How Real Are Synthetic Therapy Conversations? Evaluating Fidelity in Prolonged Exposure Dialogues

Suhas BN[1] **Dominik Mattioli**[1] **Andrew M. Sherrill**[2]
Rosa I. Arriaga[3] Chris W. Wiese[4] Saeed Abdullah[1]
[1]College of Information Sciences and Technology, Penn State University, USA
[2]Department of Psychiatry and Behavioral Sciences, Emory University, USA
[3]School of Interactive Computing, Georgia Tech, USA
[4]School of Psychology, Georgia Tech, USA
{bnsuhas,saeed}@psu.edu, andrew.m.sherrill@emory.edu

Abstract

Synthetic data adoption in healthcare is driven by privacy concerns, data access limitations, and high annotation costs. We explore synthetic Prolonged Exposure (PE) therapy conversations for PTSD as a scalable alternative for training clinical models. We systematically compare real and synthetic dialogues using linguistic, structural, and protocol-specific metrics like turn-taking and treatment fidelity. We introduce and evaluate PE-specific metrics, offering a novel framework for assessing clinical fidelity beyond surface fluency. Our findings show that while synthetic data successfully mitigates data scarcity and protects privacy, capturing the most subtle therapeutic dynamics remains a complex challenge. Synthetic dialogues successfully replicate key linguistic features of real conversations, for instance, achieving a similar Readability Score (89.2 vs. 88.1), while showing differences in some key fidelity markers like distress monitoring. This comparison highlights the need for fidelity-aware metrics that go beyond surface fluency to identify clinically significant nuances. Our model-agnostic framework is a critical tool for developers and clinicians to benchmark generative model fidelity before deployment in sensitive applications. Our findings help clarify where synthetic data can effectively complement real-world datasets, while also identifying areas for future refinement.

1 Introduction

Training machine learning models in sensitive domains like healthcare remains a challenge (Giuffrè and Shung, 2023; Kokosi and Harron, 2022). Access to real clinical conversations, which are crucial for modeling tasks like mental health diagnostics and therapeutic dialogue understanding, is severely limited by high annotation costs, patient privacy concerns (Kokosi and Harron, 2022; BN and Abdullah, 2022; BN et al., 2023), and ethical constraints on data sharing. Synthetic data has emerged as a promising alternative, offering scalability and privacy preservation while reducing dependence on real-world annotations (Aher et al., 2023).

In trauma-focused mental health care, particularly Prolonged Exposure (PE) therapy for PTSD, large language models (LLMs) can generate synthetic therapy dialogues at scale. However, questions remain about whether these dialogues capture more than surface-level fluency, specifically the subtle dynamics of therapeutic fidelity such as emotional pacing, avoidance management, and protocol adherence (Shen et al., 2024). To address this, we develop and validate methods to measure clinically relevant fidelity in generated dialogues, moving beyond standard metrics like coherence or perplexity. Prior work shows that while synthetic PE sessions can convincingly mimic real sessions in tone and structure (BN et al., 2025b), they may still commit fidelity lapses, such as premature reflection or reinforcement of avoidance, which often go unnoticed by both automatic metrics and non-clinical annotators (Chiu et al., 2024; Zhang et al., 2024; Lee et al., 2024b).

For instance, a therapist saying *"That's a really powerful insight"* mid-exposure may appear empathic but violates PE protocol by derailing trauma processing (see Fig. 1). Without clinical expertise, both humans and automated metrics tend to overestimate fidelity.

We introduce a fidelity-focused lens that evaluates dialogues on multiple dimensions: linguistic coherence, adherence to PE protocols, and the therapist's navigation of key clinical interactions (e.g., managing avoidance, SUDS (Subjective Units of Distress) monitoring). This framework integrates automated scoring and expert clinical assessment for a more rigorous evaluation of synthetic dialogue quality.

We present the first large-scale, multi-dimensional comparison of real and synthetic

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20986–20995
November 4-9, 2025

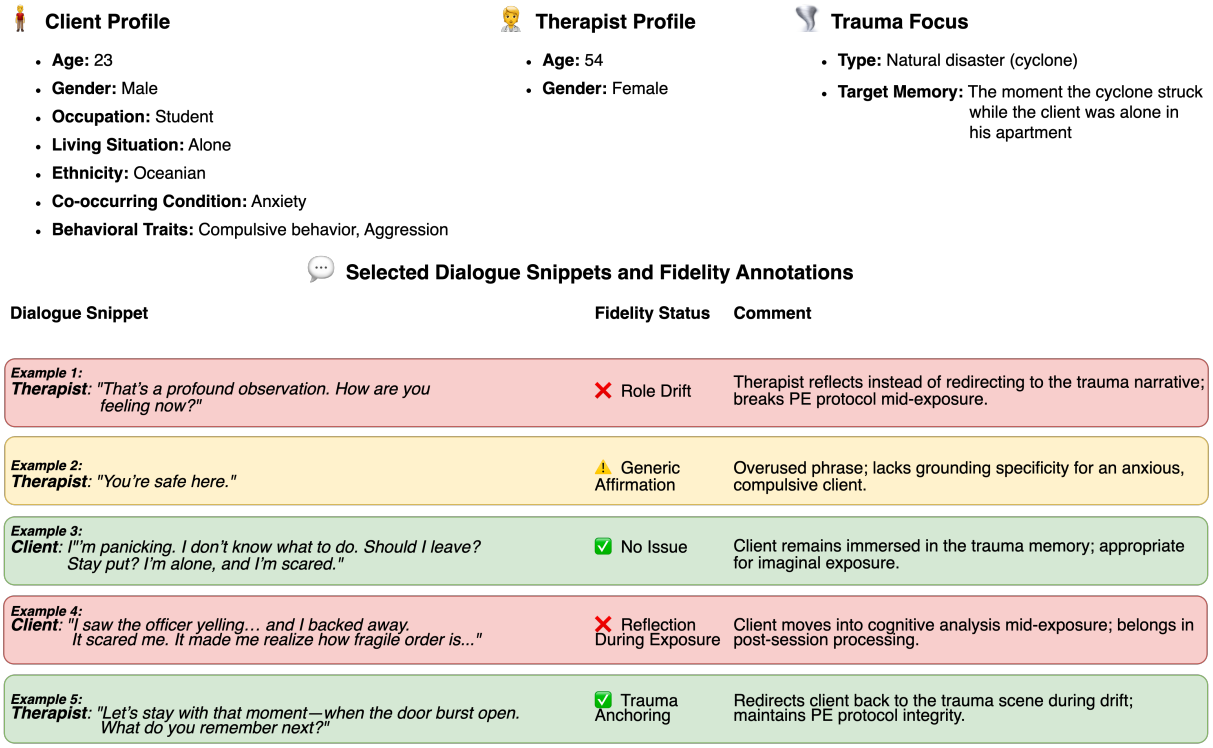

Figure 1: Selected annotated examples from synthetic PE therapy sessions. Each dialogue segment is evaluated for fidelity adherence based on PE protocol guidelines. Despite structural fluency, subtle violations like role drift (Ex. 1) and premature reflection (Ex. 4) highlight limitations in fidelity that escape automated scoring.

PE therapy sessions, analyzing their linguistic, structural, and clinical fidelity characteristics. Our findings demonstrate the current strengths and limitations of synthetic dialogues and inform future improvements in generation and evaluation methods. Our study contributes to bridging the real-synthetic data gap and provides a roadmap for advancing synthetic dialogue in sensitive healthcare domains.

2 Motivation

Developing synthetic clinical dialogues is essential to overcome persistent barriers in mental health AI. In the context of Prolonged Exposure (PE) therapy, four challenges stand out: (1) **data scarcity**, due to the high cost and effort of annotation; (2) **privacy constraints**, which limit access to sensitive patient narratives; (3) **lack of diversity**, with datasets often failing to capture varied trauma types and demographics; and (4) **evaluation inconsistency**, as fidelity assessment lacks standardized benchmarks.

This work directly addresses these issues by generating and evaluating synthetic PE dialogues that preserve protocol fidelity, enable scalable annotation, and support more equitable and robust training data for mental health applications.

3 Related Work

While synthetic clinical conversation datasets have advanced evaluations for general counseling and CBT, work specific to Prolonged Exposure (PE) and structured trauma-focused therapy is limited. Most studies focus on (1) synthetic dataset generation, (2) evaluation metrics beyond lexical similarity, and (3) human-in-the-loop validation, but generally lack PE-specific considerations.

3.1 Synthetic Dataset Generation and Evaluation

Recent work has explored synthetic clinical dialogue generation and assessment. BOLT evaluates LLM-generated therapist behaviors in general counseling (Chiu et al., 2024), while SimPsyDial benchmarks synthetic data using the Working Alliance Inventory (Qiu and Lan, 2024). CPsyCoun reconstructs dialogues for evaluation with BERTScore, GPTScore, and qualitative review (Zhang et al., 2024, 2019). Other studies focus on counselor style (Xie et al., 2024) and data augmentation (Kim et al., 2024). While The Pursuit of Empathy highlights the importance of empathy-aware evaluation (BN et al., 2025a), Thousand Voices of Trauma introduces structural variation

(BN et al., 2025b), but most evaluations compare to general counseling or CBT (Lee et al., 2024a; Shen et al., 2024) and rarely assess trauma-specific markers or PE's structural fidelity, such as avoidance handling and imaginal exposure sequences. Our work directly addresses this by evaluating whether synthetic PE sessions capture clinical realism per PE protocols (e.g., SUDS, avoidance redirection), highlighting procedural errors missed by standard metrics.

3.2 Beyond Lexical Metrics

Evaluation is shifting from lexical metrics like BLEU (Papineni et al., 2002) to semantic ones such as BERTScore and GPTScore (Zhang et al., 2024; Xie et al., 2024). Psychological measures (CTRS, PANAS) in COUNSELINGEVAL assess empathy and coherence (Lee et al., 2024a), and interactional features like PQA are explored (Shen et al., 2024). However, most metrics are not adapted for trauma therapy, and methods like Dynamic Time Warping (DTW), which are potentially useful for structured PE flows, remain underused. While some qualitative reviews consider emotional tone and goal alignment (Zhang et al., 2024; Xie et al., 2024), and structured questioning is explored (Ren et al., 2024), robust frameworks for quantitative evaluation of PE-specific structure are lacking.

3.3 Human-in-the-Loop Validation and Engagement Metrics

Clinician validation is now common, as in COUNSELINGEVAL (Lee et al., 2024a), PsyDT (Xie et al., 2024), and CPsyCoun (Zhang et al., 2024), which assess quality, empathy, and safety. Metrics for engagement and personalization are also used (Lee et al., 2024a; Zhang et al., 2024; Xie et al., 2024). Prior work covers empathy (Morris et al., 2018) and structured workflows (Ren et al., 2024). However, human evaluation remains focused on general counseling or CBT, rarely addressing PE-specific components like trauma cue processing or protocol adherence compared to real PE data.

3.4 Bridging the Gap in PE Therapy Evaluation

PE is a first-line treatment for PTSD (Sherrill and Rauch, 2019; Varkovitzky et al., 2018; Rauch et al., 2021; Ragsdale et al., 2020; Evans et al., 2020; Yasinski et al., 2017). Despite progress in synthetic dialogue generation (Qiu and Lan, 2024; Zhang et al., 2024; Xie et al., 2024), evaluation metrics (Lee et al., 2024a; Zhang et al., 2024; Xie et al., 2024), and human validation, a gap remains in fidelity assessment for trauma-focused therapies like PE (Chiu et al., 2024; Qiu and Lan, 2024; Lee et al., 2024a; Shen et al., 2024; Zhang et al., 2024; Xie et al., 2024; Hu et al., 2024). Current approaches lack metrics and focus tailored to PE's unique conversational structure, including avoidance handling, trauma cue processing, and imaginal exposure flow, and they do not address challenges such as patient emotional dysregulation (Qiu and Lan, 2024; Lee et al., 2024a; Shen et al., 2024; Zhang et al., 2024; Xie et al., 2024; Kim et al., 2024). Our work systematically evaluates synthetic data for alignment with PE structure and clinical validity, identifying areas for improving synthetic PE dialogue quality (see Table 1, Appendix A).

4 Methodology

We analyzed 400 prolonged exposure therapy conversations, consisting of 200 real-world and 200 synthetic. To demonstrate our evaluation framework, this case study uses 200 synthetic conversations from the publicly available Thousand Voices of Trauma dataset (BN et al., 2025b). The data was generated using Claude Sonnet 3.5 (Anthropic, 2024) with PE-specific prompting frameworks adapted from clinical guidelines. This dataset includes diverse therapist-client interactions across trauma types, therapy phases, and demographics.

The real-world PE sessions were collected under IRB-approved protocols, with participant consent, and cannot be released due to privacy constraints. Each session (1-1.5 hours) was transcribed using Amazon HealthScribe (Services, 2023), manually verified, and reformatted by merging consecutive speaker turns for readability. Each conversation followed a validated therapy fidelity checklist (See Table 1) to align with real-world standards. Both datasets underwent the same preprocessing:

1. Standardized formatting for consistency.
2. Processing through ModernBERT (Warner et al., 2024) for analysis.
3. Removal of non-verbal cues (e.g., pauses, laughter) to focus on dialogue, with plans to incorporate these in future work on emotional speech.

All evaluation code, pre-processing pipelines, and metric definitions will be released to support reproducibility and external validation.

Metric	Yes	No	N/A
Therapist explained rationale for imaginal?	☐	☐	☐
Therapist gave client instructions to carry out imaginal?	☐	☐	☐
Hotspots procedure and rationale introduced?	☐	☐	☐
Therapist helped patient to identify hotspots?	☐	☐	☐
Therapist oriented the client to imaginal planned for that session?	☐	☐	☐
Therapist monitored SUDS ratings about every 5 minutes?	☐	☐	☐
Therapist used appropriate reinforcing comments during imaginal?	☐	☐	☐
Therapist elicited thoughts and feelings as appropriate?	☐	☐	☐
Therapist prompted for present tense, closed eyes?	☐	☐	☐
Imaginal lasted about 30-45 minutes (or about 15 for final imaginal)?	☐	☐	☐
Therapist processed the imaginal with client?	☐	☐	☐

Table 1: Clinically validated fidelity metrics for therapist adherence to essential elements of imaginal exposure therapy (Foa et al., 2007; Powers et al., 2010; Rauch et al., 2009; Hembree et al., 2003). This checklist evaluates whether therapists consistently implement key procedural components, including providing rationale, guiding the client through the exposure process, monitoring distress levels, and reinforcing engagement. Each element is rated as 'Yes,' 'No,' or 'N/A' to ensure treatment fidelity, maintain therapeutic consistency, and identify areas for improvement in clinical practice.

4.1 Metrics and Analysis

To systematically compare real and synthetic therapeutic conversations, we selected a diverse set of linguistic, structural, and statistical metrics. These metrics provide insights into conversational dynamics, protocol adherence, and overall fidelity, ensuring a holistic evaluation of synthetic dialogue generation. To reduce subjectivity, we adopt a fidelity checklist (Table 1) and PE-specific metrics grounded in existing clinical guidelines for evaluating therapeutic adherence. Our methodology consists of four key analyses: system-level metrics comparison, correlation analysis, statistical significance testing, and PE-specific metrics.

4.1.1 System-Level Metrics Comparison

We begin by measuring fundamental characteristics of the conversations, including turn-taking patterns, verbosity, lexical diversity, and readability (see Table 2). Key metrics include:

1. Turn-taking dynamics: Metrics such as Normalized Speaker Switches, Therapist-Client Turn Ratio, and Normalized Turn Duration capture the natural flow of conversation. These are essential for evaluating whether synthetic dialogues mimic real-world engagement.
2. Linguistic complexity and coherence: Average Utterance Length, Utterance Length Std Dev, and Readability Score assess how natural and readable the synthetic text is. Significant deviations indicate poor coherence.
3. Lexical richness: Vocabulary Richness would help quantify lexical variety, providing insight into whether synthetic dialogues are overly repetitive.
4. Predictability of text: Flow Entropy and Perplexity can help measure randomness and fluency, determining whether the synthetic text is overly structured or unnatural.

4.1.2 Correlation Analysis

While mean comparisons provide a general overview, correlation analysis (see Table 2) quantifies the consistency of relationships across metrics between real and synthetic data.

4.1.3 Statistical Significance Testing

To confirm whether differences between real and synthetic data are statistically meaningful, we conduct Mann-Whitney U tests (Table 3).

To identify which properties most strongly differentiate real and synthetic sessions, we next examine feature importance from a simple predictive model (see Table 4).

4.1.4 PE-Specific Metrics

PE therapy reduces pathological fear through repeated trauma exposure. To assess the fidelity of synthetic PE sessions, we evaluate key therapeutic constructs (See Appendix A for definitions and Table 5): *Trauma Narrative Coherence* measures how structured and detailed a client's trauma account is, reflecting cognitive integration. *Emotional Engagement* captures the level of emotional expression, linked to better outcomes. *Avoidance Handling* evaluates how well avoidance behaviors are addressed. *Exposure Guidance* assesses the therapist's role in structuring effective exposure exer-

cises. *Cognitive Restructuring* tracks how clients challenge maladaptive beliefs. *Emotional Habituation* and *SUDS Progression* measure distress reduction over repeated exposures. *Avoidance Reduction* quantifies improvements in engaging with trauma-related content. *Emotion Intensity* assesses the variability and magnitude of emotional responses. These metrics are derived using linguistic analysis, semantic modeling, and interaction patterns.

4.1.5 Qualitative Fidelity Assessment

To supplement automated evaluation metrics, we conducted a manual review of synthetic dialogues, annotating exchanges for fidelity adherence or violations using established PE clinical guidelines. Manual fidelity annotations were then reviewed by a licensed clinical psychotherapist for adherence to PE protocol. While inter-rater agreement was not computed in this exploratory phase, these annotations served as a qualitative tool to identify and illustrate typical fidelity lapses, especially those potentially overlooked by automated metrics, rather than as quantitative ground truth. Representative examples of such lapses identified through this process are illustrated in Figure 1 and discussed in Section 6.2.

5 Findings

5.1 System-Level Metrics & Correlation

The system-level metrics comparison highlights alignment between real and synthetic dialogues, revealing both structural matches and model-driven differences. While turn-taking patterns (e.g., Normalized Speaker Switches, Therapist-Client Turn Ratio, and Normalized Therapist/Client Turns) remain similar, variations in utterance length, lexical diversity, and entropy-based measures arise due to PE therapy and model constraints.

1. Turn-taking fidelity: Synthetic data closely aligns with real data in Normalized Speaker Switches (0.98 vs. 0.99) and Therapist-Client Turn Ratio (0.01 vs. 0.01), preserving conversational structure. Some deviations occur as real therapy sessions involve extended client turns, which LLMs struggle to maintain due to context limitations.

2. Concise and structured responses: Synthetic dialogues are shorter (Average Utterance Length: 22.90 ± 1.74 vs. 68.72 ± 26.61) and more consistent (Utterance Length Std Dev: 18.54 ± 2.35 vs. 135.85 ± 66.25) due to LLM output constraints. Techniques like Chain-of-Thought prompting help improve coherence, though larger output contexts are needed.

3. Vocabulary Richness is marginally higher in synthetic data but may reflect repeated paraphrasing rather than authentic diversity. Real therapy involves longer client responses, naturally increasing lexical variety, but synthetic responses remain contextually appropriate.

4. Increased structural consistency: Higher Perplexity (21.22 vs. 14.73) and lower Flow Entropy (1.06 vs. 1.30) suggest a structured, and predictable flow. While real data exhibit spontaneity, synthetic dialogues maintain stability, benefiting structured evaluations.

5. Correlation analysis: High correlations in turn-taking metrics (0.65-0.78 synthetic vs. 0.85-0.92 real) confirm accurate conversational dynamics. Lower correlations in linguistic complexity metrics (e.g., Readability Score, Vocabulary Richness, Perplexity) reflect structural differences but do not hinder the coherence/naturalness of the conversation.

Table 2: Comparative analysis of real and synthetic data across multiple system-level metrics and their correlation values. The first two columns display the mean $\pm$ standard deviation for each metric, computed separately for real and synthetic datasets. The third and fourth columns provide the correlation values of these metrics within the real and synthetic datasets, respectively.

Metric	Mean ± SD		Correlation	
	Real	Synth.	Real	Synth.
Norm. Spkr. Switches	0.99 ± 0.0	0.98 ± 0.0	0.85	0.72
Norm. Total Turns	1.00 ± 0.0	1.00 ± 0.0	0.78	0.65
Avg. Utt. Len.	68.7 ± 26.6	22.9 ± 1.7	0.91	0.76
Utt. Length SD	135.9 ± 66.2	18.5 ± 2.3	0.89	0.74
Norm. Avg. Turn Dur.	0.69 ± 0.6	0.12 ± 0.0	0.82	0.67
Norm. Turn Dur. SD	1.38 ± 1.3	0.09 ± 0.0	0.79	0.64
Norm. T Turns	0.50 ± 0.0	0.51 ± 0.0	0.87	0.71
Norm. C Turns	0.50 ± 0.0	0.49 ± 0.0	0.86	0.70
Norm. T Words	21.9 ± 13.2	4.9 ± 0.3	0.92	0.78
Norm. C Words	46.8 ± 22.5	18.0 ± 1.7	0.90	0.75
Turn Ratio (T/C)	0.01 ± 0.0	0.01 ± 0.0	0.80	0.66
Word Ratio (T/C)	0.01 ± 0.0	0.00 ± 0.0	0.82	0.68
Vocab. Richness	0.13 ± 0.0	0.18 ± 0.0	0.77	0.63
Readability Score	88.1 ± 4.6	89.2 ± 1.8	0.74	0.59
Flow Entropy	1.30 ± 0.1	1.06 ± 0.0	0.88	0.73
Avg. Perplexity	14.7 ± 2.3	21.2 ± 0.5	0.79	0.65

Note: Norm. = Normalized, Utt. = Utterance, SD = Standard Deviation, Dur. = Duration, T = Therapist, C = Client.

5.2 Statistical Significance Testing

The Mann-Whitney U test results confirm that many observed differences between real and synthetic dialogues are statistically significant ($p < 0.05$). However, these differences stem from model

design choices and practical constraints rather than fundamental shortcomings. Key observations include:

1. Distinct patterns in utterance structure: The test shows differences in utterance length, turn duration, and their standard deviations ($p < 10^{-17}$). This is expected, as synthetic dialogues are designed to maintain coherence by producing more structured and concise responses. In real PE therapy, clients occasionally have extended monologues, which LLMs struggle to handle due to context window limitations. To compensate, we shorten utterances while preserving the conversational structure. Methods like Chain-of-Thought prompting have improved this, but achieving full parity would require larger output contexts.

2. Lexical properties follow a structured pattern: Differences in Vocabulary Richness ($p < 10^{-15}$) and Flow Entropy ($p < 10^{-17}$) indicate that synthetic data exhibits a more varied vocabulary but within a constrained framework. This is a direct result of the model prioritizing coherence and avoiding redundant expressions. While real conversations naturally contain more spontaneity, the synthetic approach ensures stability in generated dialogue while maintaining conversational depth.

3. Certain aspects remain comparable: Metrics such as Normalized Total Turns ($p = 1.00$) and Readability Score ($p = 0.28$) show no significant differences, meaning that despite shorter utterances, the number of conversational exchanges and overall readability remain aligned with real data. This suggests that while individual responses may be more concise, the overall flow and engagement in the conversation are well-preserved.

For example, while a real client might repeatedly use an informal phrase like "freaked out," the LLM generates diverse but semantically similar alternatives such as "felt anxious," "was overcome with fear," or "grew very distressed," thereby increasing lexical richness within a tight, clinically-appropriate boundary.

These findings reinforce that the synthetic model captures key conversational characteristics while ensuring structured, coherent responses. While differences exist, they align with known model constraints and do not compromise the overall integrity of the generated dialogues.

Table 3: Mann-Whitney U test statistics and p-values comparing real and synthetic datasets.

Metric	Statistic	p-value
Norm. Speaker Switches	2.45×10^3	$p < 0.001$
Norm. Total Turns	1.25×10^3	1.00
Norm. Conv. Length	2.50×10^3	$p < 0.001$
Avg. Utt. Length	2.50×10^3	$p < 0.001$
Utt. Length Std	2.50×10^3	$p < 0.001$
Norm. Turn Duration	2.49×10^3	$p < 0.001$
Norm. Turn Dur. SD	2.50×10^3	$p < 0.001$
Norm. T Turns	0.00	$p < 0.001$
Norm. C Turns	2.50×10^3	$p < 0.001$
Norm. T Words	2.50×10^3	$p < 0.001$
Norm. C Words	2.47×10^3	$p < 0.001$
T-C Turn Ratio	2.13×10^3	1.22×10^{-9}
T-C Word Ratio	2.37×10^3	$p < 0.001$
Vocabulary Richness	8.60×10^1	$p < 0.001$
Readability Score	1.09×10^3	0.28
Semantic Coherence	1.12×10^3	0.37
Sem. Coherence Std	2.50×10^3	$p < 0.001$
Flow Entropy	2.49×10^3	$p < 0.001$
Avg. Perplexity	0.00	$p < 0.001$
Local Coherence	1.12×10^3	0.37
Coherence Std	2.50×10^3	$p < 0.001$

Note: Norm. = Normalized, Conv. = Conversation, Utt. = Utterance, SD = Std. Dev, Dur. = Duration, T = Therapist, C = Client, Sem. = Semantic. $p < 0.001$ indicates $p < 10^{-10}$.

5.3 Feature Importance

Feature importance analysis (see Table 4) identifies the metrics most influencing synthetic data generation, confirming trends from previous evaluations. While not the central focus, these results support the model's alignment with conversational structure and linguistic patterns.

1. **Turn-taking features are most influential:** Metrics such as Normalized Speaker Switches and Therapist-Client Turn Ratio dominate, indicating the model effectively captures conversational flow and balanced exchanges. Strong correlations with real data suggest natural interaction patterns are preserved.
2. **Utterance length and entropy shape responses:** Average Utterance Length, Perplexity, and Flow Entropy are key distinguishing factors. Synthetic responses are more structured and predictable, prioritizing coherence and consistency. While often more concise, they capture essential interaction patterns, though further refinement is needed for richer emotional dynamics.
3. **Lexical richness is moderately important:** Synthetic dialogues use varied yet structured language, balancing vocabulary diversity with clarity, while real conversations are more flexible.

5.4 Prolonged Exposure Specific Metrics

Table 5 presents the statistical test results comparing synthetic Prolonged Exposure data with real therapy session data. The synthetic data shows promise, with no statistically significant difference from real sessions on key temporal metrics like Emotional Habituation ($p = 0.102$) and SUDS Progression ($p = 0.073$). Similarly, Narrative Development ($p = 0.251$) appears less robust but also shows no significant statistical difference, possibly due to limited context retention. This suggests the model is beginning to capture the therapeutic arc of distress reduction. Conversely, foundational constructs such as Trauma Narrative Coherence, Emotional Engagement, and Avoidance Handling show statistically significant differences ($p < 0.001$) when compared to real dialogues. These findings provide a clear and actionable roadmap for future work, guiding efforts to calibrate the model's consistency with the variability of real-world therapeutic interactions. Despite these discrepancies, the overall performance of synthetic data suggests it is a viable alternative for privacy-sensitive applications. Addressing dynamic narrative evolution could further improve alignment with real data. However, fidelity violations in synthetic sessions are often subtle and may not disrupt structural metrics. For example, therapist utterances can appear empathetic or affirming while inadvertently shifting the session away from trauma anchoring. These moments often go unnoticed by automated scoring or non-clinical reviewers, underscoring the need for fidelity-aware evaluation frameworks that integrate human clinical judgment.

6 Discussion

This paper evaluated the fidelity of synthetic PE therapy conversations by comparing them to real interactions using linguistic, structural, and statistical analyses. We discuss the findings from four key perspectives: system-level metrics and correlation (Table 2), statistical significance testing (Table 3), feature importance (Table 4), and PE-specific metrics (Table 5).

6.1 Clinical Implications

Our findings have several important implications for clinical practice. First, synthetic PE sessions can serve as valuable training tools for novice therapists learning to identify protocol deviations, as they offer controlled examples of both adherent and non-adherent interactions without privacy concerns. Crucially, this fidelity-focused evaluation serves as an essential ethical prerequisite. Before any AI tool for therapy can be considered for patient-facing studies, a framework like ours must be used to verify its adherence to clinical protocols and prevent the risk of clinically significant failures. Second, the consistent replication of structural elements (e.g., turn-taking, session flow) indicates that synthetic data can effectively supplement clinical training materials, particularly in settings with limited access to specialized trauma training.

However, our qualitative analysis reveals limitations that automated metrics may overlook. As shown in Figure 1, even linguistically fluent synthetic dialogues can contain subtle but clinically significant fidelity violations, such as role drift, premature processing, and improper SUDS implementation, that typically escape quantitative detection. These lapses would likely go unnoticed by non-specialist reviewers, highlighting the necessity of expert-guided evaluation for AI tools in clinical contexts. Finally, our PE-specific metrics provide a systematic framework for clinicians and developers to assess AI-generated content for trauma treatment, potentially setting minimum standards for therapeutic applications of synthetic data.

6.2 Qualitative Analysis of Fidelity Lapses

Section 6.1 highlighted the clinical implications of subtle fidelity violations that often escape automated detection. To provide a more granular understanding of these critical lapses and illustrate the necessity of expert review, Figure 1 details five representative examples identified through our qualitative analysis. These specific instances demonstrate common patterns of deviation from PE protocol:

- **Fidelity Lapse (Ex. 1): Role Drift** - The therapist reflects on the client's observation instead of redirecting them to the trauma narrative, which breaks the PE protocol mid-exposure.
- **Fidelity Issue (Ex. 2): Generic Affirmation** - The therapist uses an overused phrase that lacks the specific grounding necessary for an anxious, compulsive client.
- **Protocol Adherence (Ex. 3): No Issue** - The client's dialogue shows they are appropriately immersed in the trauma memory, which is the goal for imaginal exposure.
- **Fidelity Lapse (Ex. 4): Reflection During Exposure** - The client prematurely shifts into

cognitive analysis during the exposure, a process that belongs in post-session processing.

- **Protocol Adherence (Ex. 5): Trauma Anchoring** - The therapist correctly redirects the client back to the trauma scene after a drift, which successfully maintains the integrity of the PE protocol.

These detailed examples underscore why evaluation frameworks must integrate clinical expertise alongside computational metrics to accurately assess therapeutic fidelity in synthetic PE sessions.

7 Future Work

Future NLP research should focus on developing generative models with improved capabilities for tracking long-range dependencies and emotional dynamics (e.g., for Emotional Habituation, SUDS Progression), and creating automated metrics or classifiers trained with clinical expert annotations to better assess therapeutic process alignment.

8 Conclusions

This study evaluated synthetic therapeutic dialogues, showing that while they replicate structural features like turn-taking, they fall short in utterance length and conversational variability. Statistical analyses confirm these gaps, revealing that surface-level fluency can mask clinically meaningful fidelity lapses. We contribute a fidelity-aware evaluation framework tailored to Prolonged Exposure (PE) therapy, along with metrics highlighting that current generative models, despite their fluency, lack the nuance needed for high-stakes therapeutic contexts. The metrics and methods in this framework are model-agnostic and can be readily applied by the research community to benchmark and improve dialogues generated from any large language model. Our findings underscore the risk of overestimating quality through non-expert or automated evaluations. For NLP, this work emphasizes the need for clinically grounded benchmarks, richer linguistic modeling, and methods like expert-in-the-loop validation and time-aware metrics to improve fidelity in domain-specific generation.

9 Ethical Concerns

All annotators involved in fidelity evaluation had prior clinical training or supervision, and data access was limited to IRB-approved investigators to ensure ethical compliance. The real-world dataset used in this study was collected under IRB-approved protocols with participant consent and cannot be shared publicly. For details on the ethical safeguards, simulation design, and usage guidelines related to the synthetic dataset, we refer readers to BN et al. (2025b).

Acknowledgement

This work was supported by the National Science Foundation (NSF) under Grant No. 2326144. Any opinions, findings, and conclusions or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the NSF.

10 Data Availability

The synthetic dataset (Thousand Voices of Trauma) used in this paper is publicly available at `https://huggingface.co/datasets/yenopoya/thousand-voices-trauma` under the CC BY-NC 4.0 license. For collaborating on the real-world dataset, please contact PI Dr. Sherrill.

Limitations

Although our study demonstrates the promise of synthetic dialogues in approximating real PE therapy interactions, several limitations remain: (1) our analysis is presented as a deep case study focused on data generated by a single, state-of-the-art LLM; (2) future work should apply our framework to compare outputs from a wider range of models to assess the generalizability of these specific findings; (3) although we measure fidelity using structural, statistical, and protocol-based metrics, our evaluation does not assess therapeutic effectiveness or downstream clinical outcomes; (4) synthetic data that aligns structurally with real dialogues may still fall short in supporting meaningful therapeutic engagement or behavior change, especially in high-stakes or emotionally complex scenarios; (5) we do not evaluate inter-rater agreement on fidelity violations, which is critical for establishing the robustness of manual annotations, and this is planned for future work.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Anthropic. 2024. Claude 3.5 sonnet. *Anthropic Blog*.

Suhas BN and Saeed Abdullah. 2022. Privacy sensitive speech analysis using federated learning to assess depression. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6272–6276. IEEE.

Suhas BN, Yash Mahajan, Dominik Mattioli, Andrew M Sherrill, Rosa I Arriaga, Chris W Wiese, and Saeed Abdullah. 2025a. The pursuit of empathy: Evaluating small language models for ptsd dialogue support. *arXiv preprint arXiv:2505.15065*.

Suhas BN, Sarah Rajtmajer, and Saeed Abdullah. 2023. Differential privacy enabled dementia classification: An exploration of the privacy-accuracy trade-off in speech signal data. In *Proceedings of Interspeech 2023*, pages 346–350, Dublin, Ireland. ISCA.

Suhas BN, Andrew M Sherrill, Rosa I Arriaga, Chris W Wiese, and Saeed Abdullah. 2025b. Thousand voices of trauma: A large-scale synthetic dataset for modeling prolonged exposure therapy conversations. *arXiv preprint arXiv:2504.13955*.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of LLM therapists. *arXiv preprint arXiv:2401.00820*.

Hayley Evans, Udaya Lakshmi, Hue Watson, Azra Ismail, Andrew M Sherrill, Neha Kumar, and Rosa I Arriaga. 2020. Understanding the care ecologies of veterans with ptsd. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Edna B. Foa, Elizabeth A. Hembree, and Barbara O. Rothbaum. 2007. *Prolonged Exposure Therapy for PTSD: Emotional Processing of Traumatic Experiences Therapist Guide*. Oxford University Press, New York.

Mauro Giuffrè and Dennis L Shung. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*, 6(1):186.

Elizabeth A. Hembree, Edna B. Foa, Nicole M. Dorfan, Gordon P. Street, Joy Kowalski, and Xin Tu. 2003. Treatment adherence and therapist competence in exposure therapy for ptsd. In V.M. Follette and J.I. Ruzek, editors, *Cognitive-behavioral therapies for trauma*, pages 207–233. Guilford Press, New York.

Jinpeng Hu, Tengteng Dong, Hui Ma, Peng Zou, Xiao Sun, and Meng Wang. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation.

Jun-Woo Kim, Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, and Du-Seong Chang. 2024. Enhancing psychotherapy counseling: A data augmentation pipeline leveraging large language models for counseling conversations. *ArXiv*, abs/2406.08718.

Theodora Kokosi and Katie Harron. 2022. Synthetic data in medical research. *BMJ medicine*, 1(1).

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyoung-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024a. Cactus: Towards psychological counseling conversations using cognitive behavioral theory.

Suyeon Lee, Sunghwan Kim, et al. 2024b. Counselingeval: Towards evaluating the quality of llm-based psychological counseling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

R. Morris, Kareem Kouddous, Rohan Kshirsagar, and S. Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research*, 20.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Mark B. Powers, Jeffrey M. Halpern, Michael P. Ferenschak, Seth J. Gillihan, and Edna B. Foa. 2010. A meta-analytic review of prolonged exposure for posttraumatic stress disorder. *Clinical Psychology Review*, 30(6):635–641.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *ArXiv*, abs/2408.15787.

Katie A Ragsdale, Laura E Watkins, Andrew M Sherrill, Liza Zwiebach, and Barbara O Rothbaum. 2020. Advances in ptsd treatment delivery: Evidence base and future directions for intensive outpatient programs. *Current Treatment Options in Psychiatry*, 7:291–300.

Sheila AM Rauch, Afsoon Eftekhari, and Josef I Ruzek. 2009. Effectiveness of prolonged exposure for ptsd in a veterans affairs clinical setting. *Journal of Traumatic Stress*, 22(6):718–722.

Sheila AM Rauch, Carly W Yasinski, Loren M Post, Tanja Jovanovic, Seth Norrholm, Andrew M Sherrill, Vasiliki Michopoulos, Jessica L Maples-Keller, Kathryn Black, Liza Zwiebach, et al. 2021. An intensive outpatient program with prolonged exposure for veterans with posttraumatic stress disorder: Retention, predictors, and patterns of change. *Psychological services*, 18(4):606.

Chenyu Ren, Yazhou Zhang, Daihai He, and Jing Qin. 2024. Wundtgpt: Shaping large language models to be an empathetic, proactive psychologist. *ArXiv*, abs/2406.15474.

Amazon Web Services. 2023. Aws HealthScribe: A new generative AI-powered service that automatically creates clinical documentation. *AWS Press Release*.

Hao Shen, Zihan Li, Minqiang Yang, Minghui Ni, Yongfeng Tao, Zhengyang Yu, Weihao Zheng, Chen Xu, and Bin Hu. 2024. Are large language models possible to conduct cognitive behavioral therapy? *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3695–3700.

A. M. Sherrill and S. A. M. Rauch. 2019. Treatment and prevention of anxiety and related disorders: Trauma- and stressor-related disorders. In B. O. Olatunji, editor, *Handbook of Anxiety and Related Disorders*, pages 826–862. Cambridge University Press, New York.

Ruth L Varkovitzky, Andrew M Sherrill, and Greg M Reger. 2018. Effectiveness of the unified protocol for transdiagnostic treatment of emotional disorders among veterans with posttraumatic stress disorder: A pilot study. *Behavior modification*, 42(2):210–230.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2024. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. *ArXiv*, abs/2412.13660.

Carly Yasinski, Andrew M. Sherrill, Jessica L. Maples-Keller, Sheila A. M. Rauch, and Barbara O. Rothbaum. 2017. Intensive outpatient prolonged exposure for ptsd in post-9/11 veterans and service-members: Program structure and preliminary outcomes of the emory healthcare veterans program. *Trauma Psychology News*, 12(3):14–17.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, Xiping Hu, and Derek F. Wong. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *Association for Computational Linguistics*, pages 13947–13966.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Metric Definitions

Key PE terms (Foa et al., 2007; Powers et al., 2010; Rauch et al., 2009; Hembree et al., 2003):

1. **SUDS Progression:** Change in Subjective Units of Distress (SUDS) reported by the client across the session.
2. **Emotional Habituation:** Decrease in distress or emotional intensity from start to end of imaginal exposure.
3. **Trauma Narrative Coherence:** Syntactic and semantic coherence of the trauma narrative, via discourse metrics.
4. **Emotional Engagement:** Degree of emotional expression, associated with better outcomes.
5. **Avoidance Handling:** Effectiveness in addressing avoidance behaviors.
6. **Exposure Guidance:** Therapist's structuring of exposure exercises.
7. **Cognitive Restructuring:** Client's efforts to challenge maladaptive beliefs.
8. **Avoidance Reduction:** Increased engagement with trauma content.
9. **Emotion Intensity:** Variability and magnitude of emotional responses.

B Additional Tables

Table 4: Relative importance of various conversational features in distinguishing real data from synthetic data, based on a predictive model. The Importance Score (%) reflects the contribution of each feature to the model's decision-making process, with higher values indicating greater predictive power.

Feature	Imp. Score (%)
Average Utterance Length	18.42
Utterance Length Std Dev	15.76
Normalized Therapist Words	12.58
Normalized Client Words	10.94
Flow Entropy	8.72
Readability Score	7.89
Normalized Avg Turn Duration	6.43
Normalized Turn Duration Std	5.98
Vocabulary Richness	4.85
Average Perplexity	3.72
Therapist-Client Turn Ratio	2.91
Therapist-Client Word Ratio	2.32
Normalized Speaker Switches	1.88
Normalized Total Turns	0.98

Table 5: Mann-Whitney U tests on PE therapy metrics comparing real and synthetic datasets.

Metric	Statistic	p-value
Trauma Narrative Coherence	2.31×10^3	$p < 0.001$
Emotional Engagement	2.45×10^3	$p < 0.001$
Avoidance Handling	1.98×10^3	2.74×10^{-7}
Exposure Guidance	2.21×10^3	5.62×10^{-10}
Cognitive Restructuring	2.12×10^3	1.28×10^{-8}
Emotional Habituation	1.35×10^3	0.102
SUDS Progression	1.50×10^3	0.073
Avoidance Reduction	2.48×10^3	$p < 0.001$
Emotion Intensity	2.39×10^3	$p < 0.001$
Narrative Development	1.21×10^3	0.251

SUDS = Subjective Units of Distress Scale. $p < 0.001$ indicates $p < 10^{-10}$.