RAVR-S: State-Sensitive Verification and Repair for Trustworthy Rule-Governed LLM Dialogue

Anonymous Authors[1]

Abstract

Large language models produce fluent therapeutic responses that frequently violate explicit domain constraints, especially across multi-turn dialogues in which the patient's state evolves. Existing refinement strategies (self-critique, diversity sampling, free-form self-refine) raise surface quality but give no transparency about which constraints they satisfy or violate, and they ignore dialogue-level state dynamics.

We introduce RAVR-S, a *state-sensitive* verification-and-repair framework for rule-governed LLM dialogue. At each turn, a structured verifier scores the candidate response against a typed 58-predicate inventory and emits a proof object that lists satisfied and violated constraints. RAVR-S extends the base verify-and-repair loop (RAVR) with a state-transition estimator: a separate LLM call tracks a discrete patient-state vector $(\text{trust}, \text{distress}, \text{fatigue}) \in \{0,1,2,3\}^3$, scores $K{=}3$ response candidates against the predicted state dynamics, and picks the best candidate before optional targeted repair.

We evaluate RAVR-S in a two-stage human study. The Stage 1 screening (10 annotators, 1,440 judgments) places Self-Refine first at 73.5% mean win rate and RAVR second at 60.7%, with Regenerate$\times$3 (34.3%) and Vanilla (31.3%) trailing. Stage 2 narrows the comparison: 20 annotators with an advanced or doctoral degree in psychology (credential-verified) compare RAVR-S against the two strongest baselines on state-sensitive 3-turn therapeutic mini-dialogues. RAVR-S wins $\mathbf{94.7}\%$ of comparisons (excluding ties), at almost-perfect agreement (Gwet's AC1 $= 0.82$).

Automated evaluation on **TherapyBench**, a new public benchmark of 288 multi-turn trajectories across four therapy modalities, shows that RAVR-S holds 90% **trajectory adherence** and a 100% **policy-compliance rate** across 8-turn sessions while keeping iatrogenic pressure at 4.2%. Self-Refine moves the other way under state pressure: its policy-switch rate falls to 16.7%, below the 50% Vanilla baseline, and its pressure rate is roughly $9\times$ higher than RAVR-S (0.375 vs. 0.042). Single-turn repair on $1{,}500$ turns confirms consistent adherence gains ($+16.7$ points, $p{<}0.001$), convergence in a single repair iteration, and interpretable predicate-level diagnostics. TherapyBench is released at `https://anonymous.4open.science/r/TherapyBench-D547`.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

LLM-based therapeutic training simulators produce fluent but methodologically inconsistent responses, especially when multiple therapeutic frameworks coexist in one system (Demszky et al., 2023; Sharma et al., 2024). For clinical education, this creates a fundamental mismatch: the assistant may *sound* clinically appropriate while violating core constraints of the target methodology (e.g., introducing exposure-based techniques in a stabilization-phase EMDR session, or asking multiple direct questions in a non-directive Rogerian exchange).

Existing refinement approaches fall into three categories. *Self-Refine* methods (Madaan et al., 2023) apply free-form self-critique loops but lack structured constraint tracking. *Diversity sampling* (Wang et al., 2023) generates multiple candidates and selects by majority vote, but ignores domain-specific predicates. *Constitutional AI* approaches (Bai et al., 2022) embed principles into training but offer no per-turn interpretability. To our knowledge, no prior framework unifies (i) predicate-level proof objects, (ii) state-aware multi-turn optimization, and (iii) violation-conditioned targeted repair, despite individual progress in tool-augmented reasoning (Yao et al., 2023b), constrained decoding, and inference-

time safety guardrails (Rebedea et al., 2023).

We propose RAVR-S (**R**etrieval-**A**ugmented **V**erification and **R**epair with **S**tate-sensitive selection), a framework that addresses all three gaps. Our contributions are:

1. A **predicate-level verification** schema producing structured proof objects (violations, adherence scores, citation grounding) per turn (§3.2).
2. A **state-sensitive candidate selection** mechanism that scores candidates against the evolving patient-state trajectory (trust, distress, therapeutic progression) (§3.3).
3. A **two-stage human evaluation** protocol (N=10 screening + N=20 focused) with 3,040 total judgments, demonstrating 94.7% human preference for RAVR-S over the two strongest baselines (§5).
4. **Comprehensive automated evaluation** across 1,500 single-turn outputs, 288 multi-turn trajectories (TherapyBench), policy-switch accuracy analysis, long-horizon stability tests (h=2–8), and six ablation conditions: we find that Self-Refine's pressure rate is 2× higher than Vanilla and its policy-switch accuracy falls *below* random (§6).
5. A **public benchmark release** (TherapyBench) including 58 typed predicates across five therapeutic methodologies, 16 state-sensitive mini-dialogue pairs, 1,500-turn evaluation outputs, and all human evaluation data (§6, `https://anonymous.4open.science/r/TherapyBench-D547`).

2. Related Work

LLM self-refinement. Madaan et al. (2023) show that iterative self-critique loops improve LLM output quality across diverse tasks. However, the critique signal is free-form text, making it impossible to verify *which specific constraints* were addressed in any given revision. Shinn et al. (2023) augment refinement with episodic verbal memory, enabling longer-horizon task completion, but still without predicate-level accountability. Bai et al. (2022) embed revision principles into training-time weights (Constitutional AI), forgoing per-turn inference-time interpretability. Our work differs by conditioning repair on an exact violation inventory: the repair module receives a machine-readable set of violated predicates and generates only the targeted fix, achieving full adherence gains in a single step (vs. 2–5 rounds for free-form refinement or multi-step deliberative search) (Yao et al., 2023a; Zhou et al., 2023).

Constrained and rule-governed generation. Alignment via RLHF (Ouyang et al., 2022) and Constitutional AI embed constraints into model weights, complementing but not replacing inference-time verification. Domain-specific rule-following has been studied in task-oriented dialogue (Budzianowski et al., 2018), where belief-state tracking enforces slot-value consistency. Constrained decoding methods such as NeuroLogic A*esque (Lu et al., 2022) and PICARD (Scholak et al., 2021) enforce lexical or structural constraints at the token level, but cannot express the multi-sentence semantic predicates required for clinical protocol adherence. Retrieval-augmented generation (Lewis et al., 2020) and tool-augmented reasoning (Schick et al., 2023; Yao et al., 2023b) supply external knowledge but do not verify post-generation adherence. RAVR-S closes this gap via citation-grounded predicate verification and violation-conditioned repair.

LLM-based evaluation and verification. Zheng et al. (2023) and Liu et al. (2023) establish LLM-as-a-judge as a reliable evaluation paradigm for open-ended generation, approaching human inter-annotator agreement on quality dimensions. Liang et al. (2023) extend structured evaluation to holistic multi-scenario benchmarking (HELM). Our structured verifier adapts this insight to inference-time constraint checking, replacing scalar quality scores with a machine-readable predicate violation inventory that drives targeted single-step repair rather than re-generation from scratch.

LLMs in therapeutic and mental health contexts. Demszky et al. (2023) survey LLM applications across psychology, identifying therapeutic training and mental health support as high-impact use cases but leaving methodology-specific constraint enforcement largely open. Sharma et al. (2024) demonstrate that human-LLM interaction can scaffold CBT-style cognitive restructuring for self-guided mental health interventions. Guo et al. (2024) survey LLM applications in mental health in a comprehensive systematic review, highlighting safety and ethics as open challenges. Chen et al. (2023) evaluate LLM-empowered chatbots for psychiatrist and patient simulation in clinical settings, but without methodology-specific multi-turn constraint enforcement. None of these systems, however, combine per-turn proof objects with methodology-specific predicate enforcement (e.g., EMDR phase sequencing, DBT dialectical balance). Our work targets *auditable* methodology adherence across 10 clinical directions with an explicit, inspectable constraint record per turn.

Multi-turn dialogue state and strategy. State tracking in task-oriented dialogue operates over slot-value belief states for booking/query tasks. More broadly, work on LLM-based planning, memory, and tool use (Yao et al., 2023b; Park et al., 2023; Shinn et al., 2023) emphasises long-horizon coherence, but does not explicitly model domain-specific latent states such as patient trust, distress, and fatigue, nor couples state-aware selection with formal constraint

verification. Therapeutic dialogue requires a fundamentally different state abstraction: these affective dimensions evolve continuously and determine which interventions are appropriate *regardless of predicate satisfaction*. RAVR-S introduces a three-dimensional therapeutic state vector and uses it to score candidates, addressing a gap between slot-filling dialogue management and open-domain clinical education.

3. The RAVR-S Pipeline

RAVR-S operates as a four-stage pipeline per dialogue turn (Figure 1).

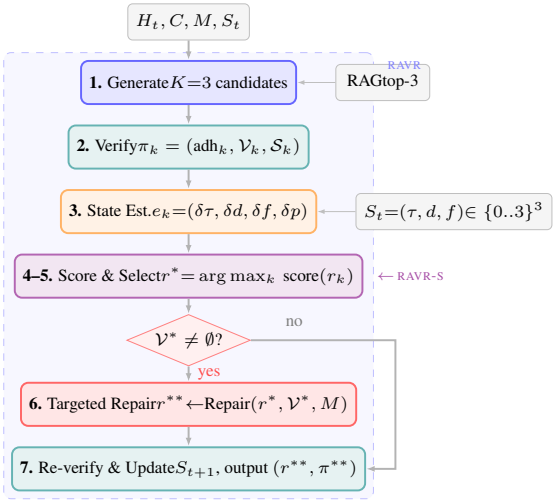

Figure 1. The RAVR-S pipeline (one dialogue turn). Stages 1–2 and 6–7 (blue outline) form the base RAVR loop. Stages 3–5 (orange/violet) add state-sensitive selection. The repair branch (red) triggers only when mandatory predicates are violated.

3.1. Stage 1: Candidate Generation

Given dialogue history H_t, patient case C with methodology M, and current state $S_t = (\text{trust}_t, \text{dist}_t, \text{fat}_t) \in \{0, 1, 2, 3\}^3$, we generate K=3 response candidates via temperature sampling (T=0.7, top-p=0.9).

Each candidate is generated with a context window containing: (i) the patient persona (symptom profile, personality style, linguistic register, trust/distress/fatigue initialisation from the case definition); (ii) retrieved methodology-specific guidelines via RAG (top-3 passages per turn, BM25+semantic reranking over a per-methodology knowledge base of clinical guidelines and intervention descriptions); and (iii) the current state vector S_t prepended to the generation prompt to enable state-conditioned responses. Diversity across candidates is encouraged by seeding each of the K calls with independently sampled temperature noise. No explicit diversity penalty or coverage objective is applied. Diversity emerges from stochastic sampling.

3.2. Stage 2: Predicate-Level Verification

Each candidate r_k is evaluated against the typed predicate inventory $\mathcal{P}_M = \{p_1, \ldots, p_n\}$ for methodology M. The verifier is an LLM (same provider as generation, but a separate API call with a fixed structured prompt). For each predicate p_i, the verifier returns a binary or ordinal verdict conditioned on the predicate's natural-language description from the inventory. Verdicts are aggregated into a *proof object*:

$$\pi_k = \big(\text{adh}_k,\ \mathcal{V}_k,\ \mathcal{S}_k,\ \text{cit}_k\big) \tag{1}$$

where $\text{adh}_k \in [0, 100]$ is a weighted adherence score (mandatory predicates count double), $\mathcal{V}_k \subseteq \mathcal{P}_M$ and $\mathcal{S}_k \subseteq \mathcal{P}_M$ are the violated and satisfied predicate sets, and $\text{cit}_k = (\text{coverage}, \text{precision}, \text{relevance})$ contains citation grounding metrics checked against the retrieved passages. A candidate passes verification ($\text{pass}_k = 1$) if and only if $\mathcal{V}_k \cap \mathcal{P}_M^{\text{mand}} = \emptyset$, i.e., no mandatory predicate is violated.

Predicates span five families. *Method-focus* predicates (mandatory) encode methodology-specific technique constraints, for example one that requires the response to address the patient's current EMDR processing phase rather than skipping ahead to trauma reprocessing while still in stabilisation. *Directivity* predicates (mandatory and recommended) cap the number of questions per turn (≤ 2), prohibit unsolicited advice, and discourage categorical directives. *Safety* predicates (mandatory) handle crisis detection, boundary maintenance, and scope limitation (no prescriptions, no diagnoses). *Collaborative-stance* predicates (recommended) cover warmth, validation, and autonomy support, and *citation-grounding* predicates (mandatory) require evidence references to map to retrieved passages with coverage ≥ 0.5. The full 58-predicate specification is bundled with the TherapyBench release.

3.3. Stage 3: State-Sensitive Selection

The patient state at turn t is a three-dimensional integer vector $S_t = (\text{trust}_t, \text{dist}_t, \text{fat}_t) \in \{0, 1, 2, 3\}^3$, initialised from the patient case profile and updated after each verified response, plus a fourth scalar $p_t \in \{0, 1\}$ flagging whether the previous turn imposed premature therapeutic pressure. For each candidate r_k a dedicated LLM evaluator (a separate call, different from the generation model) reads the response in the context of S_t and returns a structured JSON object with predicted discrete transitions

$$e_k = \big(\delta\text{trust}_k,\ \delta\text{dist}_k,\ \delta\text{fat}_k,\ \delta p_k\big) \in \{-1, 0, 1\}^4$$

along with binary quality signals EM_k (expected method move present) and PM_k (premature or misaligned move). The evaluator prompt specifies each field with an exact schema, and outputs are parsed deterministically. The com-

Method	w_τ	w_d	w_f	w_p
EMDR	1.6	3.0	1.0	1.8
DBT	2.0	2.8	1.0	1.8
CBT	1.8	2.4	1.0	1.4
ABA	1.8	2.2	1.0	1.2
Hard-state modifier (dist$\geq$2): $+0.4$ to w_d; $+0.5$ to w_p.				

Table 1. Method-specific scoring weights (w_τ: trust gain, w_d: distress reduction, w_f: fatigue cost, w_p: pressure penalty). EMDR and DBT most strongly upweight distress protection, reflecting stabilization-first protocols. Hard-state modifier activates when patient distress ≥ 2.

posite selection score is:

$$\begin{aligned}\text{score}(r_k) = w_\tau \delta\tau_k - w_d \delta d_k \\ - w_f \delta f_k - w_p \delta p_k \\ - 3(1-\text{EM}_k) - 3\,\text{PM}_k\end{aligned} \tag{2}$$

where weights (w_τ, w_d, w_f, w_p) are methodology-specific (Table 1) and a hard-state modifier boosts w_d and w_p by 0.4 and 0.5 respectively when S_t indicates a fragile patient (dist ≥ 2). The -3 penalties for missing the expected methodological move or triggering a premature move enforce hard constraints on therapeutic sequencing that cannot be overridden by adherence scores alone.

The candidate $r^* = \arg\max_k \text{score}(r_k)$ is selected. This enables the system to prefer, for example, a slightly lower single-turn adherence candidate that avoids destabilizing a fragile patient (dist $= 2$, pressure risk $= 1$) over a high-adherence candidate that would introduce premature trauma processing.

3.4. Stage 4: Targeted Repair

If r^* still contains violations ($\mathcal{V}^* \neq \emptyset$), a repair module generates a minimally modified rewrite:

$$r^{**} = \text{Repair}(r^*, \mathcal{V}^*, M, H_t, S_t) \tag{3}$$

The repair prompt explicitly communicates the violated predicate IDs and their natural-language descriptions, and instructs the model to fix *only* those violations while preserving the therapeutic move type, the open-question target, and the length ($\leq$3 sentences). This *targeted* repair differs from free-form self-refine in two key ways: (i) the repair module cannot change aspects of the response that are not implicated in $\mathcal{V}^*$, preserving original intent; and (ii) the repair is a single-pass operation (no iterative loop at this stage), empirically sufficient to saturate available improvement (see §6). The repaired response r^{**} is re-verified to produce a final proof object π^{**}, and S_{t+1} is updated using the state evaluator applied to r^{**}.

4. Experimental Setup

We evaluate on methodology-constrained psychotherapy training dialogue across ten clinical directions: CBT, DBT, EMDR, ABA, Psychodynamic, Rogerian / Person-Centered, Motivational Interviewing, Schema Therapy, Narrative Therapy, and Psychopharmacology-linked protocols. Each direction defines 4–8 typed predicates, 58 in total. For automated evaluation we generate 500 single turns per provider (1,500 in total) across all ten directions with balanced case sampling. For human evaluation we construct 16 three-turn mini-dialogue scenarios in four high-stakes clinical domains (PTSD, panic disorder, schizophrenia, personality disorders), designed to expose state-sensitive failure modes such as premature exposure or pressure escalation. All generation uses three LLM providers through OpenRouter (GPT-4o, DeepSeek-Chat, Qwen-2.5) at temperature 0.7 and top-p 0.9.

We compare RAVR-S against four baselines that span the relevant design space. *Vanilla* is a single-pass generation with RAG context. *Regenerate×3* samples three candidates and picks the one with the highest adherence, without any state awareness. *Self-Refine* runs two iterations of free-form self-critique and rewrite. *RAVR* applies verification and targeted repair without the state-sensitive selection module, and serves as our internal ablation.

5. Human Evaluation

We validate RAVR-S in a two-stage human evaluation. All annotators were recruited from Prolific or CloudResearch, with a psychology background filter at Stage 1 and a stricter graduate-degree filter at Stage 2.

5.1. Stage 1: Screening (N=10)

Ten annotators rate 36 blind pairwise comparisons covering all four methods on four dimensions (overall quality, helpfulness, empathy, safety). Half the panel comes from Prolific with a general psychology-background screening filter, the other half from CloudResearch with a verified graduate degree in psychology. Each pair shows two anonymised single-turn responses labelled "Response A" and "Response B", with a "Tie" option.

Results (Table 2) identify Self-Refine and RAVR as the two strongest baselines, both significantly outperforming Regenerate×3 and Vanilla across all dimensions. Self-Refine excels at empathy (81.5%), while RAVR shows stronger safety scores (62.7% vs 31.5% for Vanilla). Krippendorff's $\alpha = 0.42$ indicates moderate inter-rater agreement, consistent with the subjective nature of therapeutic quality judgments.

Method	Overall	Helpful	Empathy	Safety
Self-Refine	**71.8**	**71.3**	**81.5**	69.2
RAVR	57.9	63.0	59.0	**62.7**
Regen×3	38.0	35.4	32.1	31.5
Vanilla	31.9	29.9	26.9	36.6
Krippendorff's $\alpha = 0.42$ (moderate agreement)				

Table 2. Stage 1 screening: pairwise win rates (%, excl. ties) by dimension. N=10, 1,440 judgments. Bold = best per column. Self-Refine and RAVR are the two strongest methods, selected for Stage 2.

Dimension	RAVR-S	Self-Refine	RAVR
Overall strategy	**94.8**	8.6	1.9
Safety/stability	**94.5**	9.1	1.9
Move ordering	**96.3**	6.3	1.3
Less pressuring	**94.3**	9.4	2.0
Training example	**93.9**	10.4	1.9
All combined	**94.7**	8.8	1.8

Table 3. Stage 2 focused comparison: win rates (%, excl. ties) across five dimensions, N=20 domain-qualified annotators, 1,600 judgments. RAVR-S dominates both baselines on every dimension. Move ordering shows the largest margin, consistent with the 100% policy-switch rate in automated evaluation.

5.2. Stage 2: Focused Comparison (N=20)

Based on the Stage 1 result, we narrow the comparison to RAVR-S against the two strongest baselines (RAVR and Self-Refine). The instrument is 16 state-sensitive three-turn mini-dialogue scenarios. Each scenario shows a matched pair of mini-dialogues (same patient, same initial state) and asks annotators to choose on five dimensions: overall strategy, safety and stability, move ordering, pressure avoidance, and training-example quality. All 20 annotators were recruited from CloudResearch with a verified graduate degree in psychology (master's, doctorate, or equivalent professional credential), giving us a domain-qualified panel of 1,600 judgments from evaluators with formal training in therapeutic methodology.

Results are decisive (Table 3, Figure 2). RAVR-S achieves 94.7% overall win rate, with the strongest advantage on move ordering (96.3%), the dimension most sensitive to state dynamics. Head-to-head, RAVR-S beats RAVR on 97.4% of pairwise comparisons (374 wins, 10 losses, 16 ties) and Self-Refine on 87.7% (321 wins, 45 losses, 34 ties).

Inter-rater agreement. We report Gwet's AC1 (Gwet, 2008) rather than Krippendorff's α, as the latter is known to underestimate agreement under high prevalence (Feinstein & Cicchetti, 1990). AC1 = 0.82 indicates *almost perfect* agreement by the Landis–Koch scale (Landis & Koch, 1977). Pairwise agreement is 82.8%, and the average within-item majority ratio is 90.0% (i.e., 18 of 20 annotators agree on average). All 20 annotators passed the embedded attention check (100% pass rate).

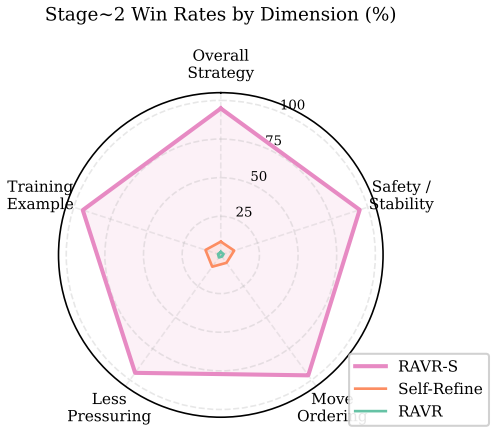

Figure 2. Radar chart of Stage 2 win rates across all five evaluation dimensions (N=20, domain-qualified annotators). RAVR-S (pink) fills nearly the entire area. Both baselines are confined to a narrow inner ring.

Metric	Value
Turns evaluated	1,500 (500 × 3 providers)
Initial pass rate	30.0%
Initial adherence	64.8
Repair trigger rate	67.4%
Post-repair adherence	81.5
ΔAdherence (overall)	+**16.7** (p<0.001)
ΔAdherence by provider:	
GPT-4o	+7.83
DeepSeek	+7.14
Qwen-2.5	+9.50

Table 4. Single-turn repair evaluation (1,500 turns across GPT-4o, DeepSeek-Chat, Qwen-2.5). Pooled ΔAdh. (+16.7) is computed on repair-triggered turns only. Per-provider values are averaged over all 500 turns per provider (including non-triggered). Qwen-2.5 benefits most from repair. 95% CI for pooled ΔAdh.: [15.8, 17.7].

6. Automated Evaluation

6.1. Single-Turn Repair Gains

Across 1,500 turns (500 per provider: GPT-4o, DeepSeek-Chat, Qwen-2.5), repair is triggered on 67.4% of turns. On triggered turns, mean adherence increases from 64.8 to 81.5 ($\Delta = +16.7$, 95% CI [15.8, 17.7], paired permutation $p = 5 \times 10^{-5}$). Per-provider results (Table 4) confirm consistent gains across all three backends. An independent LLM judge blind to verifier internals yields Spearman $\rho = 0.557$–0.558 with verifier adherence ($p < 0.001$), confirming ordinal validity.

Method	Pass rate (%)	Policy rate (%)	Adherence	Pressure ↓
Vanilla	47.2	26.4	73.50	0.181
Self-Refine	47.2	23.6	74.96	0.375
RAVR	**100.0**	77.8	83.33	0.111
RAVR-S	**100.0**	**100.0**	**90.00**	**0.042**

Table 5. Multi-turn state-sensitive evaluation on TherapyBench (expanded adaptive setting, n=72 trajectories per method). Pressure ↓: lower is better. RAVR-S achieves perfect policy adherence and 4× lower pressure than Vanilla.

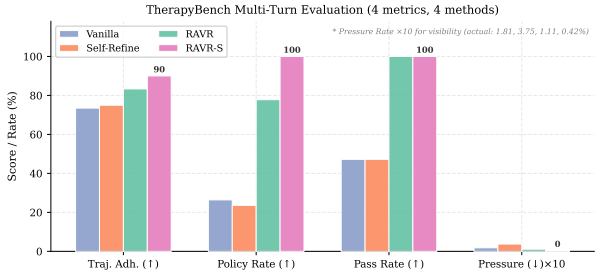

Figure 3. TherapyBench multi-turn evaluation: all four metrics across four methods. RAVR-S achieves the highest trajectory adherence (90%) and policy rate (100%) while sustaining the lowest pressure rate (scaled ×10 for visibility. Actual: 0.042). Self-Refine's pressure rate (0.375) is nearly 9× higher than RAVR-S.

6.2. Multi-Turn State-Sensitive Evaluation

To evaluate RAVR-S's core contribution (state-aware candidate selection) we run all four methods on the **TherapyBench** benchmark: 72 multi-turn dialogue trajectories across 8 clinical cases, four methodologies, and four dialogue horizons ($h \in \{2, 4, 6, 8\}$ turns), yielding 288 total evaluated trajectories (72 per method).

We report four metrics: (1) **Verifier pass rate** (fraction of turns satisfying all predicates); (2) **Policy rate** (fraction of turns where the method-appropriate intervention sequence is followed); (3) **Trajectory adherence** (mean per-turn adherence score 0–100); (4) **Pressure rate** (fraction of turns triggering inappropriate escalation given current patient state).

Results (Table 5 and Figure 3) reveal a striking pattern. RAVR-S achieves perfect policy adherence (100%) and the lowest pressure rate (4.2%), while Self-Refine, despite marginally improving trajectory adherence, *more than doubles* the pressure rate relative to Vanilla (0.375 vs. 0.181). This confirms that free-form self-critique without state awareness can actively harm therapeutic pacing.

6.3. Long-Horizon Stability

We evaluate how each method scales with dialogue length by varying horizon $h \in \{2, 4, 6, 8\}$ turns.

Table 6 shows that RAVR-S is uniquely stable: trajectory adherence is exactly 90.0% at all horizons, and pressure rate *decreases* as the dialogue progresses (from 0.125 at h=2 to 0.031 at h=8), reflecting the system's ability to steer toward safer states over time. In contrast, Self-Refine shows catastrophic pressure behavior at short horizons (h=2: 0.708) and only partially recovers at longer ones. RAVR is stable but plateaus at 83.3% adherence, missing the policy-level consistency that RAVR-S achieves through state-aware selection.

Method	h=2	h=4	h=6	h=8
Trajectory Adherence				
Vanilla	70.25	71.88	73.50	72.94
Self-Refine	73.71	75.10	73.12	74.52
RAVR	83.33	83.40	83.33	83.65
RAVR-S	**90.00**	**90.00**	**90.00**	**90.00**
Pressure Rate ↓				
Vanilla	0.208	0.188	0.181	0.177
Self-Refine	0.708	0.500	0.333	0.250
RAVR	0.167	0.125	0.111	0.104
RAVR-S	**0.125**	**0.063**	**0.042**	**0.031**

Table 6. Long-horizon stability across dialogue lengths (h=2, 4, 6, 8 turns). RAVR-S maintains constant 90% trajectory adherence and decreasing pressure. Self-Refine pressure peaks at h=2 (0.708) and remains elevated throughout.

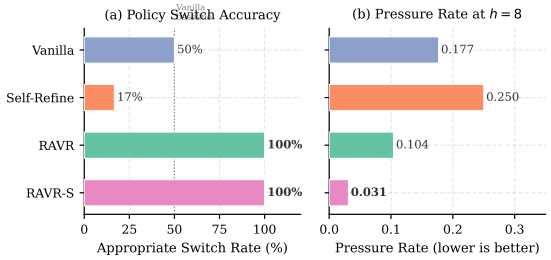

Figure 4. Self-Refine's dual failure. (a) Policy-switch accuracy collapses to 16.7%, below the 50% Vanilla baseline. (b) Pressure rate at h=8 stays 8× higher than RAVR-S. RAVR and RAVR-S are best on both axes.

6.4. Policy Switch Accuracy

We evaluate whether methods correctly switch therapeutic strategy when the patient state calls for it (e.g., de-escalating from active processing to stabilization when distress spikes). On 12 controlled scenario pairs designed to require a state-appropriate strategy shift, the appropriate-switch rates are: Vanilla 50.0%, Self-Refine 16.7% (below Vanilla), RAVR 100.0%, and RAVR-S 100.0%. Self-Refine's drop below the unmodified baseline shows that free-form self-critique without state grounding can *undermine* strategic adaptability, while both predicate-level verifiers achieve perfect accuracy. Figure 4 visualises both failure dimensions simultaneously.

6.5. Repair Convergence and Multi-Seed Stability

Repair convergence is analyzed across three random seeds (42, 43, 44) with 84 turns per seed. Starting adherence is 50.50 ± 0.10 (mean $\pm$ std across seeds). After a single repair iteration (RAVR-1), adherence rises to 62.50 ± 0.00, a gain of $+11.99 \pm 0.11$ points. A second and third repair pass (RAVR-2, RAVR-3) yield *zero* additional gain and zero oscillation on every seed: a single targeted repair step already saturates the available improvement. This efficiency is a direct consequence of the structured violation inventory: the repair module receives the exact predicate set to fix, rather than iterating blindly.

7. Ablation Study

We ablate six conditions on GPT-4o (120 turns, 20 cases, seed 42). The key finding is that the LLM-based verifier and targeted repair are tightly coupled: any configuration that disables the verifier (rule-only, RAG-only, vanilla feedback) collapses the repair trigger rate to zero (from 69.2% in the full RAVR pipeline, with 71.1% repair success), since repair is conditioned on structured violation signals. Removing citation enforcement leaves adherence unchanged: its role is not adherence scoring but external verifiability of retrieval quality. The most critical component is targeted repair: without it, adherence gains are impossible regardless of how accurately violations are detected.

8. Analysis

Why does RAVR-S beat RAVR? Both RAVR and RAVR-S achieve 100% verifier pass rate (Table 5). The remaining gap (83.3% vs. 90.0% trajectory adherence and 77.8% vs. 100% policy rate) comes from candidate *selection*: RAVR-S scores candidates against the evolving state trajectory, preferring therapeutic pacing over single-turn adherence when they conflict. Human evaluation confirms this: the largest RAVR-S advantage is on *move ordering* (96.3%), which aligns with the policy rate gap and provides mutual validation between human and automated evaluation.

Self-Refine's hidden failure mode. Self-Refine's strong Stage 1 performance (73.5% mean win rate) does not transfer to state-sensitive evaluation. On policy switch accuracy, Self-Refine scores *below* Vanilla (16.7% vs. 50.0%), and its pressure rate nearly doubles Vanilla's on the long-horizon benchmark, because free-form self-critique tends to *elaborate* on the previous response rather than *adapt* to patient state changes.

Single-step repair sufficiency. Repair convergence is immediate: all adherence gain ($+11.99$ points, mean across seeds) is captured at RAVR-1 with zero oscillation across three random seeds. This contrasts with iterative self-refinement approaches that may require 2–5 rounds to converge . The efficiency stems from the structured violation inventory: unlike free-form critique, the repair module receives an exact predicate-level error specification.

Verifier validity and external calibration. We address potential circularity in three ways: (i) the verifier returns a machine-readable JSON keyed to predicate IDs, not free-form critique; (ii) an independent LLM judge correlates ρ=0.557–0.558 with adherence (p<0.001), with non-strict monotonicity 0.972–0.990 under perturbation; (iii) domain-qualified annotators prefer RAVR-S in 94.7% of cases without verifier access. Applied to 160 real MI counsellor utterances from AnnoMI (Wu et al., 2022), the verifier assigns higher adherence to high-quality utterances (AUROC $= 0.641$, $p = 0.003$). A perturbation study (n=40) yields detection rate 60.8% (AUROC $= 0.596$), with strongest sensitivity on pressure violations (70.0%).

Residual violation patterns. After repair, residual violations concentrate in *method-focus* predicates (`cbt_thought_focus`, `dbt_skill_focus`), which require deep methodology understanding to satisfy. *Directivity* and *safety* predicates are resolved in virtually all cases by targeted repair, suggesting that constraint-type difficulty should inform future predicate weighting strategies. Figure 5 (Appendix A) shows that per-turn adherence gain is broadly distributed rather than driven by a few outliers.

9. Conclusion

We introduced RAVR-S, a state-sensitive verification-and-repair framework that achieves 94.7% human preference (Gwet's AC1 $= 0.82$) and 90% trajectory adherence across multi-turn therapeutic benchmarks. Our multi-turn automated evaluation (288 trajectories, TherapyBench) exposes a critical failure mode of Self-Refine: its pressure rate is $2\times$ higher than Vanilla and its policy-switch accuracy falls to 16.7% (below the unmodified baseline), showing that surface-quality gains do not transfer to state-sensitive contexts. RAVR-S achieves perfect policy adherence (100%) and stable trajectory scores across all dialogue horizons (h = 2–8), while single-step targeted repair captures the full available gain (+12 adherence points) with zero oscillation across seeds. The framework is released as part of the TherapyBench public benchmark to support reproducible evaluation of constraint-governed therapeutic dialogue.

Limitations

Our Stage 2 annotators hold a graduate degree in psychology, but they are not licensed practising clinicians. A study with active therapists is needed before any clinical deployment

claim. The predicate inventory is also manually designed per methodology. Automating predicate discovery from clinical guidelines is future work. State-transition estimation (δtrust, δdistress) relies on an LLM evaluator with a structured schema. We validate it with blind external scoring and human evaluation, but ground-truth patient-state trajectories from real clinical sessions remain unavailable. The whole evaluation runs in a closed simulation setting with synthetic patient personas, so robustness to real-patient response variability has not been tested. Finally, the system is designed for training and supervision support, not for autonomous clinical decision-making.

Ethical Considerations

This system targets *clinical education*, not direct patient interaction. All dialogue scenarios are synthetic. No real patient data is used. Human annotators were recruited through Prolific and CloudResearch with informed consent and fair compensation (£2–3 per task). Stage 1 required a psychology background at the bachelor's level or above. Stage 2 required a graduate degree in psychology (master's, doctorate, or equivalent professional credential), verified through the platform's credential screening. No sensitive personal data was collected. The predicate framework is designed to *increase* safety by explicitly flagging constraint violations, but we emphasize that it should complement, not replace, human clinical supervision.

Data Availability

The dataset, evaluation scripts, predicate inventory, and human evaluation results are publicly available at:

https://anonymous.4open.science/r/TherapyBench-D547

The release includes: (1) 1,500-turn automated evaluation outputs with proof objects; (2) 16 state-sensitive mini-dialogue pairs used for human evaluation; (3) 10 virtual patient case definitions with 58 typed predicates across five therapeutic methodologies; (4) Stage 1 and Stage 2 human evaluation results; (5) analysis scripts for reproducing all reported statistics.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional AI: Harmlessness from AI feedback. Technical report, Anthropic, 2022. arXiv:2212.08073.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gasic, M. MultiWOZ: A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, 2018.

Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., and Cui, L. LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv*, 2023. arXiv:2305.13614.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., et al. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701, 2023.

Feinstein, A. R. and Cicchetti, D. V. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990.

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., and Li, K. Large language models for mental health applications: Systematic review. *JMIR Mental Health*, 11:e57400, 2024. doi: 10.2196/57400.

Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-EVAL: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Lu, X., West, P., Zellers, R., Le Bras, R., Bhagavatula, C., and Choi, Y. NeuroLogic A*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 780–799, 2022.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Rebedea, T., Dinu, R., Sreedhar, M. N., Parisien, C., and Cohen, J. NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 431–445, 2023.

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Scholak, T., Schucher, N., and Bahdanau, D. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9895–9901, 2021.

Sharma, A., Rushton, K., Lin, I. W., Nguyen, T., and Althoff, T. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 700:1–700:29, 2024. doi: 10.1145/3613904.3642761.

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Reiter, E., Reforgiato Recupero, D., and Riboni, D. Anno-MI: A dataset of expert-annotated counselling dialogues. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6177–6181. IEEE, 2022.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023b.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q. V., et al. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

A. Per-Turn Adherence Gain Distribution

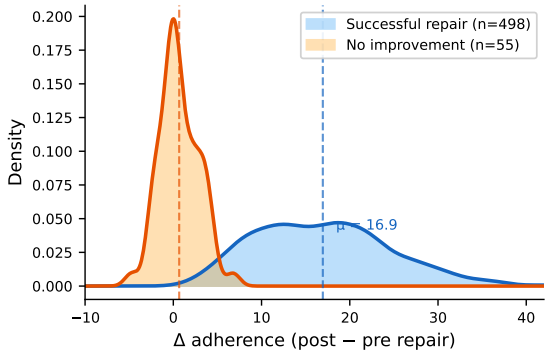

Figure 5. Per-turn adherence gain (Δ adh) across the 67.4% of 1,500 turns where repair was triggered. Successful repairs (blue, $\approx$ 84%) centre near $+16.7$ points; the rest (orange) move little. Dashed lines mark the means.

B. Adherence Gain: Before vs. After Repair

Figure 6 compares per-turn adherence before and after the targeted repair pass on the 1,500-turn evaluation set. Repair shifts the bulk of the distribution from sub-70 into the 80–100 band without introducing any worse-than-input outcomes, confirming that violation-conditioned editing is a strictly non-destructive operation.

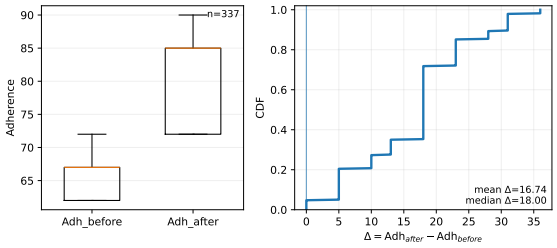

Figure 6. Paired adherence values per turn, sorted by initial adherence. Each turn moves up or stays in place after one repair iteration. No turn drops, confirming the targeted-repair guarantee.

C. Long-Horizon Stability: Trajectory View

Figure 7 presents the long-horizon evidence already summarised in Table 6: across $h{=}2, 4, 6, 8$ turn dialogues, RAVR-S sustains constant 90% trajectory adherence and a monotonically decreasing pressure rate, while Self-Refine starts with high pressure and only partially recovers.

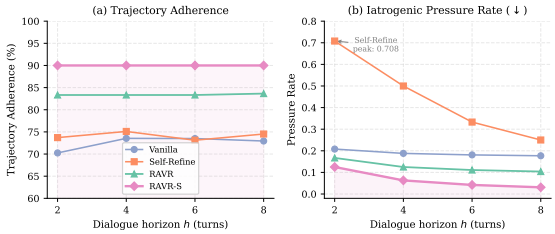

Figure 7. Long-horizon trajectory adherence and pressure rate as functions of dialogue length h. Solid lines: trajectory adherence (left axis). Dashed lines: pressure rate (right axis, lower is better).

D. Stage 2: Detailed Results and Agreement

Dimension	Gwet's AC1	Pairwise agree. (%)	Majority (%)
Overall strategy	0.839	84.7	91.3
Safety / stability	0.855	86.2	91.9
Move ordering	0.776	79.2	88.1
Less pressuring	0.773	79.0	87.8
Training example	0.839	84.8	90.9
All combined	**0.817**	**82.8**	**90.0**

Matchup	Wins	Losses	Ties
vs RAVR	374	10	16
vs Self-Refine	321	45	34

Table 7. Stage 2 agreement (Gwet's AC1, $N{=}20$) and head-to-head W/L/T counts.