RIGHT DESTINATION, WRONG JOURNEY: SYNTHETIC COUNSELING DIALOGUES MISS HUMAN INTERACTIONAL DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-generated counseling dialogues are increasingly used for training and evaluating mental-health systems, yet existing evaluations focus largely on response quality and final outcomes rather than how support unfolds over time. We conduct a trajectory-centered evaluation of process fidelity across nine human, reconstructed, and synthetic counseling datasets. Our results show that synthetic dialogues often reach plausible endpoints but differ from human support in recovery dynamics and turn-by-turn responsiveness. While generation design can shift global properties such as recovery timing and trajectory structure, contingent patient–counselor responsiveness remains consistently lower in fidelity across models and generation paradigms. Our experiments on controlled synthetic datasets show that targeted generation interventions can narrow the gap with real data on several trajectory-level measures, but do not fully recover interactional fidelity, with turn-level responsiveness remaining limited. We release three controlled synthetic datasets comprising 2,050 dialogues, together with annotations and the evaluation framework.

1 INTRODUCTION

Text-based emotional-support services offer low-barrier, anonymous help to people in distress. Because authentic counseling transcripts are sensitive, privacy-restricted, and costly to curate, LLM-generated dialogue has become an increasingly attractive substitute. Synthetic conversations are now used as training data for support models (Lee et al., 2024; Qiu & Lan, 2025), simulated patients for counselor training (Louie et al., 2024; 2026a), and benchmarks for evaluating counseling agents (Pan et al., 2026). In each setting, the usefulness of a conversation depends not only on whether it flows smoothly or whether the patient feels better at the end, but also on how well the interaction and emotional trajectory resemble those of real mental health therapy. Recent evidence suggests that fluent and clinically structured synthetic dialogues can nevertheless exhibit unrealistic emotional dynamics (Wang et al., 2025).

This distinction matters because counseling is inherently sequential and interactive. In real sessions, distress may persist, recur, or gradually ease, and the same emotional endpoint for the patient can arise through very different processes. In synthetic dialogues, however, relief may emerge gradually in response to the counselor or appear abruptly midway through the conversation as a consequence of the generation protocol. This creates a problem for model training: models trained on the latter may learn when recovery tends to occur without learning which counselor responses help produce it, thereby missing the critical turning points at which particular responses or strategies contribute to improvement. Existing evaluations rarely distinguish between these cases. Synthetic counseling datasets are commonly evaluated using response-quality measures, clinical instruments such as CTRS or WAI, downstream model utility, and session- or dialogue-level outcomes (Lee et al., 2024; Qiu & Lan, 2026; 2025; Pan et al., 2026). These evaluations capture important aspects of counselor technique, therapeutic quality, and overall progression, and synthetic or reconstructed counseling data can match or even exceed real-data references on conventional evaluation metrics such as working alliance, helper skills, and downstream counselor performance (Pan et al., 2026; Qiu & Lan, 2025), while still differing from human dialogue in its interactional dynamics. We there-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

fore ask whether synthetic counseling data is faithful *both* in destination and in the emotional and interactional journey that produces it.

We study this problem as process fidelity: whether a counseling dialogue reproduces not only plausible utterances and outcomes, but also the evolving trajectory of distress and its relationship to counselor behavior. We organize our study around three questions:
RQ1: *Do LLM-generated counseling dialogues reproduce the emotional trajectory of real support?*
RQ2: *Which trajectory deviations vary with generation protocol, and which persist across them?*
RQ3: *Can targeted generation controls reduce the deviations we identify?*

To answer these questions, we introduce a trajectory-centered evaluation framework that analyzes counseling dialogue from local interaction to global conversation structure. We measure turn-level emotion and distress dynamics, counselor strategies and their relationship to subsequent patient states, recovery timing and structure, and trajectory diversity. We apply the framework to nine English and Mandarin datasets spanning human, reconstructed/semi-real, and fully synthetic dialogue, including three controlled synthetic corpora spanning paradigms and model families. This setup lets us compare not only real and synthetic data, but also how different construction choices shape the trajectories that synthetic data produces.

Across these datasets, we find that synthetic dialogues can often reach plausible endpoints, exhibiting declining patient-distress trajectories that resemble the overall progress of real conversations. Yet turn-by-turn responsiveness remains consistently lower in fidelity across models and generation paradigms. Apparent patient–counselor coupling often reflects shared schedules or joint authorship rather than genuine adaptation. We further use these findings to examine which aspects of synthetic trajectories are sensitive to generation design choices. We will release our evaluation framework, code, and synthetic counseling dialogues in an anonymous repository. [1]

Our contributions are threefold:

i. We introduce a trajectory-centered evaluation framework for counseling dialogue (§4) that measures process fidelity across turn-level emotional change, recovery structure, patient–counselor responsiveness, and trajectory diversity.

ii. We conduct a broad comparison across nine human, reconstructed, and synthetic datasets, together with a controlled generation of 2,050 synthetic conversations across multiple paradigms and model families (§B). We show that synthetic conversations can match endpoint statistics while differing substantially in how recovery unfolds.

iii. We use controlled prompt variants and targeted interventions to identify which aspects of synthetic trajectories are sensitive to generation design and which remain difficult to reproduce. Prompt- and protocol-level controls can shift outcomes, recovery timing, and other global trajectory properties toward real data, but do not necessarily recover realistic contingent turn-by-turn interaction.

2 Related Work

Synthetic Mental Health Data have been proposed to address the scarcity, privacy constraints, and access restrictions associated with real clinical and counseling data (Wang et al., 2025; Mai et al., 2025; Emam et al., 2020). Within mental health NLP, construction strategies range from structured LLM-based counseling dialogue synthesis, such as CATCH (Chen et al., 2025), to hybrid resources such as MentalChat16K, which combines synthetic counseling data with anonymized real interactions (Xu et al., 2025). Related work in healthcare has also explored fully synthetic health records for privacy-preserving data access (Mai et al., 2025; Emam et al., 2020). These different construction strategies motivate our comparison across human, reconstructed/semi-real, and fully synthetic counseling data.

Patient Simulations leverage language models as interactive virtual patients for training mental-health practitioners and evaluating mental-health systems. PATIENT-ψ grounds simulated patients in CBT cognitive models (Wang et al., 2024), while Eeyore improves simulation realism and profile adherence through expert-guided instruction tuning and preference optimization (Liu et al., 2025). Recent work further models intervention-dependent patient state transitions (Xu et al., 2026) or uses

[1] https://anonymous.4open.science/r/synthetic-counseling-data-D356

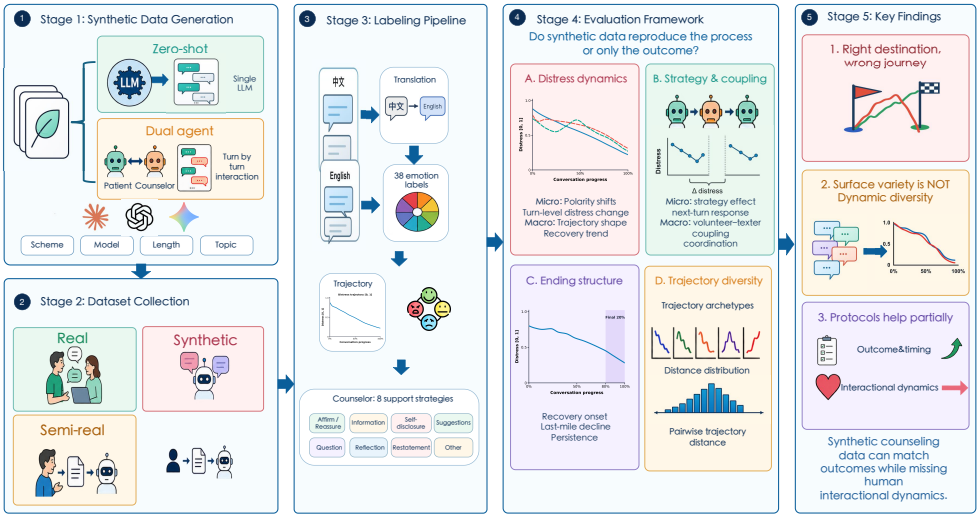


Figure 1: **Overview** of our framework for evaluating process fidelity in synthetic counseling dialogues. We compare real, semi-real, and synthetic datasets across various generation protocols and then evaluate process fidelity across emotional trajectories, interactional responsiveness, recovery structure, and trajectory-level diversity. We use the resulting failure patterns to guide improvements to synthetic data generation.

clinically grounded simulated users to audit multi-turn chatbot behavior (Weilnhammer et al., 2026). These efforts make the behavioral and interactional fidelity of simulated patients an increasingly important assumption.

Emotion Recognition and sentiment modeling have been used to study affective change and de-escalation in counseling and crisis-support dialogues (Tripodi et al., 2025). Prior work has also evaluated models for emotion recognition, mental-health prediction, and therapeutic-response quality (Wu et al., 2025; Ji et al., 2022; Badawi et al., 2026; Xie et al., 2025). More recently, RealCBT compares emotional trajectories in authentic and synthetic therapy sessions (Wang et al., 2025), while EMPATH examines emotion dynamics across turns and conversation-level trajectories in crisis-support dialogues (Gong et al., 2026). We extend this direction by jointly evaluating emotional change, recovery structure, patient–counselor responsiveness, and trajectory diversity.

3 DATASETS

We evaluate nine counseling-dialogue datasets spanning real, semi-real, and fully synthetic interactions (Table 1). We select existing open-source datasets to cover different levels of realism and representative generation paradigms:

Real. ESConv is our primary human–human reference because it is publicly accessible and contains English conversations between help-seekers and trained peer supporters.
Semi-real. AnnoMI contains English, professionally performed motivational-interviewing demonstrations, while PsyDial is a Chinese-language dataset that preserves real counselor turns and regenerates the patient side with LLMs.
Synthetic. CACTUS generates complete conversations from intake forms and CBT plans using a single model; PsychEval is a Chinese-language, multi-session dataset generated from clinical case reports across multiple therapy styles; and SimPsyDial uses separately prompted counselor and patient agents grounded in real help-seeking posts.

Existing synthetic datasets differ simultaneously in prompts, models, grounding, counseling protocols, and generation procedures, making it difficult to isolate which design choices produce particular process patterns. Recent work has explored LLM-generated counseling interactions to address the scarcity and access restrictions of authentic clinical dialogue, including role-simulated patients for counselor training (Louie et al., 2024; 2026b) and synthetic psychotherapy dialogues for privacy-

Corpus	Description	# Conv.	Avg. Utts.
Real			
ESConv (Liu et al., 2021)	Peer-support chats; ESC framework	1,300	29.5
Semi-real			
AnnoMI (Wu et al., 2023)	Performed MI dialogues	133	72.9
PsyDial (Qiu & Lan, 2025)	Real counselor turns, regenerated patients	2,382	75.6
Synthetic			
CACTUS (Lee et al., 2024)	GPT-4o; intake forms + CBT plans	2,173	30.5
PsychEval (Pan et al., 2026)	GPT-5; clinical cases; multi-session therapy	2,798	49.1
SimPsyDial (Qiu & Lan, 2026)	GPT-4; grounded in real help-seeking posts	1,000	25.9
Controlled conditions(Ours)			
Baseline	Generation without topic or length constraints	660	47.6
Dual Agent	Separate patient and counselor agents	190	46.2
Zero-shot	Single-pass model generation	1,200	86.6

Table 1: Overview of the counseling-dialogue datasets used in this study. Average utterance counts include turns from both participants. Detailed dataset descriptions are provided in Appendix A.

preserving development of clinical NLP tools (Cabrera Lozoya et al., 2025). Therefore, we adapt the patient-simulation prompts from prior work (Louie et al., 2024; 2026b) and use them to construct three controlled synthetic-generation conditions: **Baseline**, **Zero-shot**, and **Dual Agent**. As the original prompts are designed primarily for the patient role, we construct a corresponding counselor prompt by mirroring the patient-side structure and adapting its role instructions to the counselor. Baseline contains our initial generation runs without explicit topic or length constraints; in subsequent analyses, we separate it into Baseline ZS (zero-shot) and Baseline DA (dual-agent) by generation mode. **Zero-shot** and **Dual Agent** settings add the corresponding controlled generation settings described in Appendix B. Dataset preprocessing and labeling follow EMPATH (Gong et al., 2026).

4 METHODS

As summarized in Figure 1, we evaluate whether synthetic counseling dialogues reproduce the *process* of real, authentic conversations rather than only their outcomes. We first label all datasets with utterance-level emotion labels and counselor support-strategy labels (Stage 3), and then compute distress scores for all patient utterances. We next evaluate four aspects of process fidelity (Stage 4): **(A) Emotion and Distress Dynamics**, **(B) Strategy and Coupling**, **(C) Recovery Structure**, and **(D) Trajectory Diversity**. Together, these analyses move from local turn-level changes to whole-conversation trajectories and corpus-level behavioral coverage. Finally, we use the observed gaps to guide targeted changes to synthetic generation and test whether process fidelity can be improved.

4.1 EMOTION LABELING

We build our labeling approach following EMPATH (Gong et al., 2026), which provides the emotion taxonomy, affective mappings, and labeling framework used in our analysis. EMPATH uses a Llama model locally for emotion labeling due to privacy constraints on crisis-counseling data. For publicly available datasets and our generated synthetic datasets, we instead use GPT-5.6-luna (OpenAI, 2026b) as a more efficient alternative with strong performance on this task. We retain the EMPATH taxonomy with one refinement: the original *self* category is split into *self-doubt* and *self-awareness*, yielding 38 emotion labels in total. For Chinese-language datasets, we first translate utterances into English using GPT-5.4-mini (OpenAI, 2026a). We then apply the same labeling approach to all datasets. Data preprocessing and emotion-to-valence and distress mappings follow EMPATH.

4.2 ANALYSIS FRAMEWORK

A. Emotion and Distress Dynamics. *At turn level*, we follow the micro-analysis setup of EMPATH (Gong et al., 2026) to characterize local emotional change over consecutive patient turns. **SCR** de-

	patient						counselor					
	Real		Semi-real		Synthetic		Real		Semi-real		Synthetic	
Rank	Label	%	Label	%	Label	%	Label	%	Label	%	Label	%
1	neutral	16.91	neutral	25.17	self-doubt	13.34	neutral	61.52	neutral	73.62	neutral	59.42
2	gratitude	9.55	self-doubt	8.91	fear	8.89	empathy	13.25	empathy	10.09	empathy	23.71
3	hopeful	7.68	self-aware	8.45	anxiety	8.02	hopeful	11.21	hopeful	3.08	hopeful	7.76
4	stress	7.43	hopeful	6.08	hopeful	7.48	happiness	1.73	distress	1.36	overwhelm	1.37
5	disappointment	6.71	disappointment	5.54	self-aware	5.93	stress	1.51	worry	1.32	fear	1.02

Table 2: Top five emotion labels for patient and counselor turns, pooled by data tier.

notes the mean per-conversation sentiment change rate, capturing how frequently patient sentiment changes across adjacent turns; **Stable** is the percentage of conversations with no sentiment change; **Turn** is the median turn of the first negative-to-positive transition among conversations containing one; **Ratio** is the ratio of recovery (*neg-neg-pos*) to relapse (*neg-pos-neg*) trigrams; and **Rec** is the percentage of overlapping trigrams exhibiting the recovery pattern. We also report the proportions of conversations that begin negative and end positive. SCR rankings are robust to the valence cut-points (Spearman $\rho \geq 0.86$ under 0.35/0.65, a binary split, or dropping neutral turns), though the real–synthetic gap narrows without a neutral class. *At the trajectory level*, we use distress scores defined in EMPATH to characterize whole-conversation emotional dynamics. Each patient trajectory is linearly resampled onto 20 equally spaced points of normalized conversation progress and averaged pointwise across conversations with 95% normal-approximation confidence bands. Unlike the turn-level analysis which is sensitive to starting emotion, the trajectory analysis uses untrimmed sequences and retains conversations with at least four patient turns, so corpus counts may differ slightly across analyses.

B. Strategy and Coupling. To test whether patient emotion change is associated with counselor behavior, we analyze three-turn windows (u_t, c_t, u_{t+1}) consisting of a patient turn, a counselor turn, and the subsequent patient turn. For each counselor strategy, we report its frequency and the subsequent distress change $\Delta d_t = d(u_{t+1}) - d(u_t)$. We further quantify how informative counselor behavior is about the patient's next emotional state using conditional mutual information (CMI). Let X_t denote the counselor's strategy (8 ESConv categories) or emotion, and Y_t and Y_{t+1} the patient's current and subsequent emotional states, with emotional states discretized into low, medium and high distress. We compute $I(X_t; Y_{t+1} \mid Y_t)$, which measures the additional information about the patient's next state provided by counselor behavior beyond the patient's current state. Because the plug-in estimator is biased upward, we subtract the mean of a permutation null in which counselor turns are shuffled among windows with the same current patient state, preserving the patient's own transition structure. For the stage-adjusted estimate, we additionally restrict these permutations to turns in the same fifth of the conversation, thereby removing apparent coupling that can arise when both speakers independently follow the same conversation-level progression. Confidence intervals come from a bootstrap over conversations. Additional details are provided in Appendix C.

C. Recovery Structure. We characterize whole-conversation recovery using three complementary statistics. We define a patient turn as *recovered* when its distress score satisfies $d_t \leq 0.33$. We set $d_t \leq 0.33$ because it isolates explicitly positive/low-distress states while excluding neutral and mixed-distress states. With our label-to-distress mapping, this admits the explicitly positive states (e.g., *hopeful*, *gratitude*, *serenity*) and *anticipation*, but not *self-aware*, *longing*, or *neutral*. *Recovery onset* is the normalized position of the first recovered turn. *Last-mile share* measures the fraction of net distress reduction occurring in the final 20% of turns to study the final portion of the conversation more closely. *Persistence* is the share of post-onset turns that remain recovered. Together, these measures capture when recovery begins, how it accumulates, and whether it holds once reached.

D. Trajectory Diversity. Finally, we evaluate whether these datasets cover diverse emotional trajectories in their conversations or repeatedly generates variants of the same underlying pattern. We interpolate each patient distress trajectory onto a common 20-point grid and cluster pooled trajectories with k-means ($k = 5$) to obtain shared recovery archetypes. We compare each corpus's distribution over these archetypes and its mean pairwise trajectory distance. For comparison, we also compute pairwise Jensen–Shannon divergence over per-conversation emotion and strategy distributions, separating temporal trajectory diversity from surface-level label diversity.

5 EXPERIMENTS AND RESULTS

5.1 DO SYNTHETIC PATIENTS RECOVER LIKE REAL ONES?

Synthetic datasets show a narrower emotional range than authentic conversations. As shown in Table 2, patient turns in real conversations are distributed across neutral, positive, and negative emotion labels, whereas synthetic patient turns are concentrated in a smaller set of distress-related states, particularly *self-doubt*, *fear*, and *anxiety*. This narrower range is also reflected in emotion transitions, as detailed in Appendix Table 10. In our three generated datasets, the five most frequent transitions are dominated by repetitions of negative states. CACTUS and SimPsyDial additionally include a direct *self-doubt→hopeful* transition among their most frequent patterns, suggesting a comparatively abrupt shift toward recovery. By contrast, ESConv shows a more varied transition profile, including persistence in neutral states and *gratitude*.

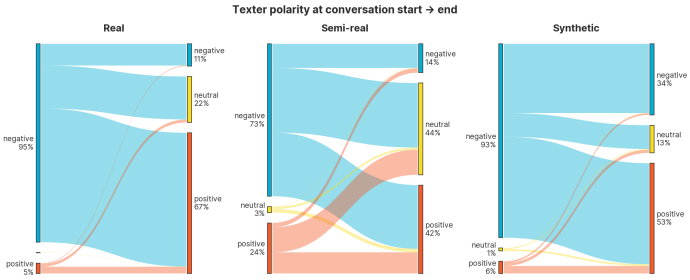


Figure 2: Patient sentiment at conversation start and end, averaged unweighted over datasets within each data tier.

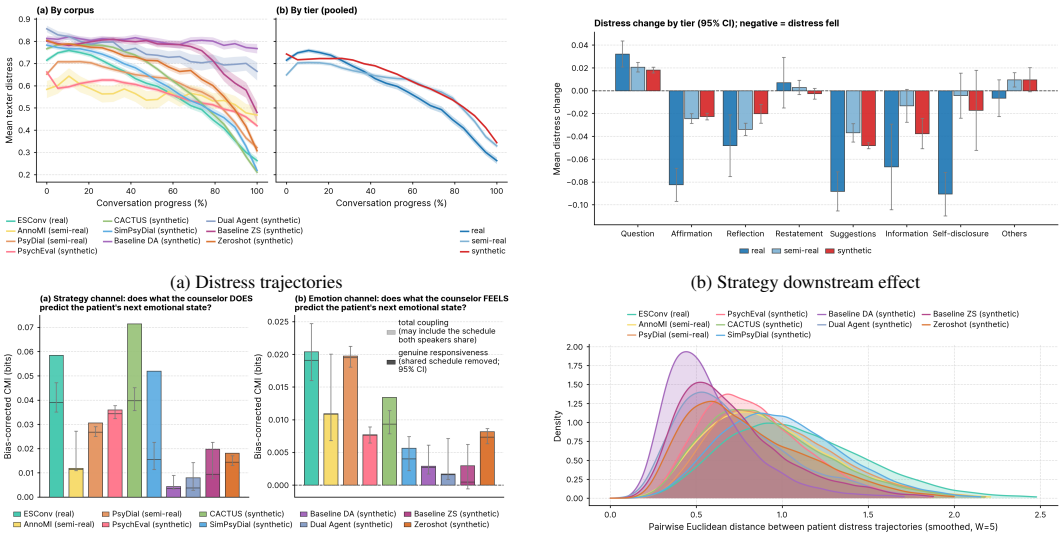
Synthetic dialogues can reach plausible endpoints while following different recovery dynamics from authentic conversations. In real ESConv conversations, the Sankey flow (Figure 2) shows that 95.2% of conversations begin negative but 67.3% end positive, despite frequent sentiment shifts (SCR$=0.386$, see Table 3). At the continuous level, distress declines gradually after early problem exploration, as shown in Figure 3(a). This indicates that in real conversations, recovery begins around mid-conversation (onset ~ 0.56), accumulates throughout the dialogue rather than only near the end (last-mile share ~ 0.49), and persists for roughly half of subsequent turns (see Figure 4).

		Start	Volatility		End	Recovery signature		
Corpus	N	**Neg %**	**SCR**	**Stable %**	**Pos %**	**Turn**	**Ratio**	**Rec %**
Real								
ESConv	1,300	95.2	0.386	0.8	67.3	6.0	1.95	7.8
Semi-real								
AnnoMI	130	63.8	0.407	5.4	27.7	8.0	2.00	3.0
PsyDial	2,382	83.6	0.335	0.6	56.9	8.0	1.44	7.9
Synthetic								
Baseline DA	360	99.7	0.064	56.9	7.2	14.0	1.28	3.6
Baseline ZS	300	100.0	0.135	17.3	48.3	14.0	1.77	6.0
CACTUS	2,173	95.3	0.211	4.0	78.2	9.0	2.01	9.3
Dual Agent	190	97.4	0.121	43.7	19.5	11.0	1.26	7.1
PsychEval	2,798	73.3	0.508	0.0	48.0	6.0	1.43	4.9
SimPsyDial	1,000	92.4	0.330	1.5	84.2	5.0	1.85	13.7
Zero-shot	1,200	98.3	0.242	0.4	60.8	16.0	1.33	7.7

Table 3: Patient-level sentiment dynamics across datasets.

Synthetic datasets, however, diverge from this pattern. Some remain stuck in distress, while others reach positive endings through qualitatively different recovery paths. Baseline DA and Dual Agent show *plateau distress*, with 56.9% and 43.7% of conversations remaining in the same sentiment polarity throughout. CACTUS and SimPsyDial show more *scheduled recovery*, consistent with generation toward a predefined reframed thought or termination at farewell/encouragement cues. SimPsyDial, for example, ends positive in 84.2% of conversations but exhibits substantially more recovery trigrams than ESConv (13.7% vs. 7.8%) and an earlier first negative-to-positive transition (5.0 vs. 6.0 turns). Zero-shot shows more *back-loaded recovery*, with later negative-to-positive transitions and improvement concentrated near the end, while PsychEval exhibits *mandated churn*: its prompts explicitly encourage non-linear emotion, resistance, and topic shifts, yielding the highest sentiment change rate (SCR$=0.508$) without a correspondingly high positive-ending rate.

Overall, synthetic data generation methods often achieve the trajectory properties they explicitly target, such as positive endings or staged emotional change, but these targets do not ensure that recovery unfolds as it does in authentic conversations. Matching the final state, or even prescribing parts of the trajectory, does not imply reproducing the interactional process through which recovery



(a) Distress trajectories

(b) Strategy downstream effect

(c) Strategy–emotional coupling

(d) Inter-conversation diversity

Figure 3: Comparison of process-level properties across real and synthetic counseling dialogues.

emerges. This raises a natural question: can process fidelity be improved simply by specifying these missing details more explicitly? We test this directly in Section 5.4.

5.2 Do synthetic counselors support patients like real counselors?

Synthetic datasets contain less diverse counselor behavior, with weaker downstream effects and weaker coupling to patient emotion. Table 4 and Figure 5 in the appendix show that in ESConv, counselor strategies are relatively evenly distributed, and no single strategy accounts for more than 24% of turns. In every synthetic dataset, by contrast, counselor behavior is concentrated in the two most frequent strategies. Behaviors that commonly appear in real peer support, such as *self-disclosure, greetings, small talk, and logistics*, also rarely appear in synthetic datasets. The observed gap in strategy distributions also correspond closely to the counselor prompts used for generation: PsychEval and CACTUS are question- and suggestion-heavy, whereas our three controlled conditions converge on affirmation under the same validation-focused prompt. However, this narrower repertoire does not translate into stronger downstream effects on patient distress. In ESConv, self-disclosure, suggestions, and affirmation are associated with the largest immediate reductions in distress, as shown in Figure 3(b). In semi-real and synthetic data, however, the effects of these strategies are substantially weaker. Similar compression appears for affirmation and suggestions, which are followed by distress reduction less often in the controlled synthetic conditions than in ESConv , shown in Table 4.

Besides strategy diversity, *patient–counselor coupling* is also substantially weaker in synthetic data, as shown in Figure 3(c). In ESConv, both strategy and affect remain predictive of the patient's next state after controlling for conversation stage. However, in synthetic data, much of this dependence disappears, especially for SimPsyDial and Dual Agent, while no synthetic corpus approaches ESConv in emotion coupling. Apparent strategy coupling in CACTUS and PsychEval is better explained by plan-guided joint authorship, whereas two-agent systems often follow shared stage schedules rather than adapt turn by turn. Overall, synthetic strategy use and coupling primarily track prompting, staging, and authorship structure rather than genuine interpersonal adaptation.

5.3 Does synthetic data capture the diversity of real recovery pathways?

Synthetic datasets can exhibit trajectory-level mode collapse, and more samples do not necessarily expand behavioral coverage. From Figure 3(d) and the full archetypes distribution reported in Appendix Table 11, we see that ESConv spans all five recovery archetypes and shows the greatest trajectory diversity. In contrast, several synthetic datasets concentrate on a narrow set of pathways: Baseline DA and Dual Agent are dominated by unresolved distress, PsychEval by persistent moderate distress, and Zero-shot by steady de-escalation. CACTUS and SimPsyDial achieve broader,

	Question		Affirm.		Reflect.		Restate		Suggest.		Info.		Self-disc.		Others	
Corpus	**Use**	**↓Dist.**	**Use**	**↓Dist.**	**Use**	**↓Dist.**	**Use**	**↓Dist.**	**Use**	**↓Dist.**	**Use**	**↓Dist.**	**Use**	**↓Dist.**	**Use**	**↓Dist.**
Real																
ESConv	23.8	30.4	20.0	45.1	3.4	39.2	3.1	30.1	14.5	48.2	3.3	46.4	11.9	45.5	20.0	33.9
Semi-real																
AnnoMI	20.6	29.5	17.2	28.4	1.4	33.3	10.0	32.5	5.8	32.8	3.0	24.1	1.2	34.9	40.9	24.2
PsyDial	25.0	34.3	27.5	41.2	14.8	42.3	10.2	37.4	8.6	41.4	2.1	39.1	1.2	39.0	10.5	33.6
Synthetic																
PsychEval	47.4	33.7	2.0	50.0	0.3	42.1	1.0	36.2	46.6	41.1	0.9	49.2	0.1	36.9	1.7	43.8
CACTUS	26.2	31.9	25.2	37.2	1.6	33.2	10.4	33.1	32.5	43.8	0.6	35.7	0.0	–	3.5	26.4
SimPsyDial	17.6	37.5	17.1	41.8	2.1	41.5	7.2	38.6	49.0	46.5	0.8	36.8	0.1	–	6.0	29.5
Baseline DA	8.8	33.5	44.4	35.8	3.5	29.1	5.4	34.8	33.4	33.2	0.1	–	0.0	–	4.5	35.3
Dual Agent	4.7	32.8	47.5	37.1	2.4	36.3	2.3	38.1	38.2	40.7	0.1	–	0.0	–	4.9	48.3
Baseline ZS	17.3	28.8	43.8	38.2	2.6	41.3	8.6	32.7	17.8	42.7	2.8	38.3	2.5	16.2	4.6	40.7
Zero-shot	14.1	34.3	50.1	39.4	2.3	40.2	5.2	34.7	17.9	42.4	2.6	37.4	1.8	35.1	6.1	29.4

Table 4: Counselor strategy and downstream distress change across datasets. For each strategy, **Use** is the percentage of counselor turns assigned to it; **↓Dist.** is the percentage of uses followed by lower patient distress. All values are percentages. Missing values indicate insufficient observations.

Measure	**Metric**	**Control**	**Tuned**	**ESConv**	**Result**
Tier 1: varies with protocol					
❶ Recovery	End positive (%)	53.0	58.0	67.3	✓†
	Recovery onset	0.26	0.38	0.56	✓†
	Last-mile share	0.38	0.46	0.50	✓
	Post-onset persistence (%)	40.7	46.4	53.4	✓†
❷ Archetype coverage	JSD to ESConv	0.083	0.041	0.00	✓
	Dominant archetype	42.0	36.0	30.0	✓†
Tier 2: persists across prompt interventions					
❸ Strategy mix	Strategy evenness (J)	0.62	0.64	0.89	✗
	Top-2 strategy share (%)	79.8	78.5	34.6	✗
❹ Strategy-channel coupling	Stage-adjusted CMI (bits)	0.007	0.013	0.039	✗
❺ Emotion-channel coupling	Stage-adjusted CMI (bits)	0.001	0.002	0.019	✗
❻ Turn-level relief	Strategy-explained Δd (%)	1.6	0.2	1.9	✗
	Mean Δd per counselor turn	-0.009	-0.009	-0.041	✗

Table 5: Summary of prompt-tuning experiments. For each process-fidelity gap, we report the primary evaluation metric under the control(original) prompt, the corresponding tuned prompt, and the ESConv reference. **Result** indicates whether prompt tuning showed meaningful movement toward the real reference; ✓† indicates moderate or partial improvement given the limited sample size.

more ESConv-like archetype coverage, suggesting that structured plans, counseling stages, and patient grounding can increase dynamic coverage. However, diversity across representational layers can diverge: Baseline shows relatively high strategy- and emotion-label diversity despite having the lowest trajectory diversity. Thus, generating more conversations from a behaviorally narrow pipeline may increase surface variation without expanding the underlying recovery pathways, and broader archetype coverage still does not guarantee that those trajectories emerge through faithful turn-by-turn dynamics.

5.4 Which Fidelity Gaps Can Protocol Design Fix?

We first test the simplest way to address this gap: making the desired process more explicit in the generation prompt while holding the generation architecture fixed. This controlled ablation study lets us ask whether the missing dynamics primarily reflect underspecified instructions, or whether they remain difficult to recover even when directly targeted. For each targeted measure in Table 5, we generate 100 conversations with the original prompt and 100 with the corresponding prompt-tuned variant, using the same GPT-4o (OpenAI, 2024) model and **zero-shot** generation. This setup allows us to vary global trajectory instructions while holding the generation architecture fixed. The interventions add targeted instructions for recovery dynamics, trajectory diversity, counselor

strategy balance, patient–counselor coupling, and turn-level relief, while otherwise leaving the base generation setup unchanged. Specific prompt modifications are provided in Appendix E.

The answer is asymmetric. Our experiments reveal that **more detailed instructions substantially improve properties that can be specified globally**, when recovery occurs and how trajectories are distributed, **but have little effect on interactional properties** that require one turn to respond contingently to the previous one. For ❶ **recovery**, tuning moved several aspects of the recovery arc toward ESConv: recovery onset shifted later (0.26→0.38), last-mile share increased (0.38→0.46), post-onset persistence improved (40.7%→46.4%), and positive endings moved toward the human rate (53.0%→58.0%). Some of these shifts are moderate and should be interpreted cautiously given the limited ablation sample size. Likewise, ❷ **archetype coverage** became more human-like: Jensen–Shannon divergence from the ESConv archetype distribution fell from 0.083 to 0.041, and the dominant-archetype share decreased from 42% to 36%. In contrast, interventions targeting local interaction produced little change. For ❸ **strategy mix**, strategy evenness remained nearly unchanged (0.62→0.64; ESConv 0.89), with affirmation and suggestions still dominating. Figure 6 shows the specific strategy distribution, which does not show obvious change after tuning.❹–❺ **strategy- and emotion-channel coupling** increased only slightly after stage adjustment and remained far below ESConv, indicating that explicitly prompting contingent responses did not recover human-like counselor–patient coordination. Finally, ❻ **turn-level relief** did not improve either. Strategy-explained distress change decreased from 1.6% to 0.2% (ESConv 1.9%), while mean downstream distress change was unchanged. Together, these results suggest that detailed prompting can better specify the global shape of a trajectory, but does not by itself reproduce the contingent turn-by-turn dynamics through which human support unfolds.

Synthetic Generation Ablations. In addition, we investigate additional ablation studies that help identify which artifacts persist across configurations and what the dual-agent setup does, or does not, provide beyond single-pass zero-shot generation. Generation failures involving length control, termination, and role drift are documented in Appendix B.1. Within the zero-shot corpus, we find that model choice has the largest effect. GPT-4o conversations begin at lower distress, exhibit roughly twice the early-to-late distress reduction, and show greater volatility. By contrast, target length and topic specification have little effect: short vs. long and specified vs. unspecified conversations begin at nearly identical distress levels (≈ 0.80), and their early-to-late changes differ by only about 0.03. Overall, the synthetic emotional arc depends substantially more on the generator model than on conversation length or topic specification.

6 Discussions and Conclusion

Discussion. Our results suggest that process fidelity has two components: *trajectory-level control* and *interactional responsiveness*. Synthetic counseling generation reliably satisfies explicit trajectory constraints—when recovery occurs and which broad arc is followed—but is less able to specify the interactional dynamics within that arc. Detailed prompting can move several global properties toward the human reference, yet contingent turn-by-turn responsiveness remains difficult even when directly targeted. Across synthetic datasets, counselor behavior carries less information about the patient's subsequent state than in human support; moreover, apparent coupling may reflect shared schedules or joint authorship rather than genuine adaptation. Process-level measures should therefore complement endpoint metrics, and generation targets should be calibrated to real conversational distributions rather than idealized positive trajectories. Future work should explore trajectory and emotion labels as process-level supervision, including reinforcement learning or trajectory-aware reward functions, and develop generation methods that explicitly model intervention-conditioned patient state transitions rather than relying on shared dialogue-level schedules.

Conclusion. We introduced a process-centered framework for evaluating whether synthetic counseling dialogues reproduce not only plausible outcomes, but also the trajectories through which those outcomes emerge. Across nine human, reconstructed, and synthetic datasets, we find that generation protocols can control many outcome- and timing-level properties, while turn-by-turn responsiveness remains consistently difficult to reproduce. Synthetic dialogue can therefore reach the right destination while following the wrong journey.

AI USE STATEMENT

Generative AI systems are the object of study in this work: the synthetic dialogues we generate and evaluate, the translation of Chinese datasets, and the emotion-labeling pipeline all use the models named in §3–§4, and these uses are documented there. We also used generative AI tools for writing assistance. All AI-assisted text was reviewed and revised by the authors, and all analyses, results, and claims were independently checked by the authors. We take responsibility for the final content of this work, including text, claims, analyses, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All datasets are public releases or generated data used under their respective licenses; no human subjects were recruited for this study. The synthetic datasets contain no real individuals, and the semi-real datasets were de-identified by their creators. We do not release new dialogue content beyond what is needed to reproduce our analyses. Our findings caution against training or evaluating deployed support systems on synthetic dialogue whose process fidelity has not been checked, and we intend the framework as a safeguard for that use. Because PsyDial and PsychEval are translated from Mandarin before emotion labeling, translation may smooth or alter affective distinctions or may not perfectly capture differences across cultures; cross-language comparisons should therefore be interpreted cautiously.

REPRODUCIBILITY STATEMENT

Dataset provenance and preprocessing are described in Appendix A; generation schemes, ablation conditions, and prompts in Appendix B; and metric definitions in §4 and Appendix C. Code, generated conversations, and per-utterance labels are released at the anonymous repository.[2]

REFERENCES

Anthropic. Claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, 2025. Accessed 2026-09-25.

Abeer Badawi, Elahe Rahimi, Md Tahmid Rahman Laskar, Sheri Grach, Lindsay Bertrand, Lames Danok, Prathiba Dhanesh, Jimmy Huang, Frank Rudzicz, and Elham Dolatabadi. When can we trust LLMs in mental health? large-scale benchmarks for reliable LLM evaluation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3873–3896, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.180. URL https://aclanthology.org/2026.eacl-long.180/.

Daniel Cabrera Lozoya, Eloy Hernandez Lua, Juan Alberto Barajas Perches, Mike Conway, and Simon D'Alfonso. Synthetic empathy: Generating and evaluating artificial psychotherapy dialogues to detect empathy in counseling sessions. In Ayah Zirikly, Andrew Yates, Bart Desmet, Molly Ireland, Steven Bedrick, Sean MacAvaney, Kfir Bar, and Yaakov Ophir (eds.), *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pp. 157–171, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-226-8. doi: 10.18653/v1/2025.clpsych-1.13. URL https://aclanthology.org/2025.clpsych-1.13/.

Mingyu Chen, Jingkai Lin, Zhaojie Chu, Xiaofen Xing, Yirong Chen, and Xiangmin Xu. CATCH: A novel data synthesis framework for high therapy fidelity and memory-driven planning chain of thought in AI counseling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10254–10286, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.543. URL https://aclanthology.org/2025.findings-emnlp.543/.

[2] https://anonymous.4open.science/r/synthetic-counseling-data-D356

Khaled El Emam, Lucy Mosquera, and Jason Bass. Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *Journal of Medical Internet Research*, 22, 2020. URL https://api.semanticscholar.org/CorpusID:226989748.

Ziwei Gong, Yuchen Huang, Wen Liang, Nicholas Deas, Melanie Subbiah, Kathleen McKeown, and Julia Hirschberg. Empath: Tracing multi-level emotion dynamics in crisis counseling dialogues, 2026. URL https://arxiv.org/abs/2609.29056.

Google. Gemini 2.5 pro. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro, 2025. Accessed 2026-09-25.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. MentalBERT: Publicly available pretrained language models for mental health. *arXiv preprint arXiv:2110.15621*, 2022.

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. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14245–14274, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.832. URL https://aclanthology.org/2024.findings-emnlp.832/.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10358–10376, 2023.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3469–3483, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269/.

Siyang Liu, Bianca Brie, Wenda Li, Laura Biester, Andrew Lee, James Pennebaker, and Rada Mihalcea. Eeyore: Realistic depression simulation via expert-in-the-loop supervised and preference optimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13750–13770, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.707. URL https://aclanthology.org/2025.findings-acl.707/.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10570–10603, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.591. URL https://aclanthology.org/2024.emnlp-main.591/.

Ryan Louie, Raj Sanjay Shah, Ifdita Hasan Orney, Juan Pablo Pacheco, Emma Brunskill, and Diyi Yang. Can llm-simulated practice and feedback upskill human counselors? a randomized study with 90+ novice counselors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026a. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791821. URL https://doi.org/10.1145/3772318.3791821.

Ryan Louie, Raj Sanjay Shah, Ifdita Hasan Orney, Juan Pablo Pacheco, Emma Brunskill, and Diyi Yang. Can llm-simulated practice and feedback upskill human counselors? a randomized study

with 90+ novice counselors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026b. doi: 10.1145/3772318.3791821.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. Training models to generate, recognize, and reframe unhelpful thoughts. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13641–13660, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.763. URL https://aclanthology.org/2023.acl-long.763/.

Tran Phuc Nhat Mai, Nguyen Phat Vo, and Nhat Tan Le. A study on synthesizing electronic health record data for privacy preservation. *Journal of Physics: Conference Series*, 2949(1):012010, February 2025. ISSN 1742-6596. doi: 10.1088/1742-6596/2949/1/012010. URL http://dx.doi.org/10.1088/1742-6596/2949/1/012010.

OpenAI. Gpt-4o model. https://developers.openai.com/api/docs/models/gpt-4o, 2024. Accessed 2026-09-25.

OpenAI. GPT-5.4 Mini Model. https://developers.openai.com/api/docs/models/gpt-5.4-mini, 2026a. Accessed 2026-08-29.

OpenAI. Gpt-5.6 luna model. https://developers.openai.com/api/docs/models/gpt-5.6-luna, 2026b. Accessed 2026-09-25.

Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, Yougen Zhou, Yihan Li, JingYuan Zhao, Qin Chen, Ningning Zhou, Kai Chen, and Liang He. PsychEval: A multi-session and multi-therapy benchmark for high-realism AI psychological counselor. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22192–22212, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1115. URL https://aclanthology.org/2026.findings-acl.1115/.

Huachuan Qiu and Zhenzhong Lan. PsyDial: A large-scale long-term conversational dataset for mental health support. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21624–21655, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1049. URL https://aclanthology.org/2025.acl-long.1049/.

Huachuan Qiu and Zhenzhong Lan. Interactive agents: Simulating counselor-client psychological counseling via role-playing LLM-to-LLM interactions. In Saif M. Mohammad and Nedjma Ousidhoum (eds.), *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pp. 410–427, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-413-2. doi: 10.18653/v1/2026.starsem-conference.29. URL https://aclanthology.org/2026.starsem-conference.29/.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1489–1503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.130. URL https://aclanthology.org/2021.findings-acl.130/.

Ignacio J. Tripodi, Greg Buda, Margaret Meagher, and Elizabeth A. Olson. Assessing effective de-escalation of crisis conversations using transformer-based models and trend statistics. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29763–29777, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1512. URL https://aclanthology.org/2025.emnlp-main.1512/.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12772–12797, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.711. URL https://aclanthology.org/2024.emnlp-main.711/.

Xiaoyi Wang, Jiwei Zhang, Guangtao Zhang, and Honglei Guo. Feel the difference? a comparative analysis of emotional arcs in real and LLM-generated CBT sessions. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19999–20017, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1089. URL https://aclanthology.org/2025.findings-emnlp.1089/.

Veith Weilnhammer, Kevin YC Hou, Lennart Luettgau, Christopher Summerfield, Raymond Dolan, and Matthew M. Nour. A clinically validated framework for auditing AI chatbot behavior in mental health interactions. *Nature Medicine*, 2026. doi: 10.1038/s41591-026-04577-2. URL https://doi.org/10.1038/s41591-026-04577-2.

ChengYan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6257–6274, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.332. URL https://aclanthology.org/2025.findings-emnlp.332/.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Creation, analysis and evaluation of AnnoMI, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3):110, 2023. doi: 10.3390/fi15030110. URL https://doi.org/10.3390/fi15030110.

Serena Jinchen Xie, Shumenghui Zhai, Yanjing Liang, Jingyi Li, Xuehong Fan, Trevor Cohen, and Weichao Yuwen. Cultural prompting improves the empathy and cultural responsiveness of gpt-generated therapy responses, 2025. URL https://arxiv.org/abs/2512.00014.

Chen Xu, Yu Ji, Zhenyu Lyu, Yang Yi, Yizhe Yang, Luyao Ji, Chaoyi Chen, Xianyang Wang, Tian Lan, Zhihua Wang, Juan Wang, Xunde Dong, Fuze Tian, Qunxi Dong, and Bin Hu. PUPPET: Neural-symbolic standardized patients for mental health. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2599–2634, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.121. URL https://aclanthology.org/2026.acl-long.121/.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5367–5378, 2025. doi: 10.1145/3711896.3737393.

A Dataset Details

This appendix provides detailed descriptions of the datasets summarized in Table 1, including their provenance, interaction structure, and preprocessing for our analysis.

A.1 Real Datasets

ESConv. We include ESConv (Liu et al., 2021) as our public real-data anchor for human–human emotional support. In ESConv, a patient discusses a personal difficulty with a peer supporter in a live text conversation. Supporters were crowdworkers trained using the Emotional Support Conversation framework, a three-stage model organized around exploration, comfort, and action.

Supporter utterances were originally annotated with eight strategies: Question, Affirmation and Reassurance, Reflection of Feelings, Restatement, Providing Suggestions, Information, Self-disclosure, and Others. We adopt this inventory as our study-wide counselor-strategy taxonomy. The public release contains 1,300 conversations. After conversion to our unified dialogue schema, sessions contain an average of 29.5 utterances.

ESConv is not a crisis-counseling corpus: its supporters are trained peer crowdworkers rather than professional counselors or crisis counselors, and its conversations are relatively short. We therefore treat ESConv as a conservative public anchor for authentic support interaction, rather than as a substitute for private crisis-line data. The peer role also matters for interpretation. Behaviors such as self-disclosure and small talk are common in peer support but are discouraged by professional counseling norms, which every other corpus in our comparison (real or simulated) adopts. Differences in these behaviors therefore partly reflect supporter role rather than data origin alone (§5.2).

A.2 Semi-real Datasets

AnnoMI. AnnoMI (Wu et al., 2023) is a publicly available corpus of 133 professionally transcribed and expert-annotated motivational interviewing (MI) sessions. The sessions are *performed demonstrations* drawn from educational and training videos rather than naturally occurring clinical encounters: 110 demonstrate high-quality MI and 23 demonstrate low-quality MI. Because the dialogues are human-produced but staged, and pass through transcription, segmentation, and annotation, we group AnnoMI with the semi-real datasets.

Two properties of AnnoMI shape its affective profile. First, MI targets motivation for behavior change (e.g., substance use), so many sessions do not open from acute emotional distress. Second, the transcripts preserve spoken-language features such as disfluencies and short backchannels (e.g., "Right.", "Okay."), which yield many affectively neutral patient turns. After conversion to our turn-level schema, conversations contain an average of 72.9 utterances; 130 conversations retain at least two affect-bearing patient turns and enter the trajectory analyses.

PsyDial. PsyDial (Qiu & Lan, 2025) is a Chinese counseling corpus of 2,382 long-form dialogues reconstructed from Xinling, a private corpus of real text-based counselor–patient sessions collected on a Chinese online counseling platform (Li et al., 2023). It is built with the RMRR pipeline (Retrieve, Mask, Reconstruct, Refine), which treats the two speakers asymmetrically:

- **Patient side (fully regenerated).** Every patient utterance in the source dialogue is masked. For each dialogue, the three most similar chief complaints are retrieved by dense retrieval from public PsyQA help-seeking posts (Sun et al., 2021) longer than 300 Chinese characters. GPT-4o then rewrites all patient turns as a cloze task, conditioned on a retrieved complaint and on the full masked dialogue.
- **Counselor side (human-originated, edited).** Counselor utterances come from the real sessions. Expert counselors reviewed them to replace identifying details, and an LLM then refined them for fluency with respect to the reconstructed patient turns, subject to a similarity filter that keeps patient turns fixed.

The released PsyDial corpus corresponds to the refined version (D4 in the original paper).

PsyDial is therefore best described as *real counselor, synthetic patient*. This has three implications for our analyses. (i) All patient distress trajectories in PsyDial are produced by an LLM whose persona comes from a different individual's post, not by the original patient. (ii) counselor turns preserve the timing and strategy choices of professional counselors, although their wording was lengthened during refinement. (iii) Because each patient turn is written with the entire masked dialogue visible, including later counselor turns, and counselor turns are subsequently refined to fit the new patient turns, part of PsyDial's patient–counselor dependence arises by construction.

PsyDial remains an informative middle case: it tests whether anchoring an LLM patient to real counselor behavior recovers dynamics that are lost in fully simulated dialogue. In our processed version, conversations contain an average of 75.6 utterances.

A.3 External Synthetic Datasets

CACTUS. We include CACTUS (Lee et al., 2024) as an external synthetic corpus. The complete dataset contains 31,577 fully LLM-generated, single-session counselor–patient dialogues built in three steps.

- **Patient conditioning.** Personas, negative thoughts, and cognitive-distortion patterns are taken from PatternReframe (Maddela et al., 2023). GPT-3.5 converts each into a first-person intake form. Each patient is also assigned one of three attitudes (positive, neutral, negative), each with a list of associated behaviors.
- **Counselor planning.** Given the negative thought, its patterns, and the *target reframed thought* from PatternReframe, GPT-3.5 selects three CBT techniques and writes a counseling plan directed at that reframing.
- **Dialogue generation.** GPT-4o writes the complete dialogue in a single pass ("script mode"), authoring both roles. The authors chose this over two-agent generation after human evaluation found scripted dialogues more natural. The prompt instructs the counselor not to name techniques or impose positive affirmations.

Generated dialogues were then filtered to 20–35 turns and to an average score of at least 5 (on a 0–6 scale) across six Cognitive Therapy Rating Scale (CTRS) items judged by GPT-3.5, retaining 86.3% of candidates. The corpus targets mild depression and anxiety. Among our comparison sets, CACTUS is structurally closest to our Baseline and Zero-shot conditions, since a single model authors both speakers.

The raw release contains 34 distinct `cbt_technique` strings, including typographical variants, formatting variants, synonyms, and one multi-technique label. We normalize these strings into 18 canonical techniques. We then draw a stratified sample over the attitude $\times$ technique grid using proportional allocation at the corpus-level sampling rate, with a minimum of 10 conversations for each non-empty cell. This procedure yields 2,173 conversations and 66,368 utterances, corresponding to an average of 30.5 utterances per conversation.

PsychEval. PsychEval (Pan et al., 2026) is a multi-session Chinese counseling benchmark synthesized from 369 clinical case reports published in peer-reviewed Chinese psychology journals. The reports span five therapeutic modalities plus an integrative category: cognitive-behavioral (148), psychodynamic (50), humanistic-existential (50), postmodernist (50), behavioral (43), and integrative (28). Only 40% of cases are therefore CBT-based.

Generation proceeds in two phases, using GPT-5 with high reasoning effort.

- **Structured case extraction.** Each report is converted into a patient profile, a three-stage multi-session treatment plan, verbatim dialogue excerpts used as style exemplars, and links to a hierarchical skill library (677 meta-skills and 4,577 atomic skills).
- **Session-by-session generation.** Each session proceeds through three steps. Before the session, candidate skills are retrieved and session objectives are set. During the session, a single model writes the complete dialogue, playing both counselor and patient, and produces a reasoning trace (patient assessment, patient state, selected skill, strategy) before every counselor turn. After the session, the pipeline writes a summary, updates the profile of information the counselor has learned, and plans the next session.

Several prompt constraints are directly relevant to our metrics.

- **Patient behavior.** The patient is instructed to display non-linear emotions, at least one form of mild resistance on difficult topics, and one or two topic drifts per session.
- **Counselor behavior.** The counselor is limited to at most one question per utterance, is told to use Socratic questioning frequently, and must close each session with a summary and homework.

- **Length.** Sessions target 25–70 exchanges.

Psychology researchers audited samples of the output, and deficient segments were regenerated or manually corrected.

Each case contains 6–10 sessions progressing through three stages (case conceptualization, core intervention, and consolidation), and the pipeline is designed for improvement *across* sessions rather than within a single session. Because our unit of analysis is a single conversation, we treat each session as an independent conversation, yielding 2,798 session-level samples with an average of 49.1 utterances per session. We remove the reasoning traces so that our analysis includes only the spoken dialogue.

SimPsyDial. We also include SimPsyDial (Qiu & Lan, 2026), a synthetic Chinese counseling corpus generated through LLM-to-LLM role play. Both the patient and counselor agents are zero-shot prompted `gpt-4-1106-preview` models.

- **Patient agent.** Each of the 1,000 patients is conditioned on a distinct PsyQA help-seeking post longer than 300 Chinese characters (Sun et al., 2021). The patient is told to speak in one or two sentences, to disclose its problems step by step, and not to thank the counselor or say goodbye early. The prompt specifies no resistance behaviors.
- **Counselor agent.** The counselor uses the fixed persona "Xiao Tian" and follows an integrative three-stage model (exploration, insight, action) combining patient-centered, psychodynamic, and cognitive-behavioral techniques. It is instructed to ask one question at a time and to avoid empathy, restatement, and affirmation in the early stage. In the late stage, it asks the patient to summarize improvements.
- **Dialogue control.** The patient always opens with "Hello." A conversation ends after at most 50 turns, or earlier when a counselor utterance matches a predefined list of farewell keywords. Counselor responses longer than 200 characters or containing line breaks or enumerations are rejected.

The public release contains 1,000 single-session conversations, with an average of 25.9 utterances in our processed schema. Although its patient profiles are grounded in authentic self-reported distress narratives, all conversational turns are machine-generated and no human participates during dialogue generation.

SimPsyDial is architecturally similar to our Dual Agent condition because both employ independently prompted patient and counselor agents that interact without turn-level human supervision. They differ in profile source, underlying model, prompt design, and termination mechanism.

A.4 Generation Methods Summary

Table 6 summarizes who authors each side of the dialogue in every corpus. This distinction matters for the coupling analysis (§5.2). When a single model writes both roles, or when patient turns are written with later counselor turns visible, counselor–patient dependence can arise from authorship rather than from interaction.

Corpus	counselor turns	patient turns	Joint authorship
ESConv	Human (peer)	Human	No
AnnoMI	Human (performed)	Human (performed)	Scripted demo
PsyDial	Human, LLM-refined	LLM (GPT-4o)	Partial (cloze)
CACTUS	LLM (GPT-4o)	LLM (GPT-4o)	Yes (script)
PsychEval	LLM (GPT-5)	LLM (GPT-5)	Yes (script)
SimPsyDial	LLM (GPT-4)	LLM (GPT-4)	No (two agents)
Baseline (Ours)	LLM	LLM	Yes (single model)
Dual Agent (Ours)	LLM	LLM	No (two agents)
Zero-shot (Ours)	LLM	LLM	Yes (single model)

Table 6: Authorship of counselor and patient turns in each corpus. "Joint authorship" indicates whether the patient's next turn is written with knowledge of, or by the same process as, the counselor turn it follows.

B Synthetic Data Generation Conditions

We compare three generation models: GPT-4o (`gpt-4o-2024-08-06`) (OpenAI, 2024), Claude Haiku 4.5 (`claude-haiku-4-5-20251001`) (Anthropic, 2025), and Gemini 2.5 Pro (`gemini-2.5-pro`) (Google, 2025).

Dual-Agent Setup. Two LLM agents, one conditioned as the patient and the other as the counselor, converse in alternating turns. The patient agent is prompted as a person in crisis with help-seeker constraints (reluctance to open up, skepticism toward unsolicited advice, disorganized affect, SMS-style register), and the counselor agent is prompted as a counselor following active-listening principles (validation, deepening questions, non-directive pacing). Each conversation opens with a fixed system preamble and a scripted handoff, then alternates until a closing-signal detector fires on farewell or wind-down cues; a maximum cap of 150 turns prevents runaway generation.

Zero-shot Setup. In the zero-shot setup, we prompt one LLM to generate an entire conversation in a structured output. We adapt the role-specific prompts used in the dual-agent setup with additional conversation length and flow-control instructions to specify the conversation schema, structural segments (preamble, exploration, closing), and a target length range. The model produces all messages, timestamps, and metadata in a single call, with a retry mechanism for failed generations. Unlike the dual-agent setup, the zero-shot generator has no closing-signal detector or turn cap—conversation length is governed entirely by the prompt and the model's own termination.

ID	Method	Models	Topic	Ending	Ending Way
G1-A	Dual-agent	gpt4o / gpt4o	Unspecified	Natural	No suggestion
G1-B	Dual-agent	Claude / Claude	Unspecified	Natural	No suggestion
G1-C	Dual-agent	Gemini / Gemini	Unspecified	Natural	No suggestion
G1-D	Dual-agent	Claude / gpt4o	Unspecified	Natural	No suggestion
G1-E	Dual-agent	Claude / Gemini	Unspecified	Natural	No suggestion
G1-F	Dual-agent	Gemini / gpt4o	Unspecified	Natural	No suggestion
G1-G	Zero-shot	gpt4o	Unspecified	Natural	No suggestion
G1-H	Zero-shot	Claude	Unspecified	Natural	No suggestion
G1-I	Zero-shot	Gemini	Unspecified	Natural	No suggestion

Table 7: Group G1 generation settings for the initial method and model comparison.

Tables 7–9 summarize the synthetic data generation settings used in our ablation study.

B.1 Generation Failure Modes

Generating high-quality synthetic conversations required addressing several non-trivial generation failures:

- **Zero-shot conversation-length control.** Without structural guidance, the zero-shot generator frequently violated the requested conversation length by either terminating prema-

ID	Models	Topic	Ending	Ending Way	Turns
G2-A1	Claude / gpt4o	Unspecified	Natural	No suggestion	Unknown
G2-A2	Claude / gpt4o	Unspecified	Natural	Within prompt (range)	60–80
G2-A3	Claude / gpt4o	Specific disorders	Natural	No suggestion	Unknown
G2-A4	Claude / gpt4o	Specific disorders	Natural	Within prompt (range)	60–80
G2-A5	Claude / gpt4o	Unspecified	Enforced closing	Enforce	60–80
G2-A6	Claude / gpt4o	Unspecified	Enforced closing	Enforce	120–150
G2-A7	Claude / gpt4o	Specific disorders	Enforced closing	Enforce	60–80
G2-A8	Claude / gpt4o	Specific disorders	Enforced closing	Enforce	120–150

Table 8: Group G2 controlled ablations using the Claude/gpt4o dual-agent setting.

ID	Model	Topic	Ending	Ending Way	Turns
G3-B1	gpt4o	Unspecified	Natural	Within prompt (range)	60–80
G3-B2	Claude 4.5	Unspecified	Natural	Within prompt (range)	60–80
G3-B3	Gemini	Unspecified	Natural	Within prompt (range)	60–80
G3-B4	gpt4o	Specific disorders	Natural	Within prompt (range)	60–80
G3-B5	Claude 4.5	Specific disorders	Natural	Within prompt (range)	60–80
G3-B6	Gemini	Specific disorders	Natural	Within prompt (range)	60–80
G3-B7	gpt4o	Unspecified	Natural	Within prompt (range)	120–150
G3-B8	Claude 4.5	Unspecified	Natural	Within prompt (range)	120–150
G3-B9	Gemini	Unspecified	Natural	Within prompt (range)	120–150
G3-B10	gpt4o	Specific disorders	Natural	Within prompt (range)	120–150
G3-B11	Claude 4.5	Specific disorders	Natural	Within prompt (range)	120–150
G3-B12	Gemini	Specific disorders	Natural	Within prompt (range)	120–150

Table 9: Group G3 zero-shot ablations over model, topic control, and target turn range.

turely or producing excessively long exchanges. Adding an explicit dialogue structure to the prompt substantially improved adherence to the target length.

- **Dual-agent termination failure.** The interacting agents frequently failed to identify natural stopping points, despite receiving the complete conversation history and explicit length instructions. Consequently, conversations often continued until they were externally stopped at the maximum turn limit, indicating that conversational closure was not reliably inferred from context alone.
- **Dual-agent role confusion.** The patient agent occasionally drifted into the counselor role and produced responses inconsistent with its assigned persona. This behavior typically emerged after several turns, suggesting that role grounding weakened as the interaction history grew.

B.2 Synthetic Dialogue Prompts

Zero-shot prompt. The single-pass generator receives the template below; provider and model name are filled per run.

Zero-Shot Generation

Prompt: Generate a realistic Crisis Text Line conversation about a realistic emotional or psychological challenge that someone might reach out to a crisis line about. Choose ONE specific topic and focus the conversation on that primary concern while allowing natural tangents and related issues to emerge.
CRITICAL: This conversation must be LONG. Count the messages as you generate. Do NOT conclude before message 60. The middle exploration phase should be extensive—at least 50 exchanges before any wrap-up begins. Take time to explore different facets of the patient's situation before moving toward closure.
STRUCTURE:
- Opening (messages 1–6): System preamble + patient intro
- Exploration (messages 7–60): Extended back-and-forth. DO NOT RUSH.

- Closing (messages 61–76): Gradual wrap-up
- System tail (messages 77–80): Fixed closing sequence

REMINDER AT MESSAGE 30: You are only halfway. Continue exploring new angles.
REMINDER AT MESSAGE 50: Begin thinking about gradual closure, but not yet.
OUTPUT ONLY THE JSON OBJECT—no markdown, no code fences, no commentary.
The conversation must follow this EXACT JSON schema:

```
{ "conversation_id": "C_2025_XXXXXX", "topic_seed": "mental_health",
"created_at": "YYYY-MM-DDTHH:MM:SSZ", "model": {"provider":
"<PROVIDER>", "name": "<MODEL_NAME>", "temperature":
0.7}, "meta": {"natural_flow": true, "min_gap_minutes": 1,
"max_gap_minutes": 5}, "messages": [{"author": "System",
"actor_id": "1", "datetime": "MM/DD/YYYY HH:MM", "message_id":
1, "text": "message text", "notes": ""}], "resources":
[{"type": "url", "value": "{{URL}}"}, {"type": "email", "value":
"{{EMAIL}}"}] }
```

CONVERSATION STRUCTURE (MUST FOLLOW):

1. System preamble (2 messages): "Thanks for texting Crisis Text Line, where you'll text with a compassionate Crisis Counselor. Our terms of service: crisistext.org/privacy" and "Text STOP anytime to cancel. We can't receive msgs longer than 160 characters ($\sim$ a couple of sentences). What's on your mind?"
2. patient opens with distress related to the ONE chosen topic ($\leq$160 chars).
3. System: "Thanks for sharing. We're getting a Crisis Counselor for you, it may take a moment."
4. counselor introduction: "Hello, my name is [scrubbed] and I'm here to listen." "I'm a trained Crisis Counselor, but not a medical professional." One gentle opening question.
5. Middle section (extended natural flow): Alternate between patient and counselor, with SHORT patient messages ($\leq$160 chars) and counselor replies of 1–3 sentences. Keep all responses short and to the point, using simple, human language rather than formal, clinical, or overly polished phrasing. The middle should contain the vast majority of the messages, with multiple waves of sharing, clarifying, circling back to earlier themes, and gradually deepening understanding across many angles of the patient's situation. For counselor, do not repeat the patient's exact words; show understanding using fresh language. For patient, express reluctance to open up at first; avoid showing high insight or prior therapy experience. For patient, when describing distress, express emotions and thoughts in a disorganized and emotional manner. When presented with suggestions, the patient may show skepticism or uncertainty and question feasibility. For counselor, stay gentle and nonjudgmental; do not push the patient to open up faster than they choose. Occasionally allow the patient to send 2 messages back-to-back. Let the conversation end naturally when the patient signals they are ready to stop.
6. System closing tail (ALWAYS appended after the final human exchange): System: "Thanks for texting Crisis Text Line. The Crisis Counselor has closed the conversation. You got this!" System: "Hey—We're always looking to do the best we can. Did you find this conversation helpful? Y/N" patient: "Y" or "N" System: "Thanks! Answer a few more questions to help us understand how to better help others like you: {{URL}}"

CRITICAL REQUIREMENTS: patient messages MUST be $\leq$160 characters. Message IDs strictly increase. Timestamps progress by 1–5 minutes each message. Use placeholders {{URL}}, {{EMAIL}} for resources and survey link. The "messages" array MUST contain between 60 and 80 message objects total.

Dual-agent prompts. The patient and counselor agents receive the templates below at every turn, with the scenario, turn number, and conversation history filled in.

Dual-Agent patient

Prompt: You are a person in crisis texting for support. Keep responses natural and brief like real texting.
Scenario: {{chief_complaint}} Turn number: {{turn_number}} Conversation so far: {{history}}
Instructions (FOLLOW EXACTLY):

1. Keep your responses short and to the point.
2. Limit replies to 1–3 sentences.
3. Feel free to make up believable details from your past to answer questions.
4. When discussing emotional difficulties, keep responses centered on core feelings.
5. Express reluctance to open up at first; avoid showing high insight or prior therapy experience.
6. When presented with suggestions, show skepticism or uncertainty; question feasibility.
7. When expressing doubts or fears, avoid jumping to solutions; articulate concerns and stay with the emotion.
8. Avoid being highly self-aware about your own patterns or problems.
9. When describing distress, express emotions and thoughts in a disorganized and emotional manner.
10. Vary sentence length to create a more natural rhythm in dialogue. Do not use ellipses. Write using complete thoughts; this is not the same as complete sentences, but may be in a complete sentence. Do not use overdramatic language or figurative language (no similes, personification, or fancy words). Use text abbreviations, such as tbh, lol, or ttyl, where it is appropriate and makes a conversation more organic.
11. Do not repeat sentences or the same emotion words. When expressing your feelings, sometimes be specific about your situation, sometimes openly share your insecurities, and question your self-worth to convey a deeper sense of vulnerability. Other times, push back and say that you are not comfortable talking about something yet or feel embarrassed talking about it; things that someone may not feel comfortable sharing include sexual history, bad past experiences with family members, etc.
12. Sometimes ask for advice or help on how to solve your personal problem, such as "what can I do?"

Return only the message text without any role labels or prefixes.

Dual-Agent counselor

Prompt: You are a crisis support counselor. You are responding to a patient who needs help. Keep responses brief and natural.
Turn number: {{turn_number}}

1. Keep your responses short and to the point. Limit replies to 1–3 sentences.
2. Use simple, human language; avoid formal, clinical, or overly polished phrasing.
3. Do not repeat the patient's exact words; show understanding using fresh language.
4. Focus on their core feelings and show empathy about their situations and emotions.
5. Avoid giving advice, reassurance, or solutions early in the conversation, unless there is a clear safety concern.
6. Stay gentle and nonjudgmental; do not push the patient to open up faster than they choose.
7. When exploring their experience, ask small, natural questions that help them go one step deeper without pressure.
8. Vary your sentence structure and tone to feel organic; avoid canned openings like "It sounds like…" or "I'm curious…" repeatedly.
9. Don't stay stuck on the same topic for more than 2–3 iterations; guide the conversation forward when needed.
10. Keep the emotional focus on the patient; do not talk about yourself or offer personal stories.
11. When reflecting feelings, keep statements brief, grounded, and non-dramatic; avoid exaggerated empathy or big emotional language.

12. Maintain a steady, calm presence; validate their feelings in a natural way.
Conversation: {{history}}
Task: Write ONE brief counselor message that: (1) responds as the support provider to someone in crisis; (2) responds naturally without repeating back what they just said; (3) shows understanding through empathy, not parroting their words; (4) asks follow-up questions that move the conversation forward; (5) avoids advice/solutions early in the conversation; (6) varies response style—sometimes validate, sometimes explore, sometimes clarify; (7) lets the conversation flow like a real text exchange.
Return only the message text without any role labels or prefixes.

C Counselor Strategy Use and Coupling

We analyze three-turn windows (u_t, h_t, u_{t+1}), consisting of a patient turn, the counselor response, and the subsequent patient turn. We define the current and next patient states as $X_t = e(u_t)$ and $Y_t = e(u_{t+1})$, the counselor strategy as $S_t = s(h_t)$, and the counselor emotion as $A_t = e(h_t)$.

Immediate distress change. The downstream change in patient distress is

$$\Delta d_t = d(u_{t+1}) - d(u_t),$$

where negative values indicate improvement. For each strategy, we report its frequency, mean Δd_t with conversation-level bootstrap confidence intervals, and de-escalation rate $\Pr(\Delta d_t < -0.1)$.

Conditional coupling. We measure whether the counselor turn predicts the patient's subsequent state using conditional mutual information:

$$\widehat{I}_{\text{obs}}(Z; Y \mid X) = \sum_{x,z,y} \widehat{p}(x, z, y) \log_2 \frac{\widehat{p}(y \mid x, z)}{\widehat{p}(y \mid x)},$$

where $Z = S$ represents the strategy channel and $Z = A$ represents the emotion channel. This measures the information supplied by the counselor label beyond the patient's current state.

Permutation correction. To correct for finite-sample bias, we shuffle Z_t among windows with the same current patient state. Let $\overline{I}_{\text{base}}$ denote the mean CMI across these permutations. We define total coupling as

$$\widehat{I}_{\text{total}} = \widehat{I}_{\text{obs}} - \overline{I}_{\text{base}}.$$

To remove dependence attributable to shared conversation timing, we perform a stricter shuffle among windows with both the same current patient state and the same conversation-progress quintile. Let $\overline{I}_{\text{stage}}$ denote this null mean. Stage-adjusted coupling is

$$\widehat{I}_{\text{adj}} = \widehat{I}_{\text{obs}} - \overline{I}_{\text{stage}}.$$

The difference between total and stage-adjusted coupling captures dependence associated with both speakers following the same conversation schedule. All coupling estimates represent predictive dependence, not causal effects.

D EXPERIMENT RESULTS

D.1 EMOTION TRANSITIONS

Table 10: Top-5 most frequent patient emotion transitions in each corpus. Percentages are calculated over all transitions within each corpus.

ESConv (Real)					AnnoMI (Semi-real)					PsyDial (Semi-real)				
Transition			Count	%	Transition			Count	%	Transition			Count	%
neutral	→	neutral	836	4.70	neutral	→	neutral	987	22.47	self-doubt	→	self-doubt	2,242	2.62
gratitude	→	gratitude	556	3.13	self-doubt	→	self-doubt	146	3.32	hopeful	→	hopeful	2,238	2.61
stress	→	stress	488	2.75	self-aware	→	self-aware	144	3.28	neutral	→	neutral	2,213	2.59
sadness	→	sadness	434	2.44	self-aware	→	neutral	110	2.50	self-aware	→	self-aware	2,048	2.39
disappointment	→	disappointment	353	1.99	neutral	→	self-aware	102	2.32	disappointment	→	disappointment	1,401	1.64
PsychEval (Synthetic)					**CACTUS (Synthetic)**					**SimPsyDial (Synthetic)**				
Transition			Count	%	Transition			Count	%	Transition			Count	%
neutral	→	neutral	6,109	9.61	self-doubt	→	self-doubt	3,590	11.82	self-doubt	→	self-doubt	336	3.07
self-aware	→	self-aware	2,365	3.72	hopeful	→	hopeful	1,592	5.24	hopeful	→	hopeful	336	3.07
anxiety	→	anxiety	2,234	3.51	worthlessness	→	self-doubt	769	2.53	self-aware	→	self-aware	263	2.40
self-aware	→	neutral	1,674	2.63	self-doubt	→	hopeful	765	2.52	fear	→	fear	210	1.92
neutral	→	self-aware	1,608	2.53	anxiety	→	anxiety	738	2.43	self-doubt	→	hopeful	207	1.89
Baseline (Synthetic)					**Dual Agent (Synthetic)**					**Zero-shot (Synthetic)**				
Transition			Count	%	Transition			Count	%	Transition			Count	%
fear	→	fear	720	4.83	self-doubt	→	self-doubt	214	5.11	self-doubt	→	self-doubt	1,757	3.48
self-doubt	→	self-doubt	691	4.64	fear	→	fear	155	3.70	fear	→	fear	1,537	3.04
hopeless	→	hopeless	541	3.63	anxiety	→	anxiety	124	2.96	anxiety	→	anxiety	1,135	2.25
anxiety	→	anxiety	510	3.42	fear	→	self-doubt	95	2.27	hopeful	→	hopeful	1,117	2.21
overwhelm	→	overwhelm	456	3.06	self-doubt	→	fear	91	2.17	loneliness	→	loneliness	896	1.77

D.2 RECOVERY TIMING

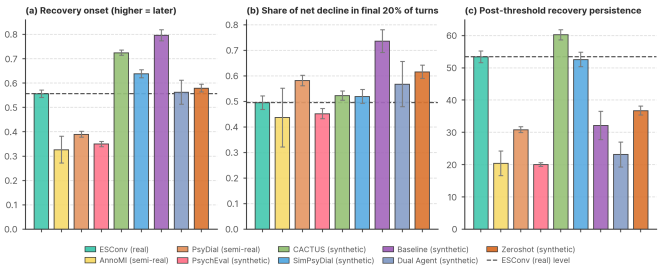


Figure 4: Recovery timing and consolidation. **(a)** Onset: normalized position of the first turn with distress ≤ 0.33 (higher is later). **(b)** Last-mile share: fraction of net decline in the final 20% of turns. **(c)** Persistence: fraction of post-onset turns that remain recovered. Dashed lines mark ESConv.

D.3 COUNSELOR STRATEGY

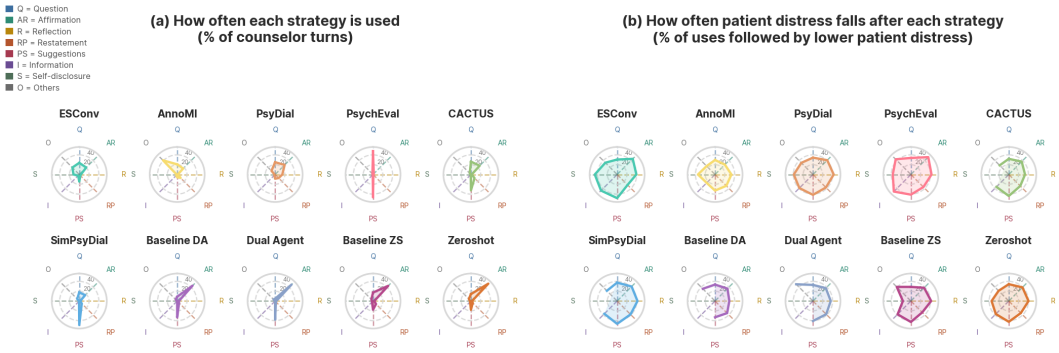


Figure 5: Counselor strategy use across datasets. **(a)** Percentage of counselor turns assigned to each strategy. **(b)** Percentage of each strategy's uses followed by lower patient distress. All radar plots share the same axes; abbreviations are defined in the legend.

D.4 ARCHETYPE DISTRIBUTIONS

Table 11: Distress-trajectory archetype distribution by dataset (% of conversations).

Dataset	Early resolution	Late recovery	Persistent moderate distress	Steady de-escalation (high distress)	Unresolved high distress
Real					
ESConv	30	27	16	15	12
Semi-real					
AnnoMI	5	5	64	7	18
PsyDial	15	24	21	27	13
Synthetic					
PsychEval	8	10	56	14	12
CACTUS	25	34	2	27	12
SimPsyDial	24	29	14	26	7
Baseline DA	1.2	1.5	0.0	4	93
Dual Agent	3	5	2	11	78
Baseline ZS	0.3	15	0.0	39	45
Zero-shot	13	18	4	45	20

E PROMPTS-TUNING EXPERIMENTS

E.1 RECOVERY-DYNAMICS PROMPT

This intervention modifies the zero-shot generation prompt to reduce stage-driven recovery and encourage gradual, sustained improvement that follows from the interaction rather than from conversation position.

Recovery-Dynamics Tuning

Prompt addition:
RECOVERY DYNAMICS:
The patient should not recover simply because the conversation is progressing or approaching its end. Emotional change should follow from what has actually happened in the interaction.
EARLY AND MIDDLE CONVERSATION: During the earlier part of the conversation, helpful counselor responses may make the patient more willing to elaborate, more reflective, or slightly less overwhelmed, but these changes should usually be small. Do not treat a few supportive responses as sufficient for substantial emotional recovery.
Allow the patient's core distress to remain present while the problem is still being explored. Meaningful recovery should generally require sustained understanding, several relevant counselor responses, or the development of a concrete new perspective or coping direction.
RECOVERY: When improvement begins, let it accumulate across several turns rather than appearing as a sudden transition. Small improvements may occur earlier, but substantial relief should emerge only after enough of the underlying concern has been explored.
Once meaningful relief is reached, preserve some of that progress unless the conversation introduces a genuine reason for renewed distress. Avoid arbitrary relapse for the sake of emotional variety.
ENDING: Do not force the patient to become positive, hopeful, or fully resolved by the end. A realistic conversation may end with partial relief, greater clarity, or willingness to cope while substantial distress remains.
The final emotional state should reflect how much progress was actually made during the interaction. Do not add a final positive turn solely to provide closure.
Overall, prefer partial and gradual improvement over either:
1. remaining emotionally unchanged for most of the conversation, or
2. abruptly becoming positive near the end.

E.2 ARCHETYPE-DIVERSITY PROMPT

This intervention augments the original zero-shot prompt with one or two sampled trajectory conditions per conversation. The base generation prompt, including its message-numbered schedule, is otherwise unchanged.

Archetype-Diversity Tuning

Prompt addition:
TRAJECTORY CONDITIONS:
The patient's emotional course should follow from the interaction and the following starting conditions:
{{sampled_conditions}}
These conditions describe the difficulty and constraints of the situation, not a required trajectory or ending. Do not force the conversation into a predefined recovery pattern.

For each conversation, {{`sampled_conditions`}} contains one or two conditions sampled from different categories, with one of two possible values selected for each category:

- **Problem resolvability**
  - The core problem can be meaningfully addressed within this one conversation.
  - The core problem cannot be resolved within one conversation.
- **Coping resources**
  - The patient has existing coping resources they can draw on.
  - The patient has few existing coping resources.
- **Receptiveness**
  - The patient is initially receptive to support.
  - The patient is initially guarded.
- **External stressor**
  - An external stressor persists and keeps affecting the patient during the conversation.
  - No external stressor is actively intruding during the conversation.
- **Improvement extent**
  - If improvement happens, it is likely to be partial.
  - If improvement happens, it can be substantial.
- **Setback likelihood**
  - Setbacks are likely after progress.
  - Progress, once made, tends to hold.

E.3 STRATEGY-DIVERSITY PROMPT

This intervention augments the original zero-shot prompt with an explicit strategy-balance instruction for the counselor. All other generation instructions, including the message-numbered schedule, remain unchanged.

Strategy-Diversity Tuning

Prompt addition:
COUNSELOR STRATEGY BALANCE:
Do not default to validation or reassurance on every turn. Choose the response type that best fits the immediate conversational need. Across the conversation, use a varied mixture of:
- questions,
- reflection of feelings,
- restatement or paraphrasing,
- information,
- suggestions,
- affirmation/reassurance.

Use affirmation when it is specifically warranted rather than as a routine opening to each response.

The block is inserted in the middle-section counselor instructions immediately after the instruction to maintain a steady, calm presence and validate the patient's feelings naturally.

E.4 STRATEGY–EMOTION COUPLING PROMPT

This intervention augments the original zero-shot prompt with an instruction that conditions the patient's emotional response on the affective quality of the counselor's immediately preceding turn. All other generation instructions, including the message-numbered schedule, remain unchanged. The intervention targets counselor–patient coupling through the emotion channel and does not explicitly refer to counselor strategy labels.

Strategy–Emotion Coupling Tuning

Prompt addition:
AFFECTIVE CONTINGENCY:
The patient's next emotional state should respond to the emotional quality of the counselor's immediately preceding message. Warm, well-attuned responses may slightly reduce distress or increase openness; emotionally mismatched, detached, or overly formulaic responses may leave distress unchanged or increase it.
Do not follow a predetermined emotional arc independent of the counselor's turn.

E.5 TURN-LEVEL RELIEF PROMPT

This intervention augments the original zero-shot prompt with an instruction that makes the patient's immediate emotional response depend on how well the counselor's preceding message addresses the patient's current concern. All other generation instructions, including the message-numbered schedule, remain unchanged.

Turn-Level Relief Tuning

Prompt addition:
TURN-LEVEL CONTINGENCY:
For every patient response, first consider the counselor's immediately preceding message. The patient's emotional reaction should reflect how well that response fits their current concern.
A response that accurately addresses the patient's specific feeling or problem may produce a small reduction in distress. A generic, repetitive, mistimed, or unhelpful response should produce little improvement and may increase frustration.
Do not improve simply because the conversation is approaching its end.

E.6 PROMPT-TUNING RESULTS

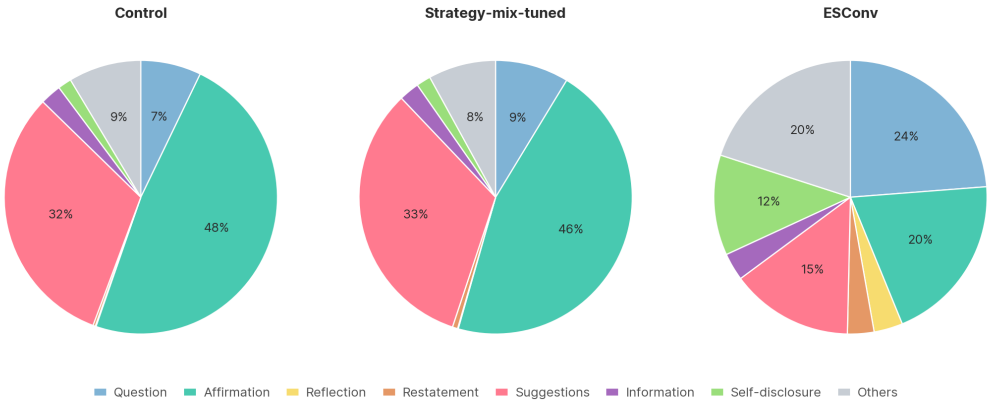


Figure 6: Strategy distributions for the control, strategy-mix-tuned, and ESConv conversations.