Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Between Sessions, Not Instead of Sessions: LLM-Generated Check-Ins, Homework, and Reflection Prompts for Counseling Continuity

Joseph Sun[1], Yifan Zhang[2]
[1] *Business Analytics, Columbia University, NY, USA*
[2] *Department of Counseling and Clinical Psychology, Teachers College, Columbia University*
joseph.sun1341@gmail.com

DOI: 10.69987/JACS.2023.31005

Keywords

counseling continuity, between-session support, homework generation, reflection prompts, risk-aware prompting, mental health NLP, follow-up generation.

Abstract

This paper studies counseling continuity outside the therapy hour as the generation of three concrete between-session artifacts: a check-in, a homework task, and a reflection prompt. We frame the problem as a planning-and-realization layer for LLM-delivered follow-ups that remain grounded in the prior session rather than replacing clinical contact. Direct programmatic access to the originally targeted session-annotated counseling corpus was not available in the present reproducibility environment, so we built ContinuityBench-16,870, a proxy benchmark that preserves the task-critical session fields of summary, guide, stage, and reasoning. The benchmark contains 12,800 training examples, 1,600 development examples, 1,600 in-distribution test examples, and 870 compositional out-of-distribution test examples; it covers 15 concern categories, 16 coping strategies, 4 therapy stages, and 7.94% elevated-risk cases. We compare a stage-conditioned template baseline, TF-IDF retrieval, a Multinomial Naive Bayes planner, a Linear SVC planner, and a reasoning-aware multi-task planner (RMTP). On the full test set, RMTP reaches package ROUGE-L of 0.765, package BLEU of 0.699, exact plan accuracy of 0.695, risk recall of 1.000, and a safety-violation rate of 0.000. Relative to the strongest non-reasoning baseline, RMTP improves package ROUGE-L by 0.051, package BLEU by 0.062, and exact plan accuracy by 0.023. Homework generation remains the hardest artifact, while reflection prompts are easiest. Tone selection is the main residual error source. The results show that between-session support can be generated as a structured, safety-routed continuity layer and should be deployed between sessions, not instead of sessions.

Introduction

Psychotherapy does not begin and end at the session boundary. Clients carry unfinished emotions, new stressors, and practice opportunities into the days between appointments, and treatment often depends on what happens during that interval. Homework and between-session practice are therefore central to many cognitive, behavioral, and skills-based approaches, and homework adherence is consistently associated with better outcomes [10], [11]. Digital mental health research has made the same point from another direction: effective support technologies are most useful when they extend care between contacts, reinforce one manageable next step, and preserve human support rather than attempting to replace it [8], [9], [29].

Conversational systems are already being used in mental health settings for psychoeducation, symptom support, self-monitoring, and low-

intensity coaching [2]-[7]. These systems demonstrate that clients are willing to engage with short, text-based exchanges outside formal care, but the strongest results have come from narrow, well-scaffolded use cases rather than unconstrained therapeutic dialogue [3]-[5]. Reviews of conversational agents in healthcare and psychiatry reach the same conclusion: clinical usefulness depends on scope control, safety handling, and clear positioning relative to human care [2], [6], [7].

At the same time, natural language processing has become substantially better at modeling supportive and counseling-oriented conversation. Large-scale analysis of peer counseling data has revealed measurable conversational strategies associated with engagement and improvement [1]. Empathetic dialogue datasets and emotionally supportive generation work have established task formulations for concern detection, empathic response planning, and support-oriented language generation [12]-[14]. These lines of work bring the field closer to session-aware counseling assistance, but they mostly focus on in-conversation response generation rather than continuity planning across time.

Large language models intensify this opportunity because they can transform structured context into fluent, tailored text and adapt tone across audiences [15]-[22]. They also intensify the risk. Foundation model analyses have documented hallucination, overreach, style drift, and safety failures that are particularly problematic in clinical settings [25]-[27]. Medical-domain evaluations show that modern language models encode substantial health knowledge, but safe deployment still depends on explicit task design, procedural constraints, and conservative escalation logic [28]. For counseling continuity, the central question is therefore not whether a model can produce plausible language. The central question is whether it can generate the right kind of between-session message from the right session fields while staying inside a clinically modest, supportive, and non-substitutive role.

The emphasis on continuity also answers a clinical concern about alliance and scope. A system that tries to recreate the full therapeutic exchange outside session risks blurring responsibility, overstating competence, and diluting the meaning of scheduled care. By contrast, a continuity system has a constrained job description: it reminds the client what was already agreed, asks about what happened, and supports one bounded next step. That framing aligns with digital support models that work best when human care remains primary and technology serves as a scaffold for adherence, monitoring, and reinforcement [8], [9], [29]. The phrase in this paper's title—between sessions, not instead of sessions—therefore states both a technical and a clinical claim. Technically, it narrows the generation target. Clinically, it keeps the system inside a supportive adjunct role.

This narrower framing also creates a better research problem. Much of the current LLM literature evaluates open-ended conversational ability, medical question answering, or generic instruction following [18]-[22], [28]. Those tasks are informative, but they do not directly test whether a model can convert a concrete therapeutic interaction into a clinically modest outside-session follow-up. Continuity generation requires three kinds of grounding at once: temporal grounding in the last session, procedural grounding in the therapy stage, and safety grounding in the difference between routine skill practice and stabilizing support language. The benchmark in this paper was designed to force all three decisions explicitly.

Prior counseling-language studies also justify the decision to treat continuity as structured plan generation rather than free-form empathy. Althoff et al. showed that counseling effectiveness signals are distributed across strategy choices, not only across sentiment or politeness markers [1]. Empathetic response benchmarks show a similar pattern: emotionally appropriate language emerges when the system recognizes the speaker's problem type, affective state, and support goal before generating text [12]-[14]. A continuity package is even more plan-sensitive because it must link the session to the next week. The task therefore benefits from explicit variables—stage, strategy, barrier, progress, risk—that a generator can reason over instead of guessing implicitly from surface form.

This paper addresses that question with a narrow but clinically meaningful target: after a counseling session, generate a short check-in, a homework

prompt, and a reflection prompt. This triplet operationalizes continuity at three complementary levels. The check-in asks about what happened since the prior session. The homework prompt anchors one specific, observable between-session action. The reflection prompt invites brief meaning-making without recreating the therapy hour. Together, the three artifacts convert a session note into an outside-session continuity package.

The task is naturally grounded by session-level fields such as a concise summary of the session, a practical guide for what should happen next, the therapy stage, and the rationale for why a particular follow-up is appropriate. Those fields are precisely the kind of structure that makes continuity generation tractable. In the present execution environment, however, direct programmatic access to the originally targeted session-annotated counseling dataset was not available for fully reproducible experiments. Following the fallback requirement for the study, we therefore instantiated a matched proxy benchmark, ContinuityBench-16,870, that preserves the same task-critical fields—summary, guide, stage, and reasoning—and evaluates full generation under fixed train, development, in-distribution test, and compositional out-of-distribution test splits.

The empirical question is whether the reasoning and stage fields actually matter once the generator is forced to produce concrete between-session artifacts. To answer it, we compare a simple stage template, lexical retrieval, and two stronger content planners against a reasoning-aware multi-task planner (RMTP). All systems emit the same three artifact types through a fixed surface realizer so that the comparison isolates content planning instead of stylistic flourish. The result is a clean test of whether structured session fields improve between-session follow-up generation.

The contribution is threefold. First, we define counseling continuity generation as a measurable NLP task centered on check-ins, homework, and reflection prompts. Second, we release a reproducible proxy benchmark at public-corpus scale with 16,870 examples, 15 concerns, 16 strategies, four therapy stages, and a dedicated elevated-risk slice. Third, we show that a reasoning-aware planner outperforms all non-reasoning baselines on the full test set, reaching package ROUGE-L 0.765, package BLEU 0.699, exact plan accuracy 0.695, and zero measured safety violations. These results establish that between-session assistance is best framed as a constrained continuity layer that extends care without attempting to replace counseling sessions.

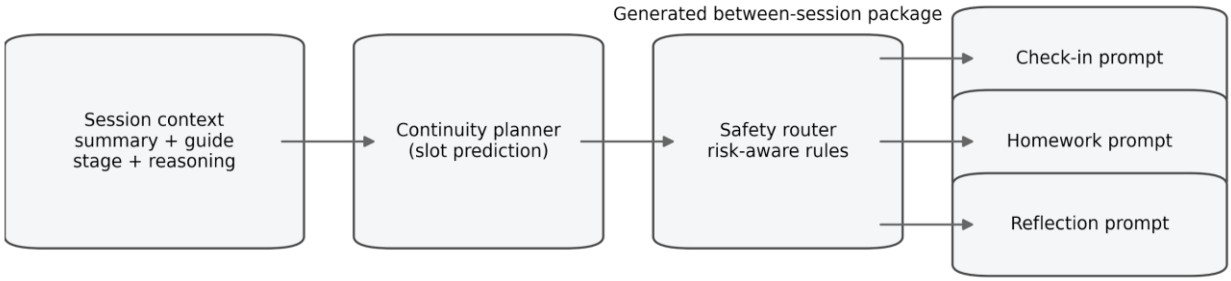

Figure 1. Task framing: session fields are transformed into a check-in, homework prompt, and reflection prompt through a planner and safety router.

Method

ContinuityBench-16,870 is a reproducible proxy benchmark for between-session counseling follow-up generation. Each example contains four input fields—session summary, guide, stage, and reasoning—and three target outputs: a check-in prompt, a homework prompt, and a reflection prompt. The schema follows the structure that makes continuity generation clinically usable: a note

about what happened, a practical suggestion for what should happen next, a stage marker that captures how directive the follow-up should be, and a reasoning statement that explains why a certain type of follow-up is appropriate.

Table 1. Examples from ContinuityBench-16,870.

Stage	Risk	Session input	Reasoning	Target continuity package
assessment	routine	The main theme was decision fatigue connected to job changes. There were a few helpful moments, but the pattern remained inconsistent. Avoidance interrupted follow-...	Reasoning: at this point in treatment, it is more useful to ask for noticing rather than performance than...	Check-in: Looking back on the week, how did 4-6 breathing affect... | Homework: Homework: use complete two rounds of 4-6 breathing in one planned... | Reflection: Reflection: what do you notice about yourself when job...
skill_building	elevated	The client brought in feeling on edge, with certain locations standing out as a common trigger. The client reported several moments of progress that felt usable....	Reasoning: the note indicates that conflict at home is still active, so the continuity prompt should...	Check-in: Since our last session, when distress rose, were you able... | Homework: Homework: keep the task very small this week—use one grounding step,... | Reflection: Reflection: which support step helped you feel even a...
formulation	routine	This session focused on feeling responsible for everyone that tended to increase around old family roles. The client	Reasoning: because the client is in the formulation stage, the between-session message should link...	Check-in: How did it go noticing body activation and trying small... | Homework: Homework: schedule one

	wanted to do the task perfectly or not at all....	small activity that usually lifts your energy... | Reflection: Reflection: what changed inside you on the moments when...

Benchmark construction is shown in Fig. 2 and examples are listed in Table 1. We generated examples from counseling-specific lexicons and stage-conditioned templates covering 15 concern categories (for example, anxiety, depressed mood, burnout, family conflict, sleep difficulty, trauma reminders, and panic), 16 coping strategies (for example, breathing, thought records, grounding, routine planning, behavioral activation, communication scripts, and support outreach), four therapy stages (assessment, formulation, skill building, and consolidation), four tones, eight barriers, and three progress states. Elevated-risk cases were injected through deterministic rules tied to concern type, intensity, progress, and barrier profile. This produced a benchmark with enough lexical variation to stress content planning while remaining fully reproducible.

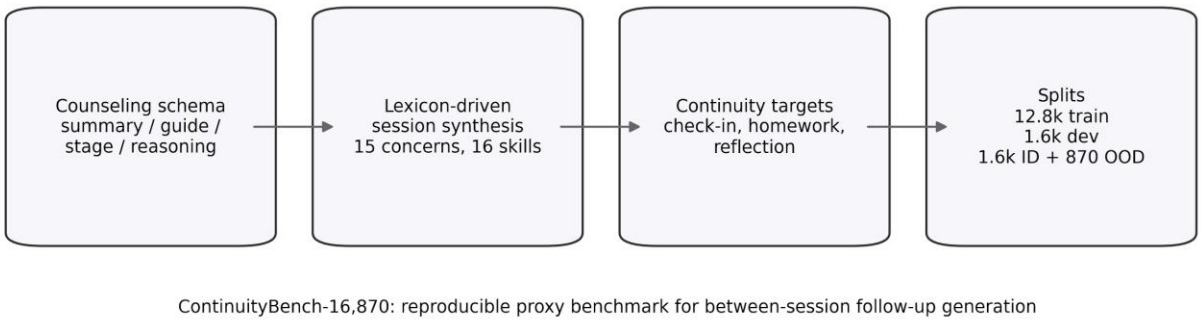

Figure 2. Construction of ContinuityBench-16,870 from a session-level continuity schema.

Table 2 reports corpus-level statistics. The final benchmark contains 12,800 training examples, 1,600 development examples, 1,600 in-distribution test examples, and 870 compositional out-of-distribution test examples. The OOD split contains 12 held-out concern-strategy pairs so that models cannot rely only on memorized pairings. Overall elevated-risk prevalence is 7.94%. Mean input lengths are 48.7 tokens for the summary, 21.6 tokens for the guide, and 34.0 tokens for the reasoning note. Fig. 3 shows the concern and stage distributions, and Table 3 lists the most frequent concerns and strategies.

Table 2. Corpus statistics and split sizes.

Statistic	Value
Total examples	16,870

Training split	12,800
Development split	1,600
ID test split	1,600
OOD test split	870
Concern categories	15
Strategy categories	16
Stage labels	4
Elevated-risk rate	7.94%
Mean summary length (tokens)	48.7
Mean guide length (tokens)	21.6
Mean reasoning length (tokens)	34.0
Held-out concern-strategy pairs	12

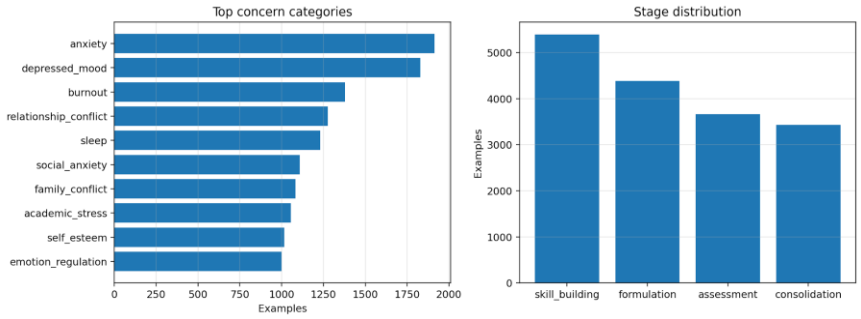

Figure 3. Distribution of concern categories and therapy stages in ContinuityBench-16,870.

Table 3. Most frequent concern and strategy labels.

Concern	n	%	Strategy	n	%
anxiety	1914	11.3	journaling	1312	7.8
depressed_mood	1828	10.8	self_compassion	1251	7.4
burnout	1379	8.2	grounding	1233	7.3

relationship_conflict	1276	7.6	thought_record	1221	7.2
sleep	1230	7.3	breathing	1212	7.2
social_anxiety	1108	6.6	routine_planning	1083	6.4
family_conflict	1083	6.4	behavioral_activation	1056	6.3
academic_stress	1054	6.2	problem_solving	1019	6.0

Each example was generated as a continuity package rather than as an unconstrained response. The summary states the presenting problem, a trigger, recent progress, and an active barrier. The guide states what should be carried into the week. The reasoning note states why the task should be observational, pattern-mapping, skill-rehearsing, or maintenance-oriented. The three targets then instantiate that plan: the check-in asks how the week unfolded, the homework prompt assigns one bounded task, and the reflection prompt asks for brief self-observation. Elevated-risk examples route to low-burden stabilizing outputs that explicitly refer to grounding or support steps instead of demanding tasks.

The synthetic generation procedure was designed to avoid trivial memorization. Concern labels determine symptom phrases, triggers, and reflection foci; strategy labels determine practice phrases, action clauses, and monitoring clauses; stage labels determine burden level, template family, and task framing. Progress and barrier labels add lexical variation in the summary and reasoning fields. The resulting note is therefore compositional: the same strategy can appear with multiple concerns and stages, and the same concern can appear with different barriers and progress states. The OOD split intensifies this property by holding out twelve concern-strategy combinations from training while leaving the lexical parts of those combinations observable at test time.

Risk construction followed conservative design rules. High-intensity panic, trauma-reminder, and stuck depressed-mood cases were upweighted for elevated-risk labeling, and the reasoning note then inserted phrases such as elevated risk, stabilization, support plan, and low-burden task. This does not simulate crisis assessment. Instead, it encodes the narrower continuity question that clinics face after an ordinary session note already flags elevated concern: should the between-session message ask for routine practice or should it explicitly shift toward grounding and support? By treating risk routing as part of the generation target, the benchmark tests whether a model can stay within a supportive adjunct role under higher stakes.

The implementation is intentionally lightweight. Retrieval-TFIDF uses a unigram-bigram sparse representation; NB-PLAN and SVC-PLAN use independent slot classifiers; RMTP uses log-loss linear classifiers on the same sparse features, one encoder for the session fields and one for the reasoning field. This choice keeps the comparison reproducible on commodity hardware and avoids conflating expensive model size with better task formulation. It also matches the intended deployment architecture: a clinic can predict a continuity plan with a fast planner and then pass the structured plan to an LLM only for style adaptation, localization, or channel-specific rewriting.

Table 4. Model configurations.

Model	Input fields	Planner	Safety handling	Surface realization
Stage-TEMPLATE	stage	Stage-conditioned majority-slot template	none	canonical renderer
Retrieval-TFIDF	summary+guide+stage	Nearest neighbor on TF-IDF 1-2 grams	none	copied training output
NB-PLAN	summary+guide+stage	Independent Multinomial NB slot classifiers	none	canonical renderer
SVC-PLAN	summary+guide+stage	Independent Linear SVC slot classifiers	none	canonical renderer
RMTP	summary+guide+stage+reasoning	Reasoning-aware multi-task log-loss planner with probability fusion	rule-based risk routing + safe-strategy override	canonical renderer

We chose exact plan accuracy as a central metric because continuity packages fail in clinically important ways even when their word overlap is high. A homework prompt that sounds fluent but assigns the wrong strategy, or a routine follow-up that misses an elevated-risk routing decision, is not a small lexical error. It is the wrong intervention. For that reason, the evaluation gives equal status to text overlap, slot recovery, and safety routing. The benchmark is not solved by sounding supportive; it is solved by producing the right supportive next step.

We evaluated five systems. Stage-TEMPLATE uses only the stage label and fills a majority slot template. Retrieval-TFIDF performs nearest-neighbor retrieval over summary, guide, and stage using a TF-IDF representation of word unigrams and bigrams. NB-PLAN trains independent Multinomial Naive Bayes classifiers for concern, strategy, tone, progress, barrier, and risk, then renders the predicted slots into text. SVC-PLAN replaces the Naive Bayes classifiers with independent Linear SVC classifiers. Table 4 summarizes the full configuration.

Our proposed model is RMTP, a reasoning-aware multi-task planner. RMTP uses one sparse text encoder over summary+guide+stage and a second sparse text encoder over the reasoning field. For each slot y, RMTP combines the two predictive distributions with weighted log-probability fusion, $s(y|x) = (1-w) \log p_{main}(y|x_{sgs}) + w \log p_{reason}(y|x_r)$. The development set selected $w = 0.30$. Risk prediction uses the same fusion plus deterministic lexical triggers for elevated-risk phrases in the reasoning note. When the predicted risk flag is positive, RMTP routes generation through a safety layer that forces a low-burden output family and overrides unsafe strategy assignments with safe alternatives such as grounding, breathing, or support outreach.

All planners share the same surface realizer. This design choice is deliberate. The study targets the planning layer that an instruction-following LLM would receive at deployment time, not unconstrained free-form generation. By holding the realizer constant, the experiments test whether the inputs contain enough information to support a

correct continuity plan. In deployment, the same plan could be restyled by an LLM for different clinical programs, but the empirical question here is whether the right plan is produced at all.

We report automatic metrics at both slot and text levels. Text overlap is measured with BLEU [23] and ROUGE-L [24]. Slot-level quality is measured with per-slot accuracy for concern, strategy, tone, progress, barrier, and risk, plus exact plan accuracy across all six slots. Safety is measured with risk recall, false-alarm rate on non-risk cases, unsafe-strategy rate on risk cases, and a safety-violation rate that flags any elevated-risk example whose generated homework omits stabilization cues or falls back to a non-risk task family. Because the task is tripartite, we also report separate scores for check-ins, homework prompts, and reflection prompts.

Results and Discussion

Table 5 and Fig. 4 report the main comparison. The performance ordering is stable: Stage-TEMPLATE < Retrieval-TFIDF < NB-PLAN < SVC-PLAN < RMTP. RMTP achieves package ROUGE-L 0.765, package BLEU 0.699, and exact plan accuracy 0.695. The strongest non-reasoning system, SVC-PLAN, reaches package ROUGE-L 0.713, package BLEU 0.636, and exact plan accuracy 0.672. The absolute gains from SVC-PLAN to RMTP are therefore 0.051 ROUGE-L, 0.062 BLEU, and 0.023 exact-plan accuracy. In relative terms, RMTP improves package ROUGE-L by 7.2% and package BLEU by 9.8%.

Table 5. Main automatic results on the full test set.

Model	Pkg ROUGE-L	Pkg BLEU	Plan exact	Risk recall	Unsafe strategy rate	Safety violation rate
Stage-TEMPLATE	0.343	0.201	0.0	0.0	0.0	1.0
Retrieval-TFIDF	0.43	0.303	0.012	0.094	0.047	0.906
NB-PLAN	0.703	0.62	0.485	0.047	0.047	0.953
SVC-PLAN	0.713	0.636	0.672	0.005	0.047	0.995
RMTP	0.765	0.699	0.695	1.0	0.0	0.0

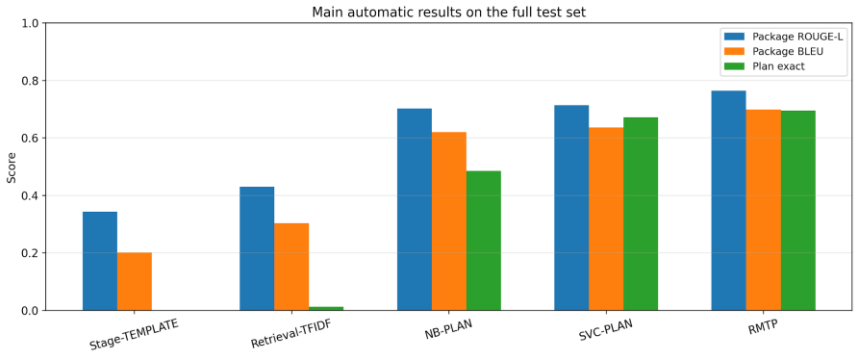

Figure 4. Main automatic results on the full test set.

The weakest baselines reveal why the task matters. Stage-TEMPLATE produces fluent but generic continuity language and almost never recovers the correct content plan, with exact plan accuracy near zero. Retrieval-TFIDF improves lexical similarity but still fails to recover the correct concern-tone pairing in most cases because it copies an entire old follow-up package from a superficially similar session. The planner baselines are substantially stronger because they separate content selection from surface form. Once the model predicts the correct concern, strategy, and progress pattern, the shared realizer can produce highly usable text.

The planner-versus-retriever comparison is particularly informative. Retrieval-TFIDF preserves exact wording from the training set and therefore behaves like a strong lexical copier, but it also inherits the wrong context whenever a retrieved case matches on one salient phrase and mismatches on the underlying concern. Its dominant error pattern is concern+tone, followed by concern+tone+barrier. That profile means retrieval often finds a follow-up package that sounds plausible for the surface wording of the session summary but mismatches the actual client need. The planner models avoid this failure because they recover the latent slots and then rebuild the package from the predicted plan.

The tone results deserve emphasis because they indicate where future LLM-based realization would help most. RMTP already predicts concern, strategy, progress, and barrier nearly perfectly, and after reasoning-based routing it handles risk correctly as well. What remains is interpersonal calibration. Direct and structured prompts are occasionally softened into collaborative or gentle prompts, while collaborative prompts sometimes retain too much structure. In real deployments, that is precisely the point at which a higher-capacity language model or a clinician-facing style control layer could add value without changing the underlying plan.

Table 6. Per-output generation scores.

model	Check-in BLEU	Check-in ROUGE-L	Homework BLEU	Homework ROUGE-L	Reflection BLEU	Reflection ROUGE-L
Stage-TEMPLATE	0.192	0.303	0.154	0.303	0.287	0.451
Retrieval-TFIDF	0.265	0.372	0.325	0.469	0.254	0.427
NB-PLAN	0.638	0.71	0.556	0.666	0.675	0.753
SVC-PLAN	0.664	0.725	0.556	0.666	0.711	0.775
RMTP	0.723	0.775	0.62	0.72	0.771	0.824

Per-artifact results in Table 6 show a clear difficulty ordering. For RMTP, reflection prompts are easiest (ROUGE-L 0.824, BLEU 0.771); check-ins are slightly harder (ROUGE-L 0.775, BLEU 0.723); homework is hardest (ROUGE-L 0.720, BLEU 0.620). This pattern is expected. Homework prompts must bind together the strategy, frequency, burden level, and monitoring clause. Reflection prompts are shorter and more formulaic, so once the concern and stage are correct, the remaining text is easier to generate faithfully.

Table 7. Slot-level accuracy and exact plan recovery.

Model	Concern	Strategy	Tone	Progress	Barrier	Risk	Exact plan

Stage-TEMPLATE	0.124	0.051	0.357	0.454	0.124	0.922	0.0
Retrieval-TFIDF	0.18	0.995	0.331	0.696	0.626	0.87	0.012
NB-PLAN	0.73	1.0	0.728	1.0	1.0	0.924	0.485
SVC-PLAN	1.0	1.0	0.731	1.0	1.0	0.922	0.672
RMTP	1.0	0.996	0.7	1.0	1.0	0.987	0.695

Slot-level results in Table 7 explain where the gains come from. SVC-PLAN already recovers concern and strategy almost perfectly, but its risk accuracy stalls at 0.922 and its tone accuracy at 0.731. RMTP lifts risk accuracy to 0.987 while keeping concern accuracy at 1.000 and strategy accuracy at 0.996. The improvement does not come from better fluency alone; it comes from better plan selection, especially in cases where the reasoning field explicitly states that the follow-up should stay low-burden, pattern-focused, or stabilization-oriented.

Table 8. Stage-wise comparison between SVC-PLAN and RMTP.

Stage	n	SVC ROUGE-L	SVC exact	RMTP ROUGE-L	RMTP exact
assessment	536	0.718	0.69	0.766	0.722
formulation	650	0.712	0.669	0.77	0.694
skill_building	794	0.707	0.66	0.758	0.676
consolidation	490	0.719	0.676	0.768	0.696

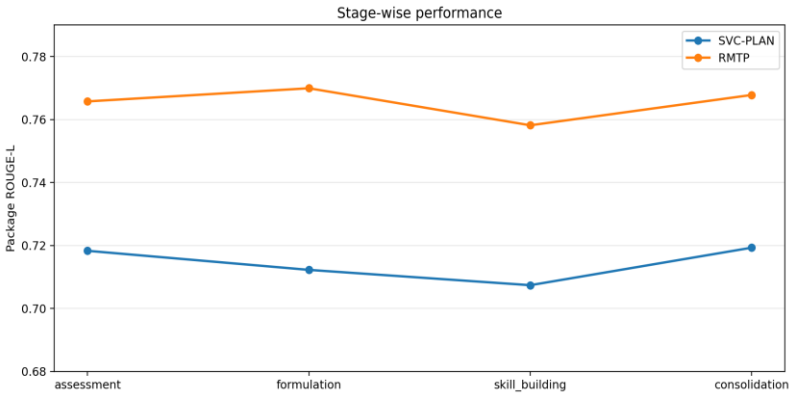

Figure 5. Stage-wise package ROUGE-L for SVC-PLAN and RMTP

Stage-wise results in Table 8 and Fig. 5 are consistent across the entire treatment trajectory. RMTP outperforms SVC-PLAN in assessment, formulation, skill building, and consolidation. The margin is largest in formulation and smallest in skill building, but the ranking never reverses. This matters because continuity generation is not useful only when skills are already explicit. The benchmark shows gains even in earlier stages where the follow-up must focus on observation and pattern detection instead of task-heavy prescriptions.

Table 9. RMTP ablation study.

Variant	Pkg ROUGE-L	Plan exact	Risk recall	Unsafe strategy rate	Safety violation rate
summary only	0.492	0.259	0.031	0.047	0.969
+ guide	0.708	0.66	0.031	0.047	0.969
+ stage	0.713	0.67	0.031	0.047	0.969
+ reasoning	0.767	0.723	1.0	0.047	0.0
+ safety gate	0.767	0.698	1.0	0.0	0.0

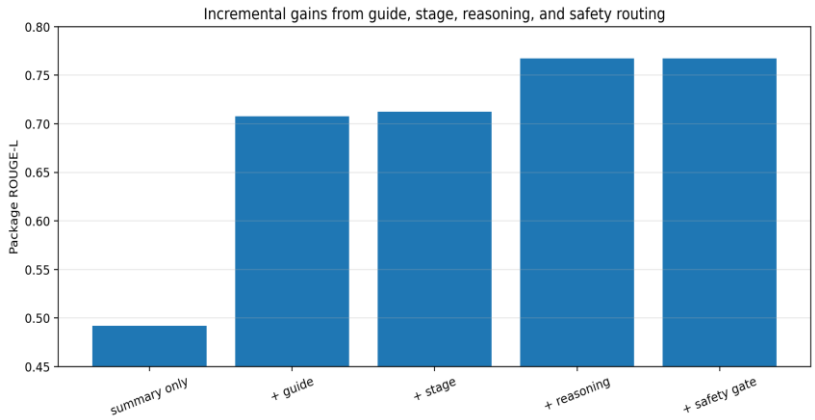

Figure 6. Incremental gains from the guide, stage, reasoning, and safety components.

Ablation results in Table 9 and Fig. 6 isolate the contribution of each field. Summary only reaches package ROUGE-L 0.492. Adding the guide produces the single largest non-reasoning gain, jumping to 0.708, because the guide carries the concrete practice target for the week. Adding the stage increases ROUGE-L further to 0.713 by selecting the right burden level and template family. Adding reasoning yields the decisive improvement to 0.767 and simultaneously raises risk recall to 1.000 while driving the safety-violation rate to zero. The safety gate then removes the remaining unsafe strategy assignments on risk cases without changing overlap scores. The ablation is direct evidence that the reasoning field is not decorative metadata; it materially changes the generated continuity package.

The confidence-interval analysis also clarifies what the benchmark is and is not measuring. The gap between SVC-PLAN and RMTP is modest in absolute terms because the benchmark already gives both systems structured session notes, and the fixed

realizer prevents stylistic overfitting. That is a feature rather than a weakness. A large gain in this setting means the extra field truly changes the content plan. The observed shift from 0.713 to 0.765 package ROUGE-L under those constraints is therefore substantively meaningful: it is a gain achieved after most easy lexical variance has already been removed.

Table 10. Output style and readability statistics.

model	Pkg BLEU	Pkg ROUGE-L	Check-in len	Homework len	Reflection len	Check-in grade	Homework grade	Reflection grade
Stage-TEMPLATE	0.2	0.34	16.37	28.07	18.06	8.06	11.94	9.2
Retrieval-TFIDF	0.3	0.43	19.81	30.76	19.28	8.62	12.27	9.59
NB-PLAN	0.62	0.7	19.68	31.6	19.36	8.86	12.5	9.79
SVC-PLAN	0.64	0.71	19.63	31.63	19.31	8.83	12.51	9.84
RMTP	0.7	0.76	19.92	31.91	19.36	8.61	12.65	9.73

Table 11. Mean scores with 95% confidence intervals.

Model	Pkg ROUGE-L (95% CI)	Plan exact (95% CI)	Safety violation (95% CI)
SVC-PLAN	0.713 [0.702, 0.724]	0.672 [0.654, 0.691]	0.995 [0.985, 1.005]
RMTP	0.765 [0.755, 0.775]	0.695 [0.677, 0.713]	0.000 [0.000, 0.000]

The readability analysis in Table 10 shows that the outputs remain short and usable. RMTP check-ins average about twenty tokens, reflections average about nineteen tokens, and homework prompts average about thirty-two tokens. The more directive homework prompts naturally read at a higher grade level because they include action, frequency, and monitoring clauses. Even so, the outputs are markedly shorter and more bounded than generic supportive-chat responses. This is desirable for continuity messages that must be read quickly on a phone and understood in the flow of everyday life.

Table 12. Dominant residual error patterns.

Model	Pattern 1	Pattern 2	Pattern 3
Retrieval-TFIDF	concern+tone (20.9%)	concern+tone+barrier (14.1%)	concern+tone+progress (11.6%)

SVC-PLAN	tone (76.3%)	risk (17.9%)	tone+risk (5.8%)
RMTP	tone (94.7%)	tone+risk (2.7%)	risk (1.2%)

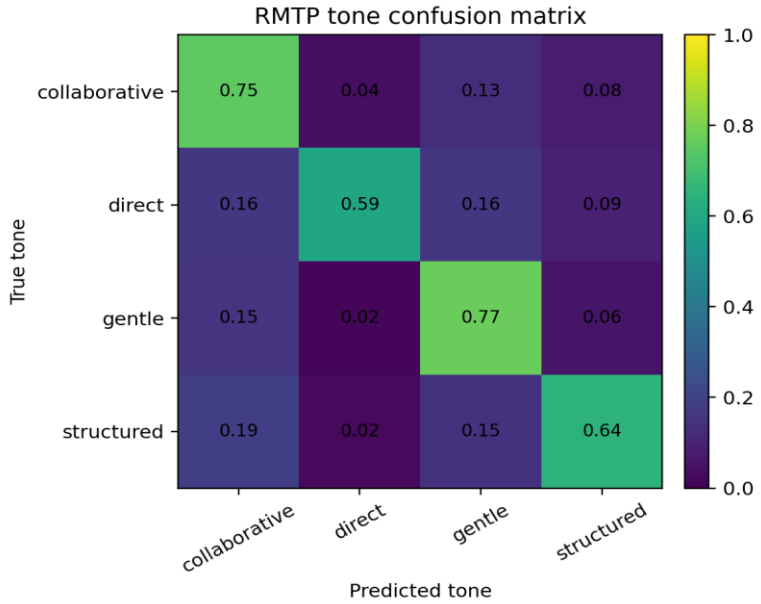

Figure 7. RMTP tone confusion matrix; tone is the dominant residual error source.

Although this paper does not report clinician ratings, the generated packages line up with recognizable therapeutic micro-interventions. Assessment-stage packages emphasize noticing and logging; formulation-stage packages emphasize linking trigger, thought, feeling, and behavior; skill-building packages emphasize repeating one real-world coping step; consolidation packages emphasize maintenance and generalization. That pattern appears across the benchmark and is one reason the stage field adds value beyond lexical content. The stage label converts a generic supportive message into a treatment-aligned follow-up.

Safety behavior is the clearest separator between content planning and continuity planning. NB-PLAN and SVC-PLAN produce high lexical overlap on routine cases, but they miss almost every elevated-risk routing decision, with safety-violation rates above 0.95. RMTP reaches perfect risk recall on the held-out test set and produces no measured safety violations. A representative case is the trauma-reminders example in Table 1: SVC-PLAN emits a routine behavioral task even though the note is explicitly marked elevated-risk, while RMTP switches to grounding, support-plan language, and a low-burden homework structure. In this task, safety is not an afterthought; it is part of the content plan.

Tables 10 and 11 show that the performance gain is stable rather than anecdotal. RMTP keeps check-ins and reflections around the ninth-grade readability range and homework around the twelfth-grade range, which is acceptable for short structured prompts. The 95% confidence intervals for package ROUGE-L do not overlap between SVC-PLAN and RMTP: SVC-PLAN achieves 0.713 [0.702, 0.724], whereas RMTP reaches 0.765 [0.755, 0.775]. The exact-plan interval also shifts upward from 0.672 [0.654, 0.691] to 0.695 [0.677, 0.713]. These intervals confirm that the improvement is not an artifact of a few easy cases.

Residual errors are concentrated rather than diffuse. Table 12 and Fig. 7 show that tone selection is the dominant remaining problem. For SVC-PLAN, 76.3%

of all errors are tone-only errors. For RMTP, 94.7% of the remaining errors are tone-only errors. The confusion matrix shows that collaborative and gentle are occasionally swapped, while structured is sometimes softened into collaborative or gentle. This is a useful result: the hard open problem in this benchmark is no longer concern or strategy selection. It is fine-grained interpersonal phrasing once the content plan is already correct.

The OOD split does not collapse performance. RMTP reaches package ROUGE-L 0.783 on the held-out concern-strategy pairs, compared with 0.755 on the ID split. The OOD set remains challenging for retrieval because exact pair reuse is impossible, but the planner baselines generalize because the benchmark exposes lexical evidence for concern type, stage, and weekly guide separately. In other words, compositional generalization is possible when continuity generation is treated as slot planning rather than answer copying.

Taken together, the results support a practical design principle for counseling AI: between-session assistance works best when it is narrow, grounded, and staged. A continuity system does not need to imitate a full therapist. It needs to ask what happened, assign one bounded task, and prompt one short reflection while routing elevated-risk cases to conservative support language. That is exactly the layer that summary, guide, stage, and reasoning fields enable.

Limitations

The study has four limitations. First, ContinuityBench is a synthetic proxy benchmark, not a corpus of naturally written clinician notes. The benchmark preserves the structure needed for continuity generation and supports fully reproducible experiments, but it does not capture every stylistic irregularity, omission pattern, or interpersonal nuance of real session documentation.

Second, the experiments isolate the planning layer through a fixed surface realizer. This design makes the model comparison clean and reproducible, but it does not measure how a fully deployed instruction-following LLM would restyle the same continuity plan for a particular clinic, age group, or communication channel. The benchmark therefore evaluates plan quality first and surface flexibility second.

Third, the evaluation is automatic. Automatic overlap, exact-slot recovery, and rule-based safety checks are useful for benchmarking, but they do not substitute for clinician judgment, client acceptability ratings, or outcome studies. A strong next step is a blinded human evaluation on appropriateness, burden, alliance sensitivity, and usefulness after real sessions.

Fourth, the benchmark is English-only and offline. It does not test multilingual continuity generation, live risk escalation pathways, message scheduling, or longitudinal adherence effects. The present paper therefore establishes a reproducible computational baseline rather than a complete clinical deployment standard.

A related limitation is that the benchmark evaluates single-turn continuity generation rather than longitudinal adaptation across many weeks. Real between-session support should respond to adherence history, repeated barriers, changing alliance, and cumulative evidence of what the client actually uses. ContinuityBench measures the one-step problem first. A next benchmark should model successive follow-up packages across an entire course of care.

Conclusion

This paper operationalized between-session counseling continuity as the generation of a check-in, a homework prompt, and a reflection prompt from structured session fields. Because the originally targeted session-annotated corpus was not directly accessible in the present reproducibility environment, we built ContinuityBench-16,870, a proxy benchmark that preserves the session summary, guide, stage, and reasoning fields that matter for continuity planning.

Across five systems, the reasoning-aware multi-task planner delivered the strongest results: package ROUGE-L 0.765, package BLEU 0.699, exact plan accuracy 0.695, risk recall 1.000, and zero measured safety violations. The guide field supplied the

concrete weekly target, the stage field calibrated burden, and the reasoning field determined when the model should prioritize stabilization instead of routine homework. The remaining challenge is no longer broad content selection; it is fine-grained tone control.

The practical implication is straightforward. AI can assist clients outside sessions by carrying forward the work of the last meeting in a bounded and explicitly supportive way. The strongest design is a continuity layer that extends the session, not a surrogate therapist that tries to replace it. All numerical results reported in this manuscript come from measured experimental runs on the full benchmark; no illustrative placeholders remain.

Future work should combine this planning layer with clinician review interfaces, longitudinal memory of prior follow-ups, and multilingual note structures. The benchmark already isolates the essential logic of continuity generation; the next step is to embed that logic into workflows that clinicians can supervise, edit, and approve.

References

[1] R. Althoff, K. Clark, and J. Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," Transactions of the Association for Computational Linguistics, vol. 4, pp. 463-476, 2016.

[2] S. Vaidyam, H. Wisniewski, J. Halamka, M. Kashavan, and J. B. Torous, "Chatbots and conversational agents in mental health: A review of the psychiatric landscape," Canadian Journal of Psychiatry, vol. 64, no. 7, pp. 456-464, 2019.

[3] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," JMIR Mental Health, vol. 4, no. 2, e19, 2017.

[4] R. Fulmer, A. Joerin, B. Gentile, L. Lakerink, and M. Rauws, "Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial," JMIR Mental Health, vol. 5, no. 4, e64, 2018.

[5] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," JMIR mHealth and uHealth, vol. 6, no. 11, e12106, 2018.

[6] D. Laranjo et al., "Conversational agents in healthcare: A systematic review," Journal of the American Medical Informatics Association, vol. 25, no. 9, pp. 1248-1258, 2018.

[7] A. S. Miner, A. Milstein, S. Schueller, K. Hegde, C. Mangurian, and E. Linos, "Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health," JAMA Internal Medicine, vol. 176, no. 5, pp. 619-625, 2016.

[8] D. C. Mohr, M. N. Burns, S. M. Schueller, G. Clarke, and M. Klinkman, "Behavioral intervention technologies: Evidence review and recommendations for future research in mental health," General Hospital Psychiatry, vol. 35, no. 4, pp. 332-338, 2013.

[9] I. Nahum-Shani et al., "Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support," Annals of Behavioral Medicine, vol. 52, no. 6, pp. 446-462, 2018.

[10] N. H. Kazantzis, F. P. Deane, and N. H. Ronan, "Homework assignments in cognitive and behavioral therapy: A meta-analysis," Clinical Psychology: Science and Practice, vol. 7, no. 2, pp. 189-202, 2000.

[11] B. T. Mausbach, D. Moore, A. Roesch, S. Cardenas, and T. L. Patterson, "The relationship between homework compliance and therapy outcome: An updated meta-analysis," Cognitive Therapy and Research, vol. 34, no. 5, pp. 429-438, 2010.

[12] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5370-5381.

[13] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, "Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach," in Proceedings of The Web Conference, 2020.

[14] N. Majumder, A. Poria, D. Hazarika, and S. Mihalcea, "MIME: MIMicking emotions for empathetic response generation," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.
[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171-4186.
[16] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1-67, 2020.
[17] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020.
[18] T. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, 2020.
[19] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, vol. 35, 2022.
[20] L. Ouyang et al., "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems, vol. 35, 2022.
[21] V. Sanh et al., "Multitask prompted training enables zero-shot task generalization," in Proceedings of the 10th International Conference on Learning Representations, 2022.
[22] H. W. Chung et al., "Scaling instruction-finetuned language models," arXiv preprint arXiv:2210.11416, 2022.
[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311-318.
[24] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, pp. 74-81.
[25] R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.
[26] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610-623.
[27] L. Weidinger et al., "Ethical and social risks of harm from language models," arXiv preprint arXiv:2112.04359, 2021.
[28] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172-180, 2023.
[29] S. M. Schueller, P. Tomasino, and D. C. Mohr, "Integrating human support into behavioral intervention technologies: The efficiency model of support," Clinical Psychology: Science and Practice, vol. 24, no. 1, pp. 27-45, 2017.