grounded comparison with synthetic dialogues and offers valuable insights into how LLM-generated therapy diverges from real interactions. By making RealCBT available, we hope to catalyze broader collaborative efforts toward building larger, more diverse, and ethically responsible resources that can better capture the complexity of therapeutic dialogue.

Looking forward, we outline plans to extend research with RealCBT in several directions:

- Expanding annotations to include dialogue acts, empathy markers, and turn-level emotional shifts;
- Training and evaluating affect-aware LLMs that model emotional dynamics more faithfully;
- Analyzing therapeutic strategies (e.g., reappraisal, validation) and their relationship to client emotional trajectories.
- Broadening scope across therapeutic approaches (e.g., MI, psychodynamic therapy), languages, and cultural contexts, while incorporating alternative synthetic datasets to test robustness across generative sources.

Together, these steps aim to strengthen the dataset's utility and support progress toward building LLM systems that are both emotionally grounded and clinically meaningful.

References

WHO Mental Disorders.

Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, and others. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204. Valletta. Number: 2010.

Judith S Beck, AT Beck, and JS Beck. 2011. Cognitive behavior therapy: basics and beyond. ed. *New York*.

Victoria Bobicev, Victoria Maxim, Tatiana Prodan, Natalia Burciu, and Victoria Angheluş. 2010. Emotions in words: Developing a multilingual wordnet-affect. In *Computational linguistics and intelligent text processing: 11th international conference, cicling 2010, iaşi, romania, march 21-27, 2010. Proceedings 11*, pages 375–384. Springer.

Edward S Bordin. 1979. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252. Publisher: Division of Psychotherapy (29), American Psychological Association.

Emily A Butler and Ashley K Randall. 2013. Emotional coregulation in close relationships. *Emotion Review*, 5(2):202–210. Publisher: Sage Publications Sage UK: London, England.

Kathleen M Carroll and Brian D Kiluk. 2017. Cognitive behavioral interventions for alcohol and drug use disorders: Through the stage model and back again. *Psychology of addictive behaviors*, 31(8):847. Publisher: American Psychological Association.

Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of large language models for mental health counseling. ArXiv: 2311.13857 [cs.CL].

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy (Chicago, Ill.)*, 55(4):399. Publisher: Educational Publishing Foundation.

Robert Elliott, Leslie S Greenberg, Jeanne C Watson, Ladislav Timulak, and Elizabeth Freire. 2013. Research on humanistic-experiential psychotherapies. *Bergin & Garfield's Handbook of psychotherapy and behavior change*, pages 495–538. Publisher: John Wiley & Sons Inc.

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3):271–299. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Will E. Hipson and Saif M. Mohammad. 2021. Emotion dynamics in movie dialogues. *PLOS ONE*, 16(9):1–19. Publisher: Public Library of Science.

Tom Hollenstein. 2015. This time, it's real: Affective flexibility, time scales, feedback loops, and the regulation of emotion. *Emotion Review*, 7:308 – 315.

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. Number: 1.

Peter Kuppens and Philippe Verduyn. 2017. Emotion dynamics. *Current Opinion in Psychology*, 17:22–26.

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. 2025. REALTALK: a 21-day real-world dataset for long-term conversation. ArXiv: 2502.13270 [cs.CL].

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Richard J Longmore and Michael Worrell. 2007. Do we need to challenge thoughts in cognitive behavior therapy? *Clinical psychology review*, 27(2):173–187. Publisher: Elsevier.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 13641–13660, Toronto, Canada. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif M. Mohammad. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741.

Saif M. Mohammad. 2025. NRC VAD lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. ArXiv: 2503.23547 [cs.CL].

Jacopo Nudo, Mario Edoardo Pandolfo, Edoardo Loru, Mattia Samory, Matteo Cinelli, and Walter Quattrociocchi. 2025. Generative exaggeration in LLM social agents: Consistency, bias, and toxicity. ArXiv: 2507.00657 [cs.HC].

Emily Ohman, Yuri Bizzoni, Pascale Feldkamp Moreira, and Kristoffer Nielbo. 2024. EmotionArcs: Emotion arcs for 9,000 literary texts. In *Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (LaTeCH-CLfL 2024)*, pages 51–66, St. Julians, Malta. Association for Computational Linguistics.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Bangzhao Shu, Isha Joshi, Melissa Karnaze, Anh C. Pham, Ishita Kakkar, Sindhu Kothe, Arpine Hovasapian, and Mai ElSherief. 2025. Fluent but unfeeling: The emotional blind spots of language models. ArXiv: 2509.09593 [cs.CL].

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. PsyQA: a Chinese dataset for generating long counseling text for mental health support. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online. Association for Computational Linguistics.

Yla R Tausczik and James W Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Daniela Teodorescu and Saif Mohammad. 2023. Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 4124–4137, Singapore. Association for Computational Linguistics.

Krishnapriya Vishnubhotla, Adam Hammond, Graeme Hirst, and Saif M Mohammad. 2024. The emotion dynamics of literary novels. *arXiv preprint arXiv:2403.02474*.

Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 4162–4176, Marseille, France. European Language Resources Association.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and

Jimin Huang. 2024. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.

Emily Öhman. 2021. The validity of lexicon-based sentiment analysis in interdisciplinary research. In *Proceedings of the workshop on natural language processing for digital humanities*, pages 7–12.

A Appendix

A.1 RealCBT Selection Criteria

We used keywords such as *CBT counseling*, *CBT case study*, *CBT role play*, *CBT session*, and *using CBT* to guide the search. To ensure that the selected videos accurately portrayed CBT-style therapy conversations, we applied the following inclusion criteria: (1) the video must feature exactly two participants—a counselor and a client; (2) the video should contain little to no narration or background music that could interfere with the dialogue; and (3) the session should focus on a behavioral or emotional issue consistent with CBT practice, such as depression, anxiety, or smoking cessation; (4) the therapy dialogue should last at least three minutes to ensure sufficient conversational context for analysis.

A.2 RealCBT Summary Statistics and Distributions

Table 3 provides an overview of the RealCBT dataset, while Table 4 presents its distribution across major categories. To highlight key patterns, Table 5 further summarizes the top three client issues represented in the dataset, and Table 6 shows their detailed distributions. Together, these tables offer a comprehensive view of both the overall dataset and the prevalence of the most common client problems.

A.3 UED Metrics in VAD across RealCBT and CACTUS

We adapt the Utterance Emotion Dynamics (UED) framework to compute emotion metrics from sequences of utterances. Table 7 reports the mean and variability of valence, arousal, and dominance (VAD) across RealCBT and CACTUS. More detailed results for each dimension are presented in Table 8 through Table 10. Together, these statistics highlight systematic differences between real and synthetic therapy dialogues, providing a basis for the comparative analyses in the paper.

A.4 Representative Examples of Emotional Arcs

This appendix provides representative examples of how emotional dynamics manifest in real and synthetic CBT dialogues. The examples span all three affective dimensions (valence, arousal, and dominance) for both counselors and clients. Selected cases, shown in Table 1 and Appendix Tables 11–15, illustrate how real and synthetic utterances differ in affective expression across varying intensity levels.

A.5 Representative Real–Synthetic Client Emotion Arc Correlation Examples

Figure 2 illustrates three representative examples of emotion arc correlation between real and synthetic clients: (a) high positive correlation (strong alignment), (b) near-zero correlation (no alignment), and (c) high negative correlation (strong misalignment).

These examples illustrate the range of relational patterns that emerge when comparing emotional trajectories across real and LLM-generated client dialogues.

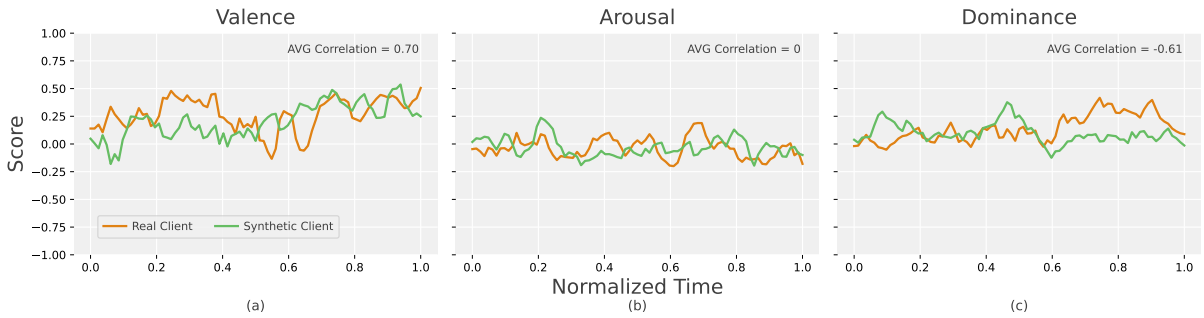

Figure 2: Emotion arcs of valence, arousal, and dominance for a client in three representative cases: (a) highest correlation, (b) near-zero correlation, and (c) lowest (negative) correlation between real and synthetic trajectories.

Number of sessions	76
Total client words count	82,436
Total counselor words count	108,278
Total word count	190,714
Average word count per session	2,516
Total Duration (min)	1,224.67
Avg. Duration (min)	16.11

Table 3: Descriptive statistics of **RealCBT** dataset

Category	**Subcategory**	**Proportion (%)**	**Count**
Client's Problem	Anxiety and fear	32.89	25
	Self-esteem and confidence issues	25.00	19
	Relationships (romantic, family, friendships)	18.42	14
	Career and work-related concerns	9.21	7
	Health-related worries	5.26	4
	Academic and educational concerns	3.95	3
	Financial concerns	2.63	2
	Health-related Worries	1.32	1
	Other miscellaneous concerns	1.32	1
Client's Attitude	Positive	90.79	69
	Neutral	6.58	5
	Negative	2.63	2
Client's Gender	Female	84.21	64
	Male	15.79	12
Total		100.00	76

Table 4: Distribution of the **RealCBT** dataset

Number of sessions	58
Total client words count	59,580
Total counselor words count	82,618
Total word count	142,198
Average word count per session	2,461
Total Duration (min)	907.18
Avg. Duration (min)	15.64

Table 5: Descriptive statistics of **RealCBT** dataset (Top 3 Client Problems)

Category	**Subcategory**	**Proportion (%)**	**Count**
Client's Problem	Anxiety and fear	43.10	25
	Self-esteem and confidence issues	32.76	19
	Relationships (romantic, family, friendships)	24.14	14
Client's Attitude	Positive	93.10	54
	Negative	3.45	2
	Neutral	3.45	2
Client's Gender	Female	87.93	51
	Male	12.07	7
Total		100.00	58

Table 6: Distribution of Top 3 Client Problems in the **RealCBT** dataset

Speaker	**Valence**		**Arousal**		**Dominance**	
	Mean	Var.	Mean	Var.	Mean	Var.
Real Dialogue	0.2334	0.1448	-0.0907	0.1161	0.0509	0.1131
Synthetic Dialogue	0.2761	0.1480	-0.0247	0.1037	0.1130	0.1154
Real Counselor	0.2302	0.1466	-0.0804	0.1148	0.0614	0.1120
Synthetic Counselor	0.3067	0.1287	-0.0240	0.0998	0.1550	0.1097
Real Client	0.2344	0.1312	-0.1006	0.1109	0.0346	0.1052
Synthetic Client	0.2339	0.1518	-0.0257	0.1012	0.0655	0.1043

Table 7: Emotion Mean and Variability in Valence, Arousal, and Dominance across RealCBT and CACTUS

metric//sperker type	Real Dialogue	Synthetic Dialogue	Real Counselor	Synthetic Counselor	Real Client	Synthetic Client
emo_mean	0.2334	0.2761	0.2302	0.3067	0.2344	0.2339
emo_std	0.1448	0.1480	0.1466	0.1287	0.1312	0.1518
emo_avg_peak_dist	0.0762	0.0779	0.0792	0.0695	0.0718	0.0818
emo_avg_disp_length	4.2100	4.2282	4.2996	3.9746	3.9735	4.0276
emo_rise_rate	0.0311	0.0332	0.0328	0.0310	0.0319	0.0349
emo_recovery_rate	0.0313	0.0325	0.0324	0.0304	0.0304	0.0369
emo_low_peak_dist	0.0896	0.0887	0.0914	0.0816	0.0838	0.0920
emo_low_disp_length	4.3939	4.2972	4.4628	4.3451	4.3002	4.0820
emo_low_rise_rate	0.0346	0.0356	0.0352	0.0321	0.0339	0.0381
emo_low_recovery_rate	0.0338	0.0352	0.0343	0.0328	0.0315	0.0379
emo_high_peak_dist	0.0666	0.0705	0.0705	0.0637	0.0664	0.0810
emo_high_disp_length	4.2266	4.3962	4.3971	4.0169	4.0794	4.6133
emo_high_rise_rate	0.0280	0.0312	0.0303	0.0304	0.0299	0.0321
emo_high_recovery_rate	0.0292	0.0301	0.0305	0.0285	0.0297	0.0367

Table 8: UED Metrics in Valence across RealCBT and CACTUS

metric//sperker type	Real Dialogue	Synthetic Dialogue	Real Counselor	Synthetic Counselor	Real Client	Synthetic Client
emo_mean	-0.0907	-0.0247	-0.0804	-0.0240	-0.1006	-0.0257
emo_std	0.1161	0.1037	0.1148	0.0998	0.1109	0.1012
emo_avg_peak_dist	0.0618	0.0557	0.0590	0.0555	0.0616	0.0564
emo_avg_disp_length	4.1024	3.6634	3.8828	3.5889	4.1044	3.5794
emo_rise_rate	0.0260	0.0252	0.0254	0.0258	0.0270	0.0262
emo_recovery_rate	0.0261	0.0254	0.0261	0.0260	0.0259	0.0262
emo_low_peak_dist	0.0526	0.0531	0.0516	0.0536	0.0542	0.0546
emo_low_disp_length	4.1279	3.7331	3.8291	3.7313	4.3172	3.6903
emo_low_rise_rate	0.0232	0.0244	0.0236	0.0249	0.0240	0.0248
emo_low_recovery_rate	0.0231	0.0243	0.0240	0.0248	0.0248	0.0253
emo_high_peak_dist	0.0732	0.0605	0.0723	0.0617	0.0727	0.0634
emo_high_disp_length	4.2461	3.7695	4.4102	3.8029	4.2717	3.8465
emo_high_rise_rate	0.0290	0.0263	0.0282	0.0270	0.0302	0.0278
emo_high_recovery_rate	0.0293	0.0268	0.0282	0.0274	0.0275	0.0273

Table 9: UED Metrics in Arousal across RealCBT and CACTUS

metric//sperker type	Real Dialogue	Synthetic Dialogue	Real Counselor	Synthetic Counselor	Real Client	Synthetic Client
emo_mean	0.0509	0.1130	0.0614	0.1550	0.0346	0.0655
emo_std	0.1131	0.1154	0.1120	0.1097	0.1052	0.1043
emo_avg_peak_dist	0.0608	0.0612	0.0595	0.0592	0.0588	0.0579
emo_avg_disp_length	4.3631	4.0982	4.1802	4.0086	4.2093	3.7704
emo_rise_rate	0.0240	0.0255	0.0247	0.0254	0.0240	0.0262
emo_recovery_rate	0.0239	0.0256	0.0250	0.0248	0.0247	0.0263
emo_low_peak_dist	0.0619	0.0645	0.0614	0.0655	0.0610	0.0619
emo_low_disp_length	4.3842	4.2835	4.2933	4.3020	4.3629	4.0320
emo_low_rise_rate	0.0247	0.0260	0.0243	0.0268	0.0237	0.0269
emo_low_recovery_rate	0.0240	0.0260	0.0247	0.0252	0.0244	0.0262
emo_high_peak_dist	0.0628	0.0609	0.0609	0.0587	0.0605	0.0604
emo_high_disp_length	4.5743	4.1729	4.4043	4.2198	4.4529	4.0397
emo_high_rise_rate	0.0237	0.0250	0.0250	0.0242	0.0242	0.0261
emo_high_recovery_rate	0.0241	0.0254	0.0255	0.0247	0.0251	0.0266

Table 10: UED Metrics in Dominance across RealCBT and CACTUS

Valence Level	Real Counselor Snippet	Synthetic Counselor Snippet
High	"Okay. And if you continue to feel like you have to be perfect in a presentation, you know, how do you think that will affect your future in this career?" *Explanation*: Connects the client's concern to long-term goals, blending positive framing with constructive challenge. This reflects emotion regulation and supports the therapeutic alliance through future-oriented collaboration.	"That's great to hear, Andrew. As a starting point, let's work on identifying and challenging these thoughts when they arise." *Explanation*: Provides encouragement but in a generic way, lacking calibration to the client's immediate narrative. This reduces opportunities for co-regulation.
Medium	"Let me ask you first. Logically, if you made a mistake in a presentation, would that make you feel like you were always going to make mistakes in presentations to follow?" *Explanation*: Uses logical probing to reduce overgeneralization, showing calibrated empathy by validating but gently challenging the client. Builds the therapeutic alliance through guided reasoning.	"Understandably, strong feelings can be difficult to manage even when we recognize them as exaggerated." *Explanation*: Normalizes emotions broadly but remains detached from the client's specific context, limiting opportunities for emotion regulation or alliance building.
Low	"Where do you feel like this fear comes from? Is there an instance that occurred that caused you to be afraid or." *Explanation*: Invites exploration of root causes, a low-valence but engaged stance that sustains interaction.	"Sometimes our mind can amplify fears beyond what's likely to happen. What might be some reasons or evidence that contradict this fear?" *Explanation*: Provides cognitive reframing with minimal emotional tone. While constructive, it lacks the relational grounding of calibrated questioning.

Table 11: Illustrative Counselor Utterances Comparison at Varying Valence Levels

Dominance Level	Real Counselor Snippet	Synthetic Counselor Snippet
High	"You've been wrestling for three years, and it is a sport you love participating in. Does it make sense that you said you were going to quit just a few matches into the season?" *Explanation*: Directly confronts the client's inconsistency, showing high directive control and authority in guiding the dialogue.	"I'm glad to hear you are open to it. Let's continue to explore and challenge these beliefs together." *Explanation*: Encourages collaboration but avoids strong challenge, reflecting lower dominance and shared control.
Medium	"Okay, so losing is not fun. No one likes to lose. But it doesn't mean you're a bad wrestler." *Explanation*: Reframes the issue and provides reassurance while still shaping the client's perspective — moderate counselor dominance.	"That sounds really challenging. It must be hard to feel like you're constantly judged. Can you recall any recent instances where you felt this way?" *Explanation*: Invites the client to elaborate on experiences, giving more control to the client and reducing directive influence.
Low	"Let me stop you right there. What if you do lose your next match? What is the worst thing that could happen?" *Explanation*: Uses hypothetical exploration to reduce anxiety; less directive than a firm challenge, reflecting lower dominance.	"Let's break it down step by step: When you noticed people staring, what specifically did you observe, and what did you think led to those thoughts and feelings?" *Explanation*: Breaks the concern into steps, but with minimal counselor control, shifting responsibility to the client – lowest dominance.

Table 12: Illustrative Counselor Utterances Comparison at Varying Dominance Levels

Valence Level	Real Client Snippet	Synthetic Client Snippet
High	"Not in a loving way, I guess, but that you should do your best, but also get the best. And I agree with that. I think that's the right thing to think." *Explanation*: Expresses agreement through reflective reasoning, positive but context-grounded.	"I can do that. I'll keep a record of those moments and I'm looking forward to our next session." *Explanation*: Optimistic and cooperative, but formulaic and less tied to personal context.
Medium	"No, because I said their opinion. It was just in that time and place that I was stupid." *Explanation*: Self-critical and context-specific, showing moderate negative valence.	"Hi, I'm doing okay, I guess. Just really anxious about some things." *Explanation*: Mild disclosure of anxiety, less intense and more generic in tone.
Low	"Not really? I've always disliked it, and I've had bad experiences that probably made my fear worse, but it's. It's always been somewhat of a fear of mine since I was a child." *Explanation*: Reflects long-term fear, personally grounded in past experiences, conveying sustained low valence.	"I just had this fear that I would end up feeling deprived or unsatisfied, like I'd somehow starve because I couldn't satisfy my cravings." *Explanation*: Negative affect framed in exaggerated hypotheticals, less grounded in lived history.

Table 13: Illustrative Client Utterances Comparison at Varying Valence Levels

Arousal Level	Real Client Snippet	Synthetic Client Snippet
High	"I really want to go out with y'all, but I just feel like I need to work on my. On my studies, so. I'm sorry." *Explanation*: Expresses conflict and relational tension, revealing socially grounded high arousal.	"I had invested a substantial amount of money, so the stress was pretty high right from the start." *Explanation*: Conveys stress directly but in a more factual, less relational manner.
Medium	"It's like peer pressure, sort of. But I can't say no. I feel like I have to say yes just. Just because they are my friends. I do care about them and I don't want to lose them." *Explanation*: Shows interpersonal anxiety and fear of rejection, reflecting moderate arousal.	"Maybe I could see it as a learning experience rather than a total failure. It doesn't cancel out my other achievements, right?" *Explanation*: Rational reframing of failure, affectively calmer and less pressured than the real example.
Low	"So I just kind of want to build up my self esteem and my confidence and accepting myself." *Explanation*: Calm, constructive goal-setting, consistent with low arousal.	"Sure. I've been feeling really unsuccessful and criticized because I've been taking losses in the stock market. It's been affecting my mood and confidence a lot." *Explanation*: Expresses discouragement with subdued tone.

Table 14: Illustrative Client Utterances Comparison at Varying Arousal Levels

Dominance Level	Real Client Snippet	Synthetic Client Snippet
High	"My parents don't really have the money to pay for my college, so I need to get good grades and do well with wrestling in order to have a better chance at getting a scholarship." *Explanation*: Shows strong agency and responsibility, reflecting high dominance.	"Thanks for having me. Well, I've been struggling with feeling judged because of my tattoos. It's really been eating at me, especially in social situations." *Explanation*: Centers on external judgment, revealing less personal control.
Medium	"I think about it all the time, which makes it hard for me to focus." *Explanation*: Rumination disrupts focus, showing partial loss of control.	"I think it would be helpful. I really don't want to feel like this every time I mess up." *Explanation*: Expresses a wish to change, indicating tentative but emerging agency.
Low	"I got so mad about losing that I wasn't thinking straight." *Explanation*: Behavior dictated by anger, emotions override reasoning and control.	"Yes, just the other day I was at the grocery store, and I caught a few people staring at me. It made me so uncomfortable that I left without buying everything I needed." *Explanation*: Actions shaped by others' perceptions, showing externally driven loss of control.

Table 15: Illustrative Client Utterances Comparison at Varying Dominance Levels