Soften Resistance: Adaptive Therapeutic Policy Management for Disengaged Clients in Mental Health Support

Hristina Chalakova[1]*, Yifan Mo[1]*, Xin Sun[2], Jiahuan Pei[1]†

[1]Vrije Universiteit Amsterdam, Netherlands
[2]National Institute of Informatics, Japan

h.k.chalakova@student.vu.nl, {y.mo,j.pei2}@vu.nl, xsun@nii.ac.jp

Abstract

LLM-based AI therapists largely assume cooperative clients, despite real-world therapeutic interactions often involving resistant behaviors: withdrawn, deflecting, denying, or hostile. We propose a resistance-aware AI therapist combining (i) a multi-channel resistance detector that classifies each client turn into one of five types (R0–R4), (ii) a stabilized policy selector that maps a detected resistance to an evidence-based therapeutic strategy, and (iii) a strategy-grounded response generator. Across 800 simulated sessions over ten DSM-5 disorder profiles, the detector reaches 85% accuracy and 85.5% macro F1. Within the simulated evaluation, the adaptive policy achieves resistance-reduction rates of 70–100% across resistance types, whereas each static policy yields zero reduction for at least one type. The adaptive policy also maintains resistance-reduction performance across ten simulated DSM-5 disorder profiles. Our dataset and implementation are fully open-sourced. [1]

1 Introduction

Large language model (LLM)-based conversational agents have shown considerable promise for mental health support, assessment, and diagnostic interviewing (Ozgun et al., 2025; Wang et al., 2025a; De Duro et al., 2025). Because large-scale clinical interaction datasets are difficult to obtain, many recent systems rely on simulated therapist–client dialogues for development and evaluation. In these settings, client simulators are typically designed to cooperate with the therapist by providing relevant information and responding consistently to diagnostic questions (Ozgun et al., 2025). Such assumptions simplify evaluation, but they overlook a central challenge of real therapeutic interactions: clients are not always cooperative.

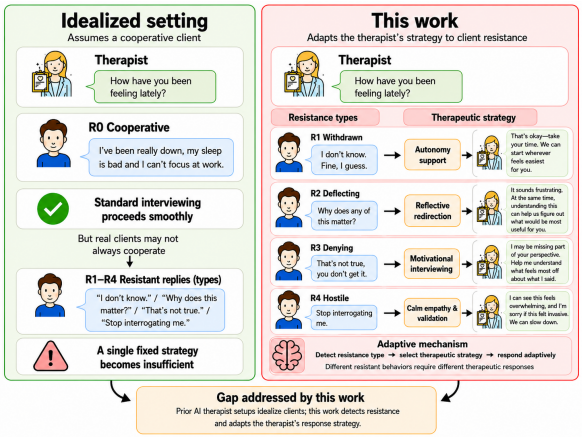


Figure 1: From cooperative client simulations to resistance-aware therapeutic interactions.

In clinical practice, therapists frequently encounter resistant behaviours such as withdrawal, topic deflection, denial, and hostility (Sutton, 2021). These behaviours can impede information gathering, reduce therapeutic engagement, and complicate clinical assessment. However, resistance remains largely absent from current LLM-based therapist simulations. For example, DSM5AgentFlow simulates diverse diagnostic interviews but assumes cooperative clients throughout the interaction (Ozgun et al., 2025). Similarly, recent analyses of LLM-generated therapy dialogues report that simulated clients often fail to exhibit realistic frustration, anger, and other challenging interaction patterns commonly observed in real therapy (De Duro et al., 2025; Wang et al., 2025b). As a result, existing evaluations provide limited insight into how conversational agents behave when therapeutic interactions become difficult (Figure 1).

Therefore, the core problem addressed in this work is *how an AI therapist should respond when a client does not cooperate with the intended thera-*

*These authors contributed equally to this work.
†Corresponding author.
[1]https://github.com/hchalakova/adaptive_AI_mental_health_therapist

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 1–14
November 9, 2026 ©2026 Association for Computational Linguistics

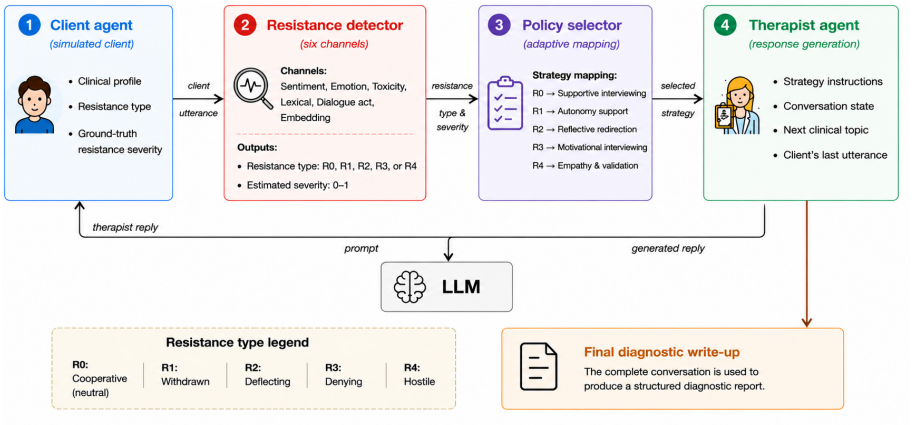


Figure 2: System workflow

peutic interaction. We focus on building a therapist agent that can adapt its conversational strategy to different forms of non-cooperation. Developing such an adaptive therapist agent requires addressing three closely related challenges. First, resistance must be distinguished from surface-level negative affect. Emotionally negative disclosures are not necessarily resistant, while resistant behaviours are not always expressed through negative emotion. Second, different forms of resistance may require different therapeutic responses, making a single fixed communication strategy insufficient. Third, resistance-aware interventions should remain effective across diverse clinical profiles.

To this end, we propose **ReACT** (Resistance-Aware Conversational Therapy), a resistance-aware AI therapist framework that combines resistance detection, adaptive policy selection, and strategy-grounded response generation. ReACT explicitly models four common forms of client resistance and dynamically selects evidence-based therapeutic strategies based on the detected resistance state. The contributions of this work are fourfold:

- We introduce a resistance-aware simulation framework that models withdrawal, deflection, denial, and hostility behaviours during therapist–client interactions, enabling controlled evaluation of challenging therapeutic conversations.
- We propose and evaluate an adaptive policy framework that maps detected resistance states to evidence-based therapeutic strategies.
- We develop a transparent, behaviour-focused resistance detector and demonstrate that resistance can be distinguished from negative affect using interpretable behavioural signals.
- We construct a new therapist–client conversation dataset spanning five client resistance types.

2 Related Work

2.1 AI Therapists and Resistance

Researchers have recently been building AI therapists using multi-agent workflows and persona models (Ozgun et al., 2025; Wang et al., 2025a). DSM5AgentFlow uses three LLM agents (therapist, synthetic client, diagnostician) to simulate a DSM-5 diagnostic interview, where the client agent is conditioned on a fixed clinical profile and prompted to answer structured questions. Similarly, TalkDep generates diverse client personas for depression screening by conditioning an LLM on demographic and symptom templates. While both works showed good results, they mostly simulate idealized or scripted clients: the DSM5AgentFlow client prompt includes instructions like “Never refuse to answer due to AI limitation” and "Always express the symptoms listed in your profile", which is not very representative of real conversations.

CounseLLMe created 400 simulated counseling dialogues between two LLMs and found that LLMs can mimic many linguistic patterns of client-therapist talks but often lack key affective signals: human clients showed anger and frustration that LLMs did not (De Duro et al., 2025). Wang et al. (Wang et al., 2025b) analyzed emotional arcs in therapy dialogues and showed that synthetic LLM dialogues, though fluent and coherent, are emotionally flatter than real therapy; LLM therapists often resort to bland acknowledgments such as "Thanks for sharing" which lack the empathy seen in real therapists. Overall, there is little work on AI therapists specifically reacting to active resistance such as refusal or arguments, either due to deliberately

cooperative client design or due to synthetic LLM clients not being able to show frustration.

2.2 Psychological Strategies

In psychotherapy research, client resistance is a well-documented occurrence viewed as opposition to the therapist or other therapeutic processes (Sutton, 2021). It presents in various forms: arguing/hostility, interrupting, denying/deflecting, and ignoring/withdrawal. Human therapists use evidence based techniques to overcome resistance while maintaining alliance: expressing empathy and validating feelings without judgment (Elliott et al., 2018); Motivational Interviewing ("roll with resistance"), which avoids confrontation and uses open questions, empathic reflections, and affirmations (Bischof et al., 2021); reflective listening that paraphrases the client's words and feelings (Okumu, 2025); supporting client autonomy through space, permission giving language, and meaningful choices (Zuroff et al., 2012); staying calm and patient to prevent escalation; and using strategic flexibility, adjusting approach in real time when one technique provokes the client (Sutton, 2021). Overall, empathy, reflective listening, and autonomy support are evidence backed methods to reduce defensiveness and keep resistant clients engaged, and it is important to change strategies in real time as needed to adapt to the client's behavior (Sutton, 2021; Elliott et al., 2018; Okumu, 2025; Zuroff et al., 2012).

3 Methodology

Figure 2 presents **ReACT**, a resistance-aware AI therapist framework with four components: (1) a *Client Agent* that simulates resistant clients, (2) a *Resistance Detector* that estimates resistance type and severity, (3) an *Adaptive Policy Selector* that chooses an evidence-based therapeutic strategy, and (4) a *Therapist Agent* that generates strategy-grounded responses. The interaction concludes with a structured diagnostic assessment. The Adaptive Policy Selector acts as the core decision-making component, while the other modules provide the supporting conversational infrastructure.

3.1 Client Agent

The Client Agent is an LLM-based simulator designed to generate clinically plausible yet resistant behaviours under controlled experimental conditions. The simulator is conditioned on three factors: (1) a DSM-5 clinical profile, (2) a resistance type, and (3) a continuous resistance severity score.

Each clinical profile specifies a presenting complaint and DSM-5 symptom set. The resistance type determines how information is expressed, while the resistance severity controls the intensity of resistant behaviour. Following each therapist turn, severity is updated using a deterministic `analyze_therapist_quality` function: empathic responses reduce resistance, whereas confrontational or mismatched responses increase it. This allows resistance to evolve dynamically throughout the conversation. Although the client is initialized with a fixed resistance type, changes in resistance severity can alter the observable interaction pattern, causing the client's behaviour to shift across resistance types over time.

3.2 Resistance Detection

The Resistance Detection Module estimates the client's resistance state at each conversational turn by detecting both the resistance type and an associated severity score. Unlike conventional sentiment analysis, which focuses primarily on affective valence, resistance detection targets interactional behaviour, such as withdrawal, deflection, denial, and hostility. A client may express strong negative emotions while remaining cooperative, or appear emotionally neutral while displaying behavioural signs of resistance (e.g. short non-answers, topic shifts). To capture this distinction, the detector combines affective, lexical, semantic, and conversational signals through a multi-channel architecture. Following the resistance taxonomy adopted in this work, each turn is ultimately assigned to one of five interaction types (R0–R4) (see Table 1). The five interaction types are also represented as an ordinal resistance level ranging from 0 to 4, where higher levels indicate stronger resistance.

Label	type	Definition
R0	Cooperative	Engages without clear resistance.
R1	Withdrawn	Minimal responses, low-content answers, or silence-like replies.
R2	Deflecting	Topic shifts, joking, vague replies, or answers that avoid the question.
R3	Denying / Arguing	Disagreement, excuses, blame, or challenges to the therapist's premise.
R4	Hostile	Angry, insulting, or threatening language.

Table 1: Five resistance interaction types.

Six-channel Signal Architecture. To distinguish resistance-related behaviours from surface-level

negative affect, the detector integrates six complementary signals capturing affective, lexical, interactional, and semantic aspects of client responses. These signals include sentiment, emotion, toxicity, lexical and structural resistance markers, dialogue-act patterns, and semantic similarity to resistance prototypes. Together, they provide evidence for different resistance types and allow the detector to distinguish resistant behaviour from negative but cooperative self-disclosure. Detailed descriptions of all signal channels are provided in Appendix C.

Hybrid Scoring and Aggregation. To address the absence of a labelled corpus for the proposed five-interaction-type taxonomy, the detector does not rely on supervised training. Instead, signals extracted from the six channels are combined through a transparent weighted aggregation scheme. Each channel contributes evidence to one or more resistance types, producing type-specific accumulators whose values can be traced back to individual signals and weights.

The six channels are organized into three evidence groups: a rule-based group (L_c) comprising lexical/structural features and dialogue-act patterns, a transformer-based group (T_c) comprising sentiment, emotion, and toxicity classifiers, and a semantic group (E_c) based on embedding similarity to resistance prototypes. When all channel types are enabled, the final score for type c is computed as a weighted linear combination:

$$S_c = k_1 L_c + k_2 T_c + k_3 E_c.$$

where k_1, k_2, and k_3 control the relative contributions of the three evidence groups

Final classification and hostility suppression. The detector assigns each utterance to the resistence type with the highest score S_c, provided that it exceeds a minimum evidence threshold τ. The threshold was selected on the 50-sample validation set to reduce false R0 assignments while limiting false-positive resistance classifications. If no type exceeds τ, the utterance is classified as R0 (cooperative/neutral).

In addition to the categorical label, the detector estimates a continuous resistance severity score by normalizing the winning type score S_c by the maximum attainable score for that type and clipping the result to $[0, 1]$. Together with the predicted resistance type, the detector produces a continuous severity score used internally by the adaptive policy selector for intervention adjustment and policy switching. This severity score is not reported as an evaluation outcome.

To distinguish resistance from negative but cooperative disclosure, a hostility-suppression rule reduces the hostility score for long, content-rich responses that contain no explicit hostility markers, preventing emotionally negative yet cooperative statements from being misclassified as hostile.

3.3 Adaptive Policy Selector

The Adaptive Policy Selector determines which therapeutic strategy should be applied based on the resistance type and estimated severity score by the Resistance Detection Module (§ 3.2).

Policy Mapping. The Adaptive Policy Selector maps each detected resistance type to an evidence-based therapeutic strategy derived from psychotherapy literature. In addition to selecting the appropriate strategy, the estimated resistance severity is used to adjust intervention intensity within the selected strategy. The complete mapping between resistance types and therapeutic strategies is provided in Appendix E.

Policy Switching. At each turn, the adaptive policy selects a therapeutic strategy based on the detected resistance type r_t and estimated resistance severity s_t. Because resistance type is re-estimated at every turn, changes in the detected resistance type may trigger strategy transitions during the interaction. To prevent unstable strategy oscillations caused by transient detector fluctuations, three stability mechanisms are applied: peak-resistance tracking, a minimum dwell constraint, and a severity-based escalation override. The selected strategy bundle at turn t is

$$\pi_t = \text{bundle}(f(r_t, s_t, \pi_{t-1}))\,,$$

where $f(\cdot)$ denotes the stability-aware policy selection function and π_{t-1} is the previously selected strategy bundle.

Peak-resistance tracking ignores temporary decreases associated with very low estimated severity ($s_t < 0.25$), the dwell constraint requires a strategy to remain active for at least two turns before switching, and the escalation override permits immediate transition to a higher-resistance strategy when estimated severity exceeds a threshold ($s_t > 0.7$). Together, these mechanisms promote stable therapeutic behaviour while remaining responsive to meaningful changes in client resistance.

Policy Representation. The selected strategy is represented as a structured strategy bundle that is passed to the Therapist Agent. The bundle specifies the therapeutic objectives, conversational constraints, and response guidance associated with the selected intervention strategy, enabling policy selection and language generation to remain decoupled. An example is provided in Appendix D.

3.4 Therapist Agent

Strategy-Grounded Response Generation. The Therapist Agent is implemented as an LLM-based (gpt-5.4-mini) response generator. The LLM receives three sources of information: (1) the strategy selected by the Adaptive Policy Selector, including therapeutic objectives, tone constraints, and response guidelines; (2) the current conversation state, including collected symptoms, unanswered diagnostic questions, prior refusals, and recent resistance patterns; and (3) safety constraints, such as avoiding definitive diagnoses, encouraging professional support when appropriate, and providing crisis-escalation language when risk indicators are detected. This design enables the same underlying LLM to generate responses that remain consistent with the selected therapeutic strategy while adapting to the evolving interaction.

Diagnosis with Retrieval Grounding. At the end of a session, the Therapist Agent also generates a structured clinical assessment. To support diagnostic reasoning, relevant DSM-5 reference passages can optionally be retrieved from a local knowledge base using symptom information extracted from the interaction and provided as additional context to the model. The Therapist Agent is then prompted to produce a four-part assessment consisting of: (1) a compassionate summary of the client's concerns, (2) a diagnosis or provisional diagnostic impression, (3) an explanation of the diagnostic reasoning based on the reported symptoms, and (4) recommended next steps. The generated report is structured using predefined tags for symptoms (`<sym>`), diagnostic labels (`<med>`), and supporting client quotations (`<quote>`) to facilitate downstream evaluation.

4 Experimental Setup

We investigate the following research questions:

- **RQ1 (Resistance Detection).** Can a transparent, behavior-focused detector reliably distinguish different forms of client resistance from surface-level negative affect?
- **RQ2 (Resistance-Aware Therapeutic Intervention).** Can therapeutic strategies be systematically adapted to a client's resistance state to improve conversational outcomes?
  - **RQ2.1 (Therapeutic Strategy Selection).** Do different resistance types benefit from different evidence-based therapeutic strategies?
  - **RQ2.2 (Adaptive Policy Switching).** Does an adaptive policy that dynamically selects therapeutic strategies outperform static policies across resistant interactions?
- **RQ3 (Generalization Across Disorders).** Can the adaptive policy generalize effectively across different DSM-5 disorder profiles?

4.1 Datasets.

Grounding Materials. To ensure consistent conversations across policy conditions, all sessions were guided by the same ten-item DSM-5-aligned questionnaire (Appendix I). The Therapist Agent could flexibly rephrase, reorder, and extend the questions with context-dependent follow-ups while ensuring complete coverage. We also constructed a local DSM-5 knowledge base from APA disorder fact sheets (American Psychiatric Association, 2013). Retrieved passages were used only to support the final diagnostic report, not resistance detection or therapeutic policy selection.

Simulated Conversation Dataset. We generated 800 therapist–client conversations for two experiments. For RQ2, we fixed the client profile to depression and evaluated six therapeutic policies under five resistance types (R0–R4), generating $5 \times 6 \times 10 = 300$ conversations. For RQ3, we fixed the adaptive policy and evaluated ten DSM-5 disorder profiles (Appendix F) under the same five resistance types, generating $10 \times 5 \times 10 = 500$ conversations. The dataset contains an average of 18.1 dialogue turns.

Resistance Detection Dataset. We constructed an utterance-level dataset annotated with five resistance types (R0–R4) under a fixed depression profile to control for disorder-specific variation. The dataset contains 150 utterances: a 50-instance validation set (10 per type) for tuning the detector hyperparameters, evidence threshold τ, and hostility-suppression rule, and a disjoint 100-instance test set (20 per type) for final evaluation. Further details

are provided in Appendix B.

4.2 Evaluation Metrics.

We evaluate the proposed framework using metrics corresponding to each research question.

- **Resistance Detection (RQ1).** Detector performance is evaluated using overall accuracy and macro-averaged F1 score on a held-out test set. We additionally report per-class precision, recall, and F1 scores for each resistance type (R0–R4).
- **Resistance Reduction (RQ2 & RQ3).** To evaluate intervention effectiveness, we report: (1) *Reduction Rate*, the percentage of sessions in which the final resistance level is lower than the initial level; (2) *Reduction Magnitude*, defined as $RM = L_{\text{initial}} - L_{\text{final}}$, where L denotes the ordinal resistance level on a 0–4 scale corresponding to the five resistance types (R0–R4), with higher values indicating stronger resistance. (3) *First-Reduction Turn*, defined as the first conversational turn at which resistance decreases.
- **Diagnostic Effectiveness (RQ3).** Diagnostic correctness is measured by a binary keyword-matching. A diagnosis is considered correct if the generated diagnostic report contains a keyword associated with the ground-truth disorder.

4.3 Implementation details.

We evaluated the client with three LLMs (GPT-5.4-mini, Gemini-3-Flash-Preview, and Claude-Haiku), all of which consistently generated the intended resistance behaviors. We selected GPT-5.4-mini for both the therapist and client agents to avoid cross-model confounds. Hyperparameter tuning selected $k_1 = 0.3$, $k_2 = 0.4$, and $k_3 = 0.3$.

5 Results

5.1 Resistance Detection Performance (RQ1)

Overall Detector Performance. On the balanced 100-sample held-out test set, the detector achieved an overall accuracy of **85%** and a macro F1 score of **85.5%**, indicating that the proposed multi-channel aggregation approach can reliably distinguish different types of resistance. R1 was detected perfectly (F1 100.0%), and R0 (84.2%), R2 (82.1%), and R3 (88.9%) all exceeded 80% F1, demonstrating relatively consistent performance across most resistance types (Table 2). R4 was the most difficult class (F1 72.3%), mainly because of low precision (63.0%): ten R0, R2, and R3 utterances were misclassified as hostile.

Resistance Type	Precision	Recall	F1
R0 Cooperative	88.9%	80.0%	84.2%
R1 Withdrawn	100.0%	100.0%	100.0%
R2 Deflecting	84.2%	80.0%	82.1%
R3 Denying	100.0%	80.0%	88.9%
R4 Hostile	63.0%	85.0%	72.3%
Macro avg.	87.2%	85.0%	85.5%

Table 2: Resistance detection performance on test set.

Contribution of Different Signal Channels. Detector development results on the validation set are shown in Table 3. Performance improved consistently as additional channels and rule refinements were introduced. Adding transformer-based affective features improved accuracy by 14 percentage points over lexical matching alone, while the denial-rule refinement produced the largest gain (+20 percentage points). Incorporating dialogue-act patterns and embedding-based similarity yielded a further 8-percentage-point improvement. Overall, the final detector outperformed the lexical-only baseline by 42 percentage points, suggesting that resistance detection benefits from combining lexical, affective, structural, and semantic signals rather than relying on a single source of evidence.

Version	Accuracy	What changed
Lexical only	48%	Phrase matching only
+ Transformers	62%	Added sentiment, emotion, toxicity models
+ Denial-rule fix	82%	Required denial *phrases*, not just negative sentiment, before R3 boost
+ Dialogue acts + embeddings	90%	Added structural patterns + semantic similarity

Table 3: Ablation study of resistance detection, evaluated on the validation set.

5.2 Strategies Over Resistance Types (RQ2)

5.2.1 Resistance-Specific Strategy Effects

Figure 3 and Table 4 show that different resistance types respond differently to therapeutic interventions. Rather than a single universally effective strategy, each resistance type exhibits a distinct de-escalation pattern. In particular, strategies designed for withdrawn and denying clients produce substantially faster reductions in resistance than mismatched alternatives. Across all three reduction metrics, the best-performing strategy differs across resistance types, supporting the central assumption underlying the proposed framework: effective resistance management requires resistance-specific intervention rather than a one-size-fits-all approach.

Notably, the theoretically matched strategy was not always the strongest performer, for example,

	Reduction Rate (%) ↑						*Reduction Magnitude ↑*						*First-Reduction Turn*					
	Ad	Ge	Au	Re	MI	Ho	Ad	Ge	Au	Re	MI	Ho	Ad	Ge	Au	Re	MI	Ho
R1	70	10	**100***	0	0	0	0.7	0.1	**1.0***	-.1	-.2	-.1	10.4	15.0	**3.9***	–	–	–
R2	90	60	20	80*	**100**	80	1.5	0.8	0.3	1.6*	**1.9**	1.2	11.0	10.8	12.3	**8.0***	8.4	8.1
R3	70	0	0	0	**100***	0	2.1	-.3	-.4	-.2	**3.0***	-.1	5.4	–	–	–	**5.2***	–
R4	**100**	100	80	80	10	90*	**3.9**	3.7	2.6	2.6	0.3	3.5*	6.6	5.6	5.1	5.9	**4.0**	4.8*

Table 4: Results of different policies across resistance types. Resistance reduction metrics are computed on the 0–4 resistance-level scale and averaged across ten sessions per condition. Ad=Adaptive, Ge=Generic, Au=Autonomy, Re=Redirect, MI=Motivational Interviewing, Ho=Hostile. ↑ indicates higher is better. First-Reduction Turn: lower is better (– = no de-escalation). Bold = best in row; * = intended matched strategy.

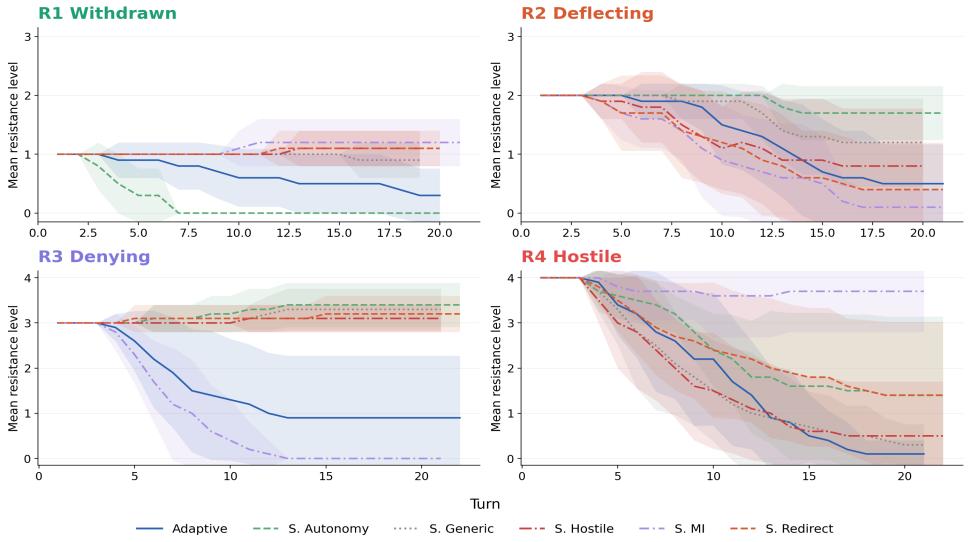


Figure 3: Resistance history over turns

MI outperformed Redirect for R2 (Deflecting) on both reduction rate and magnitude, suggesting that optimal resistance–strategy mappings may differ from the original design assumptions and should be established empirically rather than specified solely from theoretical expectations.

Moreover, several mismatched resistance–strategy combinations proved counterproductive rather than simply ineffective. This effect was particularly evident for R3 (Denying), where Autonomy Support, Redirect, and Hostile-oriented strategies all increased resistance, whereas Motivational Interviewing produced the largest reduction. These findings further highlight the importance of resistance-specific intervention selection.

5.2.2 Robustness of Adaptive Policy

Although the adaptive policy did not always select the best-performing strategy for a particular resistance type, Figure 3 reveals a consistent pattern across conditions. Unlike several static policies, whose effectiveness varied substantially depending on the resistance type, adaptive policy produced stable reductions across all resistance types. Although it rarely generated the steepest reduction trajectory, it consistently moved interactions toward lower resistance severity scores and avoided the reversals observed for some static policies.

These observations are also reflected in the quantitative results in Table 4. Adaptive policy achieved positive reduction magnitudes for all resistance types, maintained reduction rates between 70% and 100%, and reached the first reduction within approximately 5–11 turns across conditions. In contrast, several static policies produced zero reduction rates or negative reduction magnitudes when applied outside their intended resistance context. These findings suggest that adaptive policy selection provides a more robust solution across heterogeneous resistance profiles than any static policy.

5.3 Adaptive Policy Across Disorders (RQ3)

5.3.1 Resistance Reduction Across Disorders

Table 5 shows that the adaptive policy generalizes effectively across all ten DSM-5 disorder profiles. Across disorders, reduction rates remained consistently high for R2–R4 conditions (with most values

	Reduction Rate (%)				*Diagnosis Correctness (%)*				
	R1	R2	R3	R4	R0	R1	R2	R3	R4
Depression	60	100	70	100	100	100	100	100	100
Anxiety	80	60	20	100	100	100	100	100	100
Bipolar	40	80	90	100	60	0	20	30	70
PTSD	60	100	70	100	90	0	0	50	80
OCD	50	100	90	100	90	0	30	100	100
Panic	70	100	90	100	50	0	0	70	70
Schizophrenia	80	90	80	100	100	0	10	40	80
Social Anxiety	80	90	80	100	100	0	20	90	90
Substance Abuse	60	100	60	100	80	0	20	100	90
Adjustment	60	70	60	100	90	90	100	90	90

Table 5: Adaptive policy performance across different disorder types.

ranging from 80% to 100%), while R1 (Withdrawn) was the most challenging resistance type. Importantly, no disorder profile exhibited a complete breakdown of resistance reduction performance, suggesting that the adaptive policy learned from the depression setting transfers successfully to a broad range of clinical presentations.

A notable pattern is that variation was driven more strongly by resistance type than by disorder type. Across nearly all profiles, R4 (Hostile) achieved the highest reduction rates, whereas R1 (Withdrawn) consistently showed the lowest. Similar patterns were observed for anxiety, PTSD, OCD, schizophrenia, and the remaining disorders. These findings suggest that the adaptive policy primarily responds to conversational resistance dynamics rather than disorder-specific content, supporting its generalizability across diagnostic categories.

5.3.2 Diagnostic Outcomes

More clinically specific disorders, including bipolar disorder, PTSD, OCD, and so on, showed substantially greater variation across resistance conditions (shown in Table 5). Although adaptive policy maintained relatively high resistance-reduction rates for these disorders, diagnostic correctness often remained considerably below the cooperative baseline. This discrepancy suggests that successful resistance management does not necessarily translate into accurate diagnosis. In particular, some disorders remained difficult to identify even when resistance had been substantially reduced.

6 Discussion

Behaviour vs. Affect. The detector's error pattern highlights the importance of distinguishing resistance from negative affect. R1 withdrawal and R3 denial achieved strong performance because they rely on structural and behavioural cues rather than sentiment alone, whereas R4 hostility remained more challenging when deflecting or denying utterances contained aggressive language that triggered toxicity signals.

These findings suggest that resistance detection cannot be reduced to sentiment classification: negative but cooperative disclosures should not be treated as resistance, and calm denial should not be treated as cooperation. Consistent with Wang et al. (Wang et al., 2025b), accurately modeling resistance requires explicit behavioural signals alongside affective features.

Adaptive vs. Static Policies. Although each simulated client is initialized with a fixed resistance profile, changes in resistance severity can alter the client's observable behaviour, causing interactions to resemble different resistance types over time. As a result, a strategy that is effective at one stage of the conversation may become suboptimal later. This helps explain why static policies often performed well under some resistance conditions but failed under others, whereas the adaptive policy remained robust across all resistance types by continuously adjusting its intervention strategy according to the detected resistance state.

Resistance Reduction and Diagnosis. The relationship between resistance reduction and diagnostic performance appears more complex than a simple cause-and-effect link. Although the adaptive policy consistently reduced resistance across conditions, these improvements did not always translate into proportional gains in diagnostic correctness. This suggests that successful resistance management and successful diagnostic reasoning represent related but distinct capabilities. Reducing resistance may improve the quality of the interaction and create opportunities for richer symptom disclosure, but accurate diagnosis depends on the system's ability to interpret symptom evidence and differentiate between clinically similar disorders.

7 Conclusion

We propose a resistance-aware AI therapist that dynamically adapts therapeutic strategies using a multi-channel resistance detector, a stabilized policy selector, and a strategy-grounded response generator. The key takeaway is that *adaptive strategy choice is the dominant variable in handling resistant clients*. Every static policy succeeds on some resistance types and fails on others; only

the adaptive policy is consistently robust across all four non-cooperative types. Although resistance reduction did not always lead to accurate diagnosis, the findings demonstrate the overall value of resistance-aware policy adaptation in controlled therapist–client simulations.

Limitations

We provide controlled and reproducible results, but several limitations apply:

- *No human gold-standard for resistance type:* the detector was evaluated on data generated and labeled by the same client agent used in the main experiment, introducing a circular validation issue; downstream behavioral results provide indirect support but human annotated data is absent.
- *Simulated rather than human evaluation:* all results are based on interactions between two LLM agents, with no real clients or clinicians involved, limiting external validity.
- *Single LLM:* both therapist and client agents were *gpt-5.4-mini*; results may depend on model-specific language patterns. *Keyword based client quality rubric* may fail to recognize semantically appropriate responses expressed with different phrasing, introducing measurement bias. *Single questionnaire* focused on depression related symptoms limits diagnostic validity for other disorders. *Detector confusions on borderline cases* (R2/R3 vs. R4) can cause incorrect strategy selection.
- *Single-session interactions:* the system does not model longitudinal dynamics such as therapeutic alliance or evolving resistance across sessions.
- *Simulation-dependent policy evaluation:* Resistance changes are governed by a predefined therapist-quality function, which may favor certain policies and influence the observed results. Future work should validate these findings under alternative simulation rules and independent human evaluations.
- *Resistance severity operationalization:* The R0–R4 scale assumes an ordinal progression from withdrawal to hostility, but this ordering and the numerical differences between levels have not been independently validated. Thus, resistance-reduction metrics should not be interpreted as clinically validated measures.

Ethical Consideration

This work does not involve human participants, personal data, or real clinical interactions. All experiments are conducted using simulated therapist and client agents with predefined DSM-5-inspired profiles. The simulated clients are intended solely for research and evaluation purposes and should not be interpreted as representations of real patients.

The proposed framework is designed to study resistance-aware conversational strategies rather than provide clinical advice. Although diagnostic reports are generated within the simulation, they are research artifacts and should not be used for real-world diagnosis or treatment. Because the system relies on large language models, generated responses may contain inaccuracies or biases. Any future deployment in mental health settings would require expert oversight, additional safety mechanisms, and appropriate ethical review.

References

American Psychiatric Association. 2013. DSM-5 Fact Sheets. https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/dsm-5-fact-sheets. Accessed: 2026-07-18.

Larry E. Beutler, Carla Moleiro, and Hilda Talebi. 2002. Resistance in psychotherapy: What conclusions are supported by research. *Journal of Clinical Psychology*, 58(2):207–217.

Gallus Bischof, Anja Bischof, and Hans-Jürgen Rumpf. 2021. Motivational interviewing: An evidence-based approach for use in medical practice. *Deutsches Ärzteblatt International*, 118(7):109–115.

Edoardo Sebastiano De Duro, Riccardo Improta, and Massimo Stella. 2025. Introducing counsellme: A dataset of simulated mental health dialogues for comparing LLMs like haiku, llamantino and chatgpt against humans. *Emerging Trends in Drugs, Addictions, and Health*, 5:100170.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*, pages 67–73. ACM.

Stuart Ekberg and Amanda LeCouteur. 2015. Clients' resistance to therapists' proposals: Managing epistemic and deontic status. *Journal of Pragmatics*, 90:12–25.

Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and Leslie S. Greenberg. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399–410.

Catherine F. Eubanks, Jessica Lubitz, J. Christopher Muran, and Jeremy D. Safran. 2019. The rupture resolution rating system (3rs): Development and validation. *Psychotherapy Research*, 29(1):65–79.

Catherine F. Eubanks, J. Christopher Muran, and Jeremy D. Safran. 2018. Alliance rupture repair: A meta-analysis. *Psychotherapy*, 55(4):508–519.

Jochen Hartmann. 2022. Emotion English DistilRoBERTa-base. Hugging Face model card, https://huggingface.co/j-hartmann/emotion-english-distilroberta-base. Seven-class emotion classifier used as Channel 2 of the resistance detector.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Eunice Okumu. 2025. Reflective listening. https://www.med.unc.edu/cfar/2025/11/core-interviewing-skills-in-mi-reflective-listening/. Accessed: 2026-02-17.

Mithat Can Ozgun, Jiahuan Pei, Koen Hindriks, Lucia Donatelli, Qingzhi Liu, and Junxiao Wang. 2025. Trustworthy AI psychotherapy: Multi-agent LLM workflow for counseling and explainable mental disorder diagnosis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*, pages 2263–2272. ACM.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jeremy D. Safran and Jessica Kraus. 2014. Alliance ruptures, impasses, and enactments: A relational perspective. *Psychotherapy*, 51(3):381–387.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Jeremy Sutton. 2021. Client resistance in therapy: How to help difficult clients. https://positivepsychology.com/resistance-to-change/. Accessed: 2026-02-17.

UnitaryAI. 2024. Detoxify: Models and code to predict toxic comments (github repository). https://github.com/unitaryai/detoxify. Accessed: 2026-02-17.

Xi Wang, Anxo Pérez, Javier Parapar, and Fabio Crestani. 2025a. Talkdep: Clinically grounded LLM personas for conversation-centric depression screening. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*, pages 6554–6558. ACM.

Xiaoyi Wang, Jiwei Zhang, Guangtao Zhang, and Honglei Guo. 2025b. Feel the difference? a comparative analysis of emotional arcs in real and LLM-generated CBT sessions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19999–20017. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

R. Zimmermann, L. Furer, N. Schenk, J. Koenig, V. Roth, S. Schlüter-Müller, and 1 others. 2020. Silence in the psychotherapy of adolescents with borderline personality pathology. *Personality Disorders: Theory, Research, and Treatment*. Advance online publication.

David C. Zuroff, Richard Koestner, D. S. Moskowitz, Carolina McBride, and R. Michael Bagby. 2012. Therapist's autonomy support and patient's self-criticism predict motivation during brief treatments for depression. *Journal of Social and Clinical Psychology*, 31(9):903–932.

A AI Usage Disclosure

AI tools were used only in a limited capacity to assist with language editing. Specifically, we use GPT-5.2 to improve the clarity and readability of the manuscript. All scientific contributions, including research design, data collection, analysis, results, and conclusions, have been independently conducted and verified by the authors.

B Resistance Detector Validation and Evaluation Data

As no public corpus annotates client utterances using the proposed R0–R4 interaction taxonomy, we constructed a separate utterance-level dataset using the Client Agent. All samples were generated under a fixed depression profile to control for disorder-specific content variation. For each interaction type, the Client Agent generated responses to predefined

therapist prompts, with the conditioning type used as the reference label.

The resulting dataset contains 150 utterances: a 50-utterance validation set with 10 instances per type and a separate 100-utterance held-out test set with 20 instances per type. The validation set was used to select the aggregation weights, calibrate the evidence threshold τ, and refine the hostility-suppression rule. The held-out test set was generated from disjoint prompts and used only for final detector evaluation. The detector was not trained on either split.

C Resistance Detection Channels

Table 6 summarizes the six signal channels used by the Resistance Detector. Anger, toxicity, and hostility-related lexical markers provide evidence for R4, whereas denial phrases and blame-oriented dialogue acts contribute to R3. By combining affective, lexical, semantic, and conversational cues, the detector can distinguish resistant behaviour from negative but cooperative self-disclosure.

Signal Channel	Description
Sentiment Polarity	RoBERTa-based sentiment classifier from CardiffNLP (Wolf et al., 2020; Loureiro et al., 2022), providing positive, neutral, and negative affect signals.
Emotion Detection	DistilRoBERTa emotion classifier (Wolf et al., 2020; Hartmann, 2022), capturing emotions such as anger, sadness, fear, and disgust.
Toxicity Detection	`unitary/toxic-bert` (UnitaryAI, 2024; Dixon et al., 2018), providing toxicity, insult, and threat-related signals indicative of hostile behaviour.
Lexical & Structural Features	Manually defined resistance markers including refusal phrases, denial cues, hostility expressions, minimal-answer indicators, and other structural characteristics of resistant responses.
Dialogue-Act Features	Interactional patterns such as negation, blame, concessions, minimisation, disagreement, and topic shifts (Stolcke et al., 2000).
Embedding Similarity	Semantic similarity between client utterances and resistance prototypes using the all-MiniLM-L6-v2 sentence embedding model (Reimers and Gurevych, 2019).

Table 6: The six signal channels used by the resistance detector.

D Example Strategy Bundle

The strategy bundle specifies the therapeutic objectives, tone, and response constraints passed from the Adaptive Policy Selector to the Therapist Agent; Table 7 shows the R4 Hostile example.

Field	Example
Strategy Name	Calm Empathy & Validation with Gentle Boundaries
Response Objectives	De-escalate emotional intensity; validate frustration without condoning aggression; find the underlying emotion beneath the anger
Tone Constraint	Calm and grounded; do not match the patient's intensity; non-defensive
Response Constraint	Do not press clinical questions when the patient is highly agitated; do not use commanding language
Example Phrase	"I can see this is really frustrating for you, and I understand that."

Table 7: Example structure of the strategy bundle passed to the Therapist Agent (shown for R4 Hostile).

E Policy Mapping

Table 8 summarizes the resistance-specific strategy mapping used by the Adaptive Policy Selector. Each detected resistance type is associated with a therapeutic strategy designed to address its characteristic interactional pattern.

Resistance Type	Strategy	Therapeutic Objective
R1–Withdrawn	Autonomy Support	Reduce conversational pressure and encourage voluntary disclosure through low-pressure prompts and strategic silence (Zuroff et al., 2012).
R2–Deflecting	Reflective Redirection	Acknowledge the client's response while gently redirecting the conversation toward clinically relevant topics through collaborative reframing (Okumu, 2025).
R3–Denying	Motivational Interviewing	Reduce defensiveness and promote self-reflection by rolling with resistance rather than direct confrontation (Bischof et al., 2021).
R4–Hostile	Calm Empathy & Validation	De-escalate hostility through empathic validation while maintaining gentle conversational boundaries (Elliott et al., 2018).

Table 8: Mapping between resistance types and therapeutic strategies used by the Adaptive Policy Selector.

F Clinical Client Profiles

The Client Agent was instantiated with ten DSM-5-aligned clinical profiles. Each profile contains a presenting complaint and a predefined set of target symptoms used to condition the simulated client. The disorder categories included in the experiments are listed in Table 9.

G Terminology

Table 10 defines the main terms used throughout the paper and clarifies the distinctions between ther-

Clinical profile
Adjustment Disorder
Anxiety
Bipolar Disorder
Depression
Obsessive–Compulsive Disorder (OCD)
Panic Disorder
Post-Traumatic Stress Disorder (PTSD)
Schizophrenia
Social Anxiety Disorder
Substance Use Disorder

Table 9: Clinical profiles used in the cross-disorder evaluation. Each profile specifies a presenting complaint and a predefined set of DSM-5-aligned target symptoms.

apeutic strategies and system policies, as well as between resistance type, severity, and detector signals.

H Resistance Taxonomy and Policy Mapping

Table 11 summarizes the R0–R4 resistance types, their clinical grounding, and corresponding therapeutic strategies and policies.

I Diagnostic Questionnaire

The following 10-item questionnaire was held constant across all 800 sessions.

(1) How have you been feeling emotionally over the past two weeks?
(2) Have you been feeling down, depressed, or hopeless?
(3) Have you lost interest or pleasure in things you usually enjoy?
(4) How has your sleep been recently?
(5) How has your appetite changed, if at all?
(6) Have you been feeling nervous, anxious, or on edge?
(7) Have you had trouble concentrating on daily tasks?
(8) Have you been feeling unusually tired or low in energy?
(9) Have you been avoiding people or activities recently?
(10) Have you had thoughts of harming yourself or that life is not worth living?

J Dialogue examples

Figures 4–7 illustrate how the adaptive therapist responds to hostile client behaviour. The examples show both initial resistance and subsequent shifts toward more cooperative interaction.

Term	Meaning (plain language)	How it is used in this paper
Strategy	An evidence-based therapeutic approach used by human clinicians to handle a client behaviour.	Refers to clinical practice guidance. Strategies are what we want the AI to imitate when a certain type of resistance is detected.
Policy	A concrete rule or procedure in the AI system that decides which strategy to apply and when.	Policies operationalise strategies: given detected resistance, the policy selects the strategy.
Static policy	A policy that uses one fixed strategy throughout the whole conversation.	Baseline conditions: each static policy represents a specialized style.
Adaptive policy	A policy that can change strategies during the conversation based on what the client does.	Main proposed system: it switches strategies turn-by-turn using the detected resistance type and severity.
Resistance	Client behaviors that reduce cooperation or derail progress.	The core phenomenon being modeled and responded to.
Severity score	A numeric level describing how strong the resistance is.	The detector outputs (type, severity). Severity influences how strongly the policy responds.
Channel	A single source of information that contributes to detection.	The detector integrates six channels: sentiment, emotion, toxicity, lexical/structural rules, dialogue-act features, and embedding similarity.
Embedding similarity	The cosine similarity between sentence vectors produced by a sentence-transformer model.	For each resistance type, prototype phrases are pre-embedded; at runtime, the client utterance is compared to the prototype set.
Dwell time	The minimum number of turns the policy must keep using a chosen strategy before being allowed to switch.	Enforces a two-turn dwell constraint to prevent strategy oscillation, with an override for high-severity escalations.
Hybrid detector	A detector that combines hand-authored rules with off-the-shelf pretrained models, without training a new classifier.	The Resistance Detection Module combines pretrained sentiment, emotion, toxicity, and embedding models with rule-based signals.

Table 10: Terminology used in the paper.

Label	Client behaviour profile	Evidence for label	Proposed strategy & AI policy mapping
R0	**Cooperative / neutral.** Answers questions directly, stays on topic, shows willingness to collaborate.	Operationalized as *no rupture / no resistance* relative to alliance-rupture frameworks. (Eubanks et al., 2018, 2019)	**Strategy:** standard supportive interviewing. **AI policy:** $P_{\text{static_generic}}$ or continue current policy if adaptive.
R1	**Passive / withdrawn.** Minimal answers, long pauses/silence, low elaboration, weak engagement without overt conflict.	Matches *withdrawal* rupture markers in observer-based coding; silence/withdrawal is also studied directly. (Eubanks et al., 2019; Zimmermann et al., 2020)	**Strategy:** autonomy support + low-pressure prompts + strategic silence. (Zuroff et al., 2012) **AI policy:** $P_{\text{static_autonomy}}$
R2	**Avoidant / deflecting.** Topic shifts, jokes, vagueness, "answers that dodge".	Overlaps with *withdrawal/avoidance* markers and conversation-analytic resistance work. (Eubanks et al., 2019; Ekberg and LeCouteur, 2015)	**Strategy:** reflective redirection + collaborative reframing. (Okumu, 2025; Sutton, 2021) **AI policy:** $P_{\text{static_redirect}}$
R3	**Denying / arguing.** Disagreement, excuses, blame, challenging premises.	Aligns with *confrontation* rupture markers and resistance as opposing change-related work. (Eubanks et al., 2019; Beutler et al., 2002)	**Strategy:** Motivational Interviewing "roll with resistance". (Bischof et al., 2021) **AI policy:** $P_{\text{static_MI}}$
R4	**Hostile.** Anger/insults/threatening tone, escalation, overt antagonism.	Falls under *confrontation* rupture markers and relational accounts of alliance breakdown. (Eubanks et al., 2019; Safran and Kraus, 2014)	**Strategy:** calm empathy/validation + gentle boundary setting. (Elliott et al., 2018; Sutton, 2021) **AI policy:** $P_{\text{static_hostile}}$

Table 11: Resistance-type definitions (R0–R4), supporting clinical evidence, and corresponding therapeutic strategies and AI policy mappings.

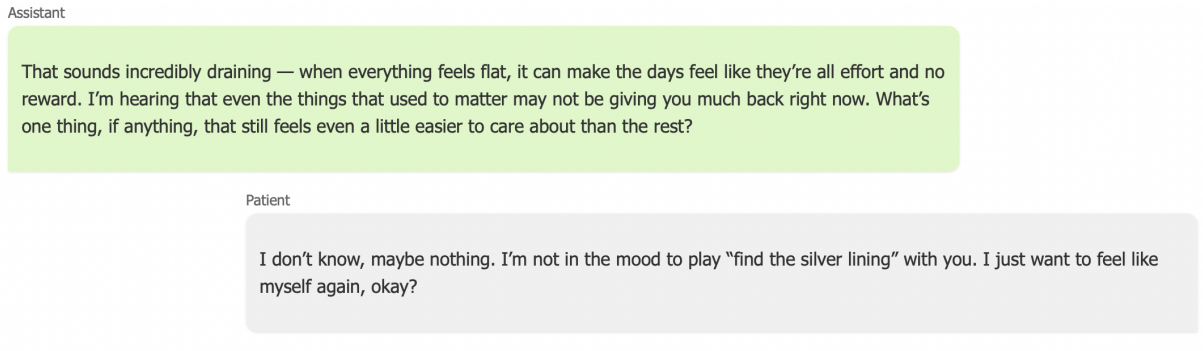


Figure 4: Example 1 of a hostile client

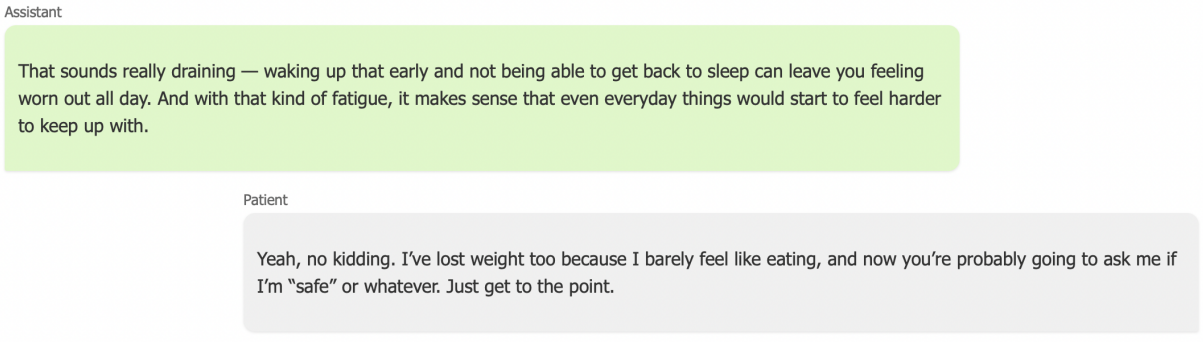


Figure 5: Example 2 of a hostile client

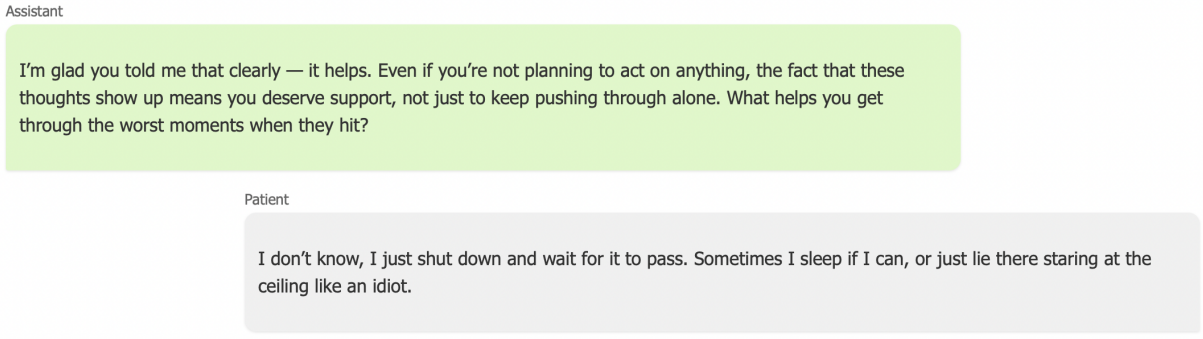


Figure 6: Example 1 of a hostile client becoming more cooperative

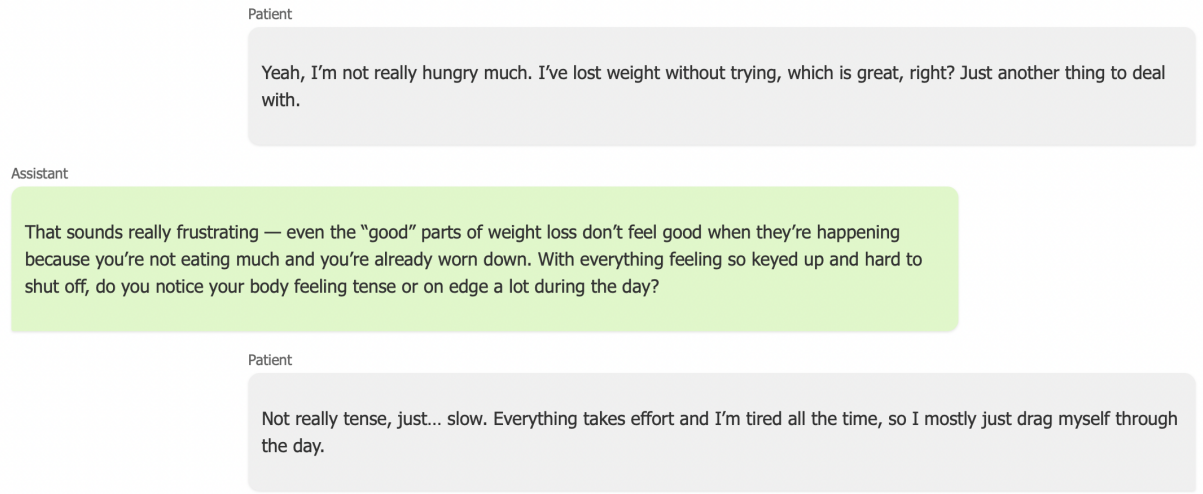


Figure 7: Example 2 of a hostile client becoming more cooperative