BEYOND COMPLIANCE: RESISTANCE-INFORMED MOTIVATION ALIGNMENT FOR CHALLENGING PSYCHOLOGICAL CLIENT SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used to simulate clients for scalable counselor training and evaluation. However, existing simulators often inherit the compliance bias of instruction-tuned LLMs, producing clients who are unrealistically open, receptive, and emotionally stable. We introduce ResistClient, a resistance-informed client simulator that models both how client resistance is expressed and why it emerges. Drawing on Client Resistance Theory, we construct the Resistance-informed Psychological Conversation (RPC) dataset by augmenting real counseling dialogues with diverse, context-dependent resistance patterns. Fine-tuning on RPC enables ResistClient to generate challenging behaviors beyond generic compliance. To ground these behaviors in coherent psychological processes, ResistClient represents client profiles using the clinical 5P case formulation, explicitly infers the client's current motivation from the profile and dialogue context, and generates responses conditioned on this motivation. Expert-informed process rewards further align motivational reasoning and response generation with human preferences while promoting decision–response consistency. Extensive automatic and expert evaluations show that ResistClient substantially outperforms existing simulators in challenge fidelity, behavioral plausibility, and reasoning coherence. Moreover, evaluations with ResistClient reveal substantial limitations in existing psychological LLMs, including frequent client resistance, therapeutic drift, and limited progress. These findings demonstrate that resistance-aware simulation provides a more realistic and demanding environment for evaluating and improving psychological LLMs.

1 INTRODUCTION

Mental health disorders affect over one billion people worldwide, yet access to qualified counselors remains severely limited (WHO, 2025). Large language models (LLMs) have increasingly been explored for counselor training (Steenstra et al., 2025), evaluation (Xiao et al., 2025), and supervision (Lin et al., 2023), where client simulators play a central role by providing scalable, repeatable practice environments (Wang et al., 2024). However, existing psychological client simulators often exhibit over-compliant behaviors compared to real clients, manifesting as unusually high openness (Kim et al., 2025a), excessive receptivity (Yang et al., 2025b), and positive emotional stability (Wang et al., 2025b). This behavioral misalignment leads to the under-exploration of challenging therapeutic scenarios, ultimately constraining the effectiveness and reliability of both counselor trainees and psychological LLMs in realistic clinical contexts.

A major challenge in real counseling is *client resistance*: opposition to therapeutic change that may manifest as self-justification, blame redirection, topic avoidance, or emotional pushback (Otani, 1989). Resistance is common across therapeutic settings, and how counselors manage it can substantially influence therapeutic processes and outcomes (Strean, 1985). Yet current LLM-based simulators rarely capture it faithfully. Because they typically generate responses directly from client profiles using pre-aligned LLMs (Wang et al., 2024), their inherent compliance bias conflicts with the non-compliant behavior observed in practice (Wang et al., 2025a). Existing prompt-based attempts to make clients more challenging (Yang et al., 2025b; Wang et al., 2025b; Kim et al., 2025a) primarily modify surface behavior without modeling the cognitive and affective mechanisms from

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

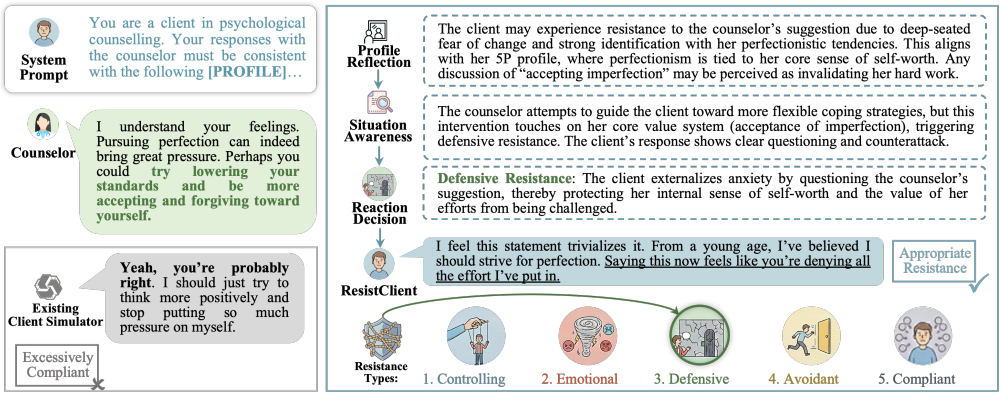


Figure 1: Existing LLM-based client simulators exhibit over-compliance. ResistClient generates authentic resistant behaviors through motivation reasoning steps.

which resistance arises (Otani, 1989; Chamberlain et al., 1984). As a result, simulated resistance may appear arbitrary or psychologically incoherent.

We introduce **ResistClient**, a resistance-informed client simulator that models both *how* resistance is expressed and *why* it emerges. For behavioral expression, we draw on *Client Resistance Theory* (Chamberlain et al., 1984), which characterizes distinct resistance types through their triggers and manifestations. Using this taxonomy, we augment real counseling dialogues to construct the **Resistance-informed Psychological Conversation (RPC)** dataset, which contains diverse, context-dependent resistance patterns. Fine-tuning on RPC shifts the model from generic compliance toward realistic challenging behavior. To model *why resistance emerges*, we adopt the **5P case formulation model** (Johnstone & Dallos, 2013)—a clinical framework explaining how clients' problems develop and persist—as a causally grounded client profile schema. Conditioned on this formulation and the evolving dialogue, ResistClient explicitly reasons about the client's current motivation before generating each response, grounding resistant behavior in both the client's psychological background and conversational context. Expert-informed process rewards further align motivational reasoning and response generation with human preferences while promoting decision–response consistency. Together, these mechanisms enable ResistClient to balance challenge intensity with behavioral plausibility, supporting realistic and interpretable simulations of difficult therapeutic interactions.

Automatic and expert evaluations show that ResistClient elicits stronger resistance while maintaining behavioral plausibility and motivation–behavior consistency. We further use ResistClient to evaluate existing psychological LLMs under resistant interactions. Results reveal substantial limitations in managing resistance and maintaining therapeutic progress. These results suggest that compliant simulators can conceal clinically important weaknesses and that resistance-informed simulation provides a more demanding test environment.

Our contributions can be summarized as follows:

- We systematically study client resistance in LLM-based psychological simulation and construct **RPC**, a theory-guided dataset covering diverse client profiles and resistance patterns. We develop **ResistClient** to shift client simulation from over-compliance toward authentic challenging interactions.
- We introduce a new modeling perspective for client simulation that moves beyond surface-level response imitation to capture clients' psychological processes, align motivational reasoning and response generation with human preferences, and promote decision–response consistency, thereby enabling challenging yet psychologically coherent client simulation.
- Extensive automatic and expert evaluations demonstrate ResistClient achieves state-of-the-art challenge intensity and behavioral plausibility. Using ResistClient to evaluate existing psychological LLMs, we further reveal their performance degradation under resistant interactions, enabling researchers to evaluate and improve the resilience of psychological LLMs in real-world scenarios.

2 RELATED WORK

Overly-Compliant Behavior in Client Simulation. Recent LLM-based client simulators enable profile-conditioned generation (Wang et al., 2024; Chen et al., 2025) and improve consistency through state-tracking or memory mechanisms (Yang et al., 2025b; Wang et al., 2025b). Despite prompt-level strategies such as information withholding (Kim et al., 2025a), low-receptivity settings (Yang et al., 2025b), and emotion injection (Wang et al., 2025b), these simulators remain overly compliant because of the inherent compliance bias of instruction-aligned LLMs (Lu et al., 2024; Xu et al., 2023). This limitation also affects counselor-side systems such as *MIRROR* (Kim et al., 2025b), which trains VLMs to manage resistance using prompt-based virtual clients. More critically, psychological theory views challenging client behaviors as manifestations of underlying cognitive–affective resistance (Otani, 1989), rather than as isolated behaviors. ResistClient addresses this limitation by grounding client simulation in Client Resistance Theory (Chamberlain et al., 1984) and jointly modeling resistant behaviors and their underlying motivational reasoning.

Building Paradigms for User Simulation. Existing user simulators follow four main paradigms. (1) *Profile-conditioned generation* (Wang et al., 2024) relies on the intrinsic reasoning capabilities of LLMs but struggles with behavioral consistency. (2) *Agent-based user simulators* (Wang et al., 2025b; Yang et al., 2025b) improve consistency through agentic frameworks but lack utterance-level realism due to limited real-world training data (Wang et al., 2025a). (3) *Conditional supervised fine-tuning* (Kong et al., 2024; Wang et al., 2025a) learns from real data but primarily achieves surface-level imitation. (4) *Non-collaborative user simulation* counteracts the inherent politeness of LLMs by modeling uncooperative behaviors (Zhang et al., 2024; Shim et al., 2026), but lacks expertise in psychological counseling. Across these paradigms, existing methods prioritize observable behaviors over latent reasoning processes, limiting their utility for psychological training.

Reasoning Reinforcement Learning for LLMs. Recent RRL frameworks, such as DeepSeek-R1 (Guo et al., 2025), explicitly model intermediate reasoning steps and use reinforcement learning to align reasoning quality with human preferences. Recent studies extend this paradigm to psychological domains: Psyche-R1 (Dai et al., 2026) enhances empathy- and expertise-oriented reasoning, while Mindora (Xiao et al., 2025) improves clinically aligned counselor reasoning. However, these approaches target counselor-side cognition, leaving client-side motivational reasoning largely unexplored. We extend process-supervised reinforcement learning to client simulation, aligning both motivational reasoning and response generation with human preferences while promoting consistency between the inferred motivation and the generated response.

3 RESISTCLIENT CONSTRUCTION

We formulate psychological client simulation as a structured conditional generation problem. At dialogue turn t, given a client profile p, dialogue history $\mathcal{H}_{t-1}$, and current counselor utterance u_t^c, the simulator first infers a latent motivation trajectory z_t and then generates the client response u_t^a:

$$z_t \sim \pi_\theta(\cdot \mid p, \mathcal{H}_{t-1}, u_t^c), \qquad u_t^a \sim \pi_\theta(\cdot \mid p, \mathcal{H}_{t-1}, u_t^c, z_t). \tag{1}$$

Here, z_t explains whether resistance should arise in the current context and, if so, how it should be expressed. This decomposition separates psychologically grounded resistance from superficial use of resistant language.

As illustrated in Figure 2, **ResistClient** is developed in two stages. First, we construct the Resistance-informed Psychological Conversation (RPC) dataset and perform resistance-informed supervised fine-tuning (SFT), teaching the model a broad repertoire of cooperative and resistant reactions. Second, we optimize the model's motivation reasoning with process supervision and consistency-aware credit assignment. The first stage primarily addresses ***how*** resistance is expressed; the second addresses ***why*** a particular reaction is plausible for a given client and interaction.

3.1 CONSTRUCTING THE RPC DATASET

Open psychological dialogue corpora contain few resistance episodes, partly because data collection favors safe and supportive interactions. Training directly on them therefore reinforces compliance bias. RPC instead recovers clinically grounded client formulations and introduces localized resistance only when counselor interventions plausibly trigger it.

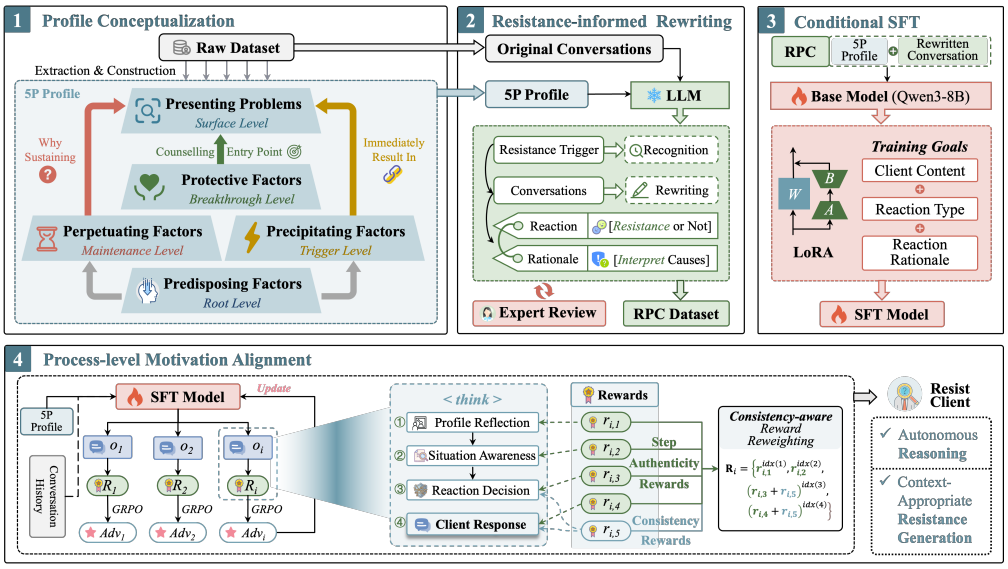


Figure 2: Overview of ResistClient. RPC augments real counseling dialogues with 5P formulations and theory-guided resistance episodes. Resistance-informed SFT learns context-dependent reaction patterns, after which process-level motivation alignment rewards the reasoning validity and decision–response consistency.

3.1.1 Client Formulation with the 5P Model

Shallow personas describe client attributes but provide limited explanation for their reactions. We instead use the 5P case formulation model (Johnstone & Dallos, 2013): (i) *presenting problems*, primary concerns; (ii) *predisposing factors*, historical vulnerabilities; (iii) *precipitating factors*, recent triggers; (iv) *perpetuating factors*, conditions maintaining the problem; and (v) *protective factors*, resources supporting recovery. The resulting profile connects stable background factors with context-dependent reactions and grounds both data construction and model training.

We derive profiles from ProPsyC, a large collection of real-world psychological conversations (Hu et al., 2025). DeepSeek-V3.2 extracts structured 5P formulations using counselor-annotated few-shot demonstrations. Following Wang et al. (2025a), we assess each profile for clinically relevant coverage and faithfulness to the source conversation, discarding low-quality samples. Prompts, filtering criteria, and statistics are provided in Appendix A.

3.1.2 Trigger-Conditioned Resistance Augmentation

We introduce resistance only when psychologically and conversationally warranted. Based on Client Resistance Theory (Chamberlain et al., 1984) and counselor interviews, we define the eliciting conditions and manifestations of each reaction category (Table 5). A counselor turn is considered a trigger when its conversational function—e.g., confronting, interpreting, or pressing for change—interacts with a relevant vulnerability or perpetuating factor in the client's 5P formulation. Conditioning on both profile and dialogue reduces arbitrary resistance insertion based on lexical cues.

For a detected trigger, we select the contextually appropriate reaction from a seven-way taxonomy: five resistant reactions (*controlling*, *emotional*, *defensive*, *avoidant*, and *compliant*) and two cooperative reactions (*non-resistant* and *facilitative*); definitions and examples appear in Table 4. We rewrite the following client utterance to realize the selected reaction while preserving the client's experience and factual content, revising at most the next three turns when needed for coherence. Each session contains at most one primary resistance episode to limit semantic drift. Each client turn is paired with a reaction label and rationale linking it to the profile and triggering context.

3.1.3 EXPERT VERIFICATION AND DATASET STATISTICS

Four professional counselors verify trigger validity and rewritten-response quality. First, they judge whether each detected counselor turn is a valid resistance trigger. Agreement reaches Cohen's $\kappa = 0.72$ between the model and counselors and Fleiss' $\kappa = 0.77$ among counselors. Second, they evaluate the quality of the rewritten client responses, which receive an average score of 2.47 out of 3. Appendix A.4 provides the evaluation criteria and annotation details.

After filtering, RPC contains 1,849 sessions across 14 counseling topics with substantially improved coverage of diverse resistance behaviors. Each session contains a validated 5P profile and annotated conversation. Full statistics and comparison with the source corpus are provided in Appendix A.

3.2 RESISTANCE-INFORMED SUPERVISED FINE-TUNING

We first fine-tune the simulator on RPC to counteract the compliance prior inherited from instruction tuning. For each client turn, the input is $x_t = (p, \mathcal{H}t - 1, u_t^c)$ and the target y_t contains the reaction type, rationale, and client response. Given $\mathcal{D}\mathrm{RPC}$, we minimize the standard autoregressive objective

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{RPC}}}\left[\sum_{j=1}^{|y|}\log \pi_\theta(y_j \mid x, y_{<j})\right]. \tag{2}$$

This stage produces the policy $\pi_{\theta_{\mathrm{SFT}}}$. Because RPC includes both resistant and cooperative reactions, the objective does not reward resistance indiscriminately; it teaches the model to condition its reaction on the client formulation and counselor behavior.

3.3 PROCESS-LEVEL MOTIVATION ALIGNMENT

Although resistance-informed SFT improves behavioral coverage, sequence-level supervision does not distinguish whether errors arise from profile reflection, situation interpretation, reaction selection, or response realization. We therefore introduce **process-level motivation alignment**, where expert judgments supervise each reasoning step and response quality, while a consistency reward encourages the response to realize the selected reaction. We optimize these signals through process-level credit assignment and group-relative policy optimization.

3.3.1 STRUCTURED MOTIVATION TRAJECTORY

Given context x, the policy generates $o = [z^{(1)}; z^{(2)}; z^{(3)}; y]$, where $z^{(1:3)}$ denote three motivation reasoning steps:

1. **Profile reflection** identifies potential resistance tendencies from cognitive and emotional factors in the 5P profile.
2. **Situation awareness** infers the client's current psychological state from the dialogue and counselor utterance.
3. **Reaction decision** selects a reaction type and specifies the behavior through which it should be realized.

The final client response then realizes this decision in natural dialogue. This structured process mirrors human-like psychological reasoning, enabling both interpretability and improved realism.

We sample trajectories from $\pi_{\theta_{\mathrm{SFT}}}$ and collect expert judgments on four criteria: profile grounding, situational validity, decision plausibility, and response quality. Experts additionally assess *decision–response consistency*, i.e., whether the final response realizes the selected reaction. The annotation protocol is provided in Appendix B.

3.3.2 CONSISTENCY-AWARE PROCESS REWARDS

The process rewards operationalize two complementary alignment goals: grounding each intermediate analysis step in the profile and dialogue context, and ensuring that the final behavior realizes

the selected reaction. Let $r_i^{(1)}, \ldots, r_i^{(4)}$ denote the process rewards for trajectory o_i and r_i^{con} its decision–response consistency reward. We attach each reward to the final token e_k of its segment and add consistency to both the *reaction decision* and *response*:

$$\begin{aligned} R_i(e_1) &= r_i^{(1)}, & R_i(e_2) &= r_i^{(2)}, \\ R_i(e_3) &= r_i^{(3)} + \lambda r_i^{\text{con}}, & R_i(e_4) &= r_i^{(4)} + \lambda r_i^{\text{con}}, \end{aligned} \tag{3}$$

where λ controls the contribution of the consistency term. For a group of G trajectories generated for the same context, each segment reward is standardized within the group, yielding $\widetilde{R}_i(e_k)$. The advantage of token j is the reward-to-go over all segments that end at or after that token:

$$\widehat{A}_{i,j} = \sum_{k:e_k \geq j} \widetilde{R}_i(e_k). \tag{4}$$

Thus, inconsistency penalizes both the final response and its preceding decision, while earlier reasoning receives credit from downstream outcomes.

3.3.3 Offline Group-Relative Policy Optimization

We optimize process-level alignment using an offline token-level variant of GRPO. For each context x, the frozen SFT policy $\pi_{\theta_{\text{SFT}}}$ samples a group $o_{i=1}^{G}$ scored by experts. With the likelihood ratio $\rho_{i,j}(\theta) = \pi_\theta(o_{i,j} \mid x, o_{i,<j}) / \pi_{\theta_{\text{SFT}}}(o_{i,j} \mid x, o_{i,<j})$, the objective is

$$\begin{aligned} \mathcal{J}_{align}(\theta) = \mathbb{E}_{x,\{o_i\}_{i=1}^{G}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{j=1}^{|o_i|} \Big(\min \left(\rho_{i,j} \widehat{A}_{i,j}, \text{clip}(\rho_{i,j}, 1-\epsilon, 1+\epsilon) \widehat{A}_{i,j} \right) \\ - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \Big) \Bigg], \end{aligned} \tag{5}$$

where ϵ is the clipping threshold, $\pi_{\text{ref}} = \pi_{\theta_{\text{SFT}}}$ is the fixed reference policy, and β controls deviation from the SFT model. The trainable policy is initialized from $\pi_{\theta_{\text{SFT}}}$. This optimization procedure translates expert process supervision into fine-grained credit for the motivation–behavior trajectory while preserving the stability of the SFT policy.

4 Experiments

We evaluate ResistClient from three perspectives: controlled resistance generation, interactive simulation, and stress testing of counseling models. We address four research questions: RQ1: Resistance simulation. How well does ResistClient simulate resistance compared with reasoning LLMs? RQ2: Stage-wise contribution. How do resistance-informed SFT and process-level motivation alignment contribute? RQ3: Interactive realism. How realistic are ResistClient's challenging behaviors compared with existing simulators? **RQ4: Evaluation utility.** What is the performance of current psychological LLMs when handling resistance in interactions with ResistClient?

Unless otherwise stated, profiles are drawn from the held-out RPC test split. RPC is partitioned by profile into disjoint train/dev/test sets (7:2:1). To ensure evaluation independence and reliability, data annotation and model evaluation use two disjoint groups of four professional counselors. Prompts, rubrics, and interaction protocols are provided in Appendix C.

4.1 Resistance Simulation Capability Analysis

Setup. We evaluate resistance simulation under controlled scenarios by reusing counselor turns from RPC. All baselines receive identical reaction descriptions and expert-curated examples (Appendix C.1.1). We sample 100 held-out profiles and generate three sessions per profile.

Baseline Models. We compare ResistClient with representative reasoning-capable LLMs, including: 1) large-scale models: GPT-5.1 (OpenAI, 2025), DeepSeek-V3.2 (Liu et al., 2025), Kimi-K2-thinking (Kimi Team et al., 2025), GLM-4.6 (Zeng et al., 2025), 2) open-source small-scale

Table 1: Resistance simulation performance of different models in psychological conversations.

Model	Automated metrics (%)				Manual metrics		
	Precision	Recall	F1 score	RTF	Fid. (↑)	Rat. (↑)	Qua. (↑)
GPT-5.1	59.31	62.88	61.04	35.94	1.42	1.35	2.52
DeepSeek-V3.2	52.87	57.56	55.12	36.91	1.29	1.21	2.40
Kimi-K2-Thinking	47.08	51.19	49.05	36.86	1.24	1.19	2.18
GLM-4.6	49.10	54.72	51.76	37.78	1.27	1.23	2.34
Qwen3-8B	36.52	48.54	41.68	45.06	1.10	1.04	1.88
DeepSeek-R1-8B	34.67	46.26	39.64	**45.23**	0.98	1.08	1.72
Qwen3-8B-SFT	63.54	73.90	68.33	39.43	1.46	1.41	2.39
ResistClient	**70.38**	**78.95**	**74.42**	38.03	**1.63**	**1.58**	**2.61**

models: Qwen3-8B (Yang et al., 2025a), DeepSeek-R1-8B (Guo et al., 2025), and 3) our ablations: Qwen3-8B-SFT.

Metrics. Automated metrics evaluate the timing and accuracy of resistance generation, including Precision, Recall, F1, and Resistance Trigger Frequency (RTF). Human evaluators rate outputs on a 0–3 scale for *resistance fidelity*, *resistance rationality*, and *reasoning quality*. Scores are averaged across expert annotators; the full rubric is given in Appendix C.1.2.

Implementation details. RESISTCLIENT uses Qwen3-8B as its backbone. We first perform resistance-informed SFT on RPC for two epochs, followed by two epochs of process-level motivation alignment using consistency-aware process rewards and offline GRPO. Both stages are trained on one NVIDIA A100 80 GB GPU. At inference, we use temperature 0.7, $\mathrm{top_p} = 0.8$, and $\mathrm{top_k} = 20$.

4.1.1 PERFORMANCE COMPARISON (RQ1)

Table 1 shows that ResistClient consistently outperforms all baselines across automated evaluations, indicating improved resistance timing and simulation accuracy. Specifically, the superior precision indicates ResistClient avoids excessive or inappropriate resistance, while strong recall suggests effective conflict recognition without over-compliance. Smaller open-source models (e.g., Qwen3-8B) exhibit higher RTF but substantially lower Precision, indicating a tendency toward over-triggering resistance. Human evaluation confirms that ResistClient achieves the highest scores in fidelity, rationality, and reasoning quality, indicating that both its surface behaviors and internal motivation reasoning are psychologically grounded and contextually appropriate. A detailed case study illustrating these dynamics is provided in the Appendix D.

4.1.2 ABLATION STUDY (RQ2)

To examine the contribution of each training stage, we compare ResistClient with two variants: 1) Qwen3-8B, with prompt-based conditioning only, and 2) Qwen3-8B-SFT, trained with SFT on RPC. Figure 3 visualizes reaction-type alignment using confusion matrices, where rows correspond to ground-truth reactions and columns to generated reactions. The diagonal mass increases from the prompt-only to SFT and to the full ResistClient model, indicating progressively improved simulation accuracy across reaction categories and confirming the contributions of both training stages. Notably, the prompt-only model shows a strong bias toward cooperative reactions, consistent with the over-compliance tendency of pretrained LLMs. SFT effectively reduces this bias by learning diverse resistance patterns from RPC. The process-level alignment stage further improves context-appropriate reaction selection, particularly by reducing confusion among resistance categories. Furthermore, the competitive performance of the SFT-only variant against large closed-source models highlights the effectiveness of the RPC dataset.

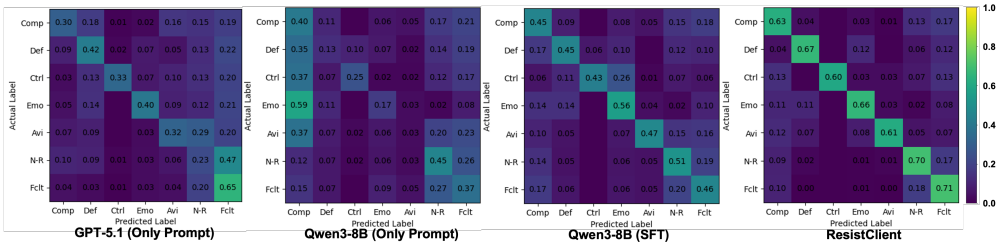


Figure 3: Confusion matrices comparing reaction-type generation across variants. Rows represent ground-truth reaction types; columns represent generated types.

4.2 Quality of Challenging Behaviors

Setup. We evaluate challenging behaviors in full sessions using the counselor–client–moderator framework of Yang et al. (2025b). SoulChat2.0 (Xie et al., 2025) serves as the fixed counselor, and all simulators use the same 50 test profiles. A moderator determines session termination using standardized criteria (Appendix C.2.1).

Baseline models. We reproduce representative challenging behaviors induction strategies: 1) Patient-ψ (Wang et al., 2024), profile-conditioned simulation without explicit challenging behavior modeling; 2) AnnaAgent (Wang et al., 2025b), which introduces challenges by injecting random emotion tags at each turn via an Emotion Perturber; and 3) the method of Yang et al. (2025b), which introduces challenges through low-receptivity control in profile (Appendix C.2.1).

Metrics. We evaluate behavior quality and challenge appropriateness using both automated and human evaluations. Automated metrics include: Client Cooperation Rate (CCR), the proportion of cooperative reactions (lower values indicate stronger challenge); Conversation Turns, the average session length (longer sessions suggest higher interaction difficulty); and Coherence (Coh.), semantic consistency measured by cosine similarity of frozen embeddings across turns. Human evaluations (0-3): Realism (Real.), assessing authenticity of the resistance-cooperation dynamics, and Consistency (Cons.), evaluating alignment between responses and profile (Appendix C.2.2). We further conduct an A/B expert evaluation under matched conditions. Four professional counselors compare ResistClient with each baseline using scores of $+1$ (win), -1 (loss), and 0 (tie), with the mean reported as the A/B score.

4.2.1 Performance Comparison(RQ3)

As shown in Table 2, ResistClient achieves the best balance between challenge intensity (lowest CCR and longest average turns) and behavioral plausibility (highest Coh. and Real. scores). In contrast, Patient-Ψ remains coherent but exhibits excessive cooperation (87.94% CCR). AnnaAgent reduces cooperation (78.62% CCR) by injecting negative emotions, but context-agnostic random perturbations compromise coherence and realism. Yang et al. further increase challenge via low-receptivity control (62.33% CCR) and maintain consistency through constraint mechanisms, but produce repetitive low-receptivity patterns lacking real resistance diversity, leading to low realism. The A/B evaluation further shows that ResistClient is consistently preferred over existing simulators while achieving simulation quality comparable to human responses. Grounded in Client Resistance Theory, ResistClient systematically models diverse resistance patterns while ensuring profile- and situation-awareness through motivation reasoning, achieving superior challenge induction without sacrificing behavioral authenticity.

4.3 Performance of Psychological LLMs with ResistClient(RQ4)

Setup. We use ResistClient to evaluate how existing psychological LLMs handle resistance in interactive counseling. We retain the counselor–client–moderator protocol from Section 4.2 with 100 held-out profiles (Appendix C.3.1).

Test models. We evaluate representative psychological LLMs, including MeChat (Qiu et al., 2024), MindChat (Yan & Xue, 2024), Psyche-R1 (Dai et al., 2026) and SoulChat2.0 (Xie et al., 2025).

Table 2: Evaluation results of client simulators.

Client simulator	Automated metrics			Manual metrics		
	CCR (↓)	Turns	Coh. (↑)	Real. (↑)	Cons. (↑)	A/B test
PATIENT-Ψ	87.94	11.24	0.51	1.87	1.32	0.469
AnnaAgent	78.62	12.65	0.62	1.95	1.60	0.327
Yang et al. (2025b)	62.33	16.67	0.68	2.01	**1.83**	0.254
ResistClient	**60.84**	**17.88**	**0.73**	**2.39**	1.75	–
vs. Human	\	\	\	\	\	−0.085

Table 3: Performance of psychological LLMs when interacting with ResistClient.

Model	Automated metrics		Manual metrics		
	RTF	Turns	Eff. (↑)	CDD (↓)	CPD (↑)
GPT-5.1	41.14	15.24	2.04	1.58	1.92
Gemini-3-Flash	43.77	21.52	1.82	2.02	1.87
DeepSeek-V3.2	40.38	18.40	1.93	1.67	2.05
GLM-4.6	44.54	24.36	1.87	1.75	1.83
MeChat	51.93	43.22	1.61	2.08	1.68
MindChat	48.75	32.76	1.72	1.81	1.79
Psyche-R1	**38.32**	21.67	2.08	**1.48**	1.98
SoulChat2.0	39.15	17.88	**2.14**	1.56	**2.05**

Motivated by evidence that many users seek mental health support from general-purpose LLMs (Guo et al., 2024), we additionally evaluate widely used general models, including GPT-5.1 (OpenAI, 2025), Gemini-3-flash (Google Gemini Team, 2025), DeepSeek-V3.2 (Liu et al., 2025), and GLM-4.6 (Zeng et al., 2025) (Appendix C.3.2).

Metrics. Automated metrics include Resistance Trigger Frequency (RTF) and dialogue turns. Experts rate *strategy effectiveness* (Eff.), *counseling drift degree* (CDD), and *counseling progress degree* (CPD) on a 0–3 scale; definitions are provided in Appendix C.3.3.

4.3.1 Overall Performance (RQ4)

Table 3 reveals that both general and specialized models frequently elicit client resistance (RTF 39-52%), underscoring the necessity for targeted resistance management training. Domain-specific models like SoulChat2.0 and Psyche-R1 achieve competitive performance with large-scale general-purpose models, suggesting that domain-specific fine-tuning can effectively maintain therapeutic focus. However, most models exhibit elevated drift and limited progress, indicating challenges in adapting intervention strategies under resistance. Overall, ResistClient provides a complementary evaluation lens for assessing the clinical robustness of psychological LLMs, exposing weaknesses in resistance handling that may not surface in standard benchmarks and providing insights into model behavior under complex therapeutic scenarios.

5 Conclusion

In this paper, we present ResistClient, a resistance-informed client simulator that addresses the over-compliance bias of existing psychological simulators. ResistClient combines RPC-based resistance-informed SFT with process-level motivation alignment: the first stage expands context-dependent cooperative and resistant behaviors, while the second uses expert process supervision to align motivational reasoning and response generation with human preferences and promote decision–response consistency. Extensive automatic and expert evaluations demonstrate improved challenge fidelity, behavioral plausibility, and reasoning coherence. Evaluations using ResistClient further reveal substantial limitations in existing psychological LLMs when handling client resistance.

AI USE STATEMENT

Aid or polish writing. We used GPT-5.6 to identify grammatical errors, improve linguistic clarity, and check formatting. It was not used to formulate the research questions, develop the methodology, interpret results, or generate scientific claims.

Retrieval and discovery. We used GPT-5.6 to identify potentially relevant recent literature based on research topics and keywords designed by the authors. The authors independently examined the original publications, verified their relevance and bibliographic information, and decided which works to include and discuss.

Generating synthetic datasets. We used DeepSeek-v3.2 to extract structured 5P client profiles from author-selected source dialogues and to assist with context-dependent rewriting and annotation when constructing the synthetic portions of the RPC dataset. The prompts and resistance framework were designed by the authors, and the resulting data underwent systematic quality control and expert verification, as detailed in Appendix A.

The authors reviewed and verified all AI-assisted text, references, annotations, and synthetic data, and take full responsibility for the final content of this work.

ETHICS STATEMENT

All source datasets are publicly available or were used with appropriate permission. The source dialogues were anonymized before use, and RPC contains no personally identifiable information. Its rewritten resistance-related utterances and the interactions generated in our experiments are synthetic and should not be interpreted as records of specific individuals or predictions of their psychological states.

Four qualified professional counselors participated in data annotation and verification. A separate group of four qualified professional counselors conducted the human evaluation of the generated responses and simulated interactions. All participating counselors had relevant clinical experience and were compensated at a rate of $14.30 per hour. Their judgments were treated as expert reference annotations and evaluations.

ResistClient is intended solely for research, education, and controlled evaluation; it must not be used for diagnosis, treatment, or real-world mental health intervention without appropriate professional oversight. Under the applicable institutional requirements, formal ethics review was not required because the study involved no patient recruitment, identifiable personal data, or real-world clinical intervention.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the RPC dataset construction and validation procedures in Appendix A, including data selection, 5P profile extraction, resistance-trigger recognition, conversation rewriting, annotation guidelines, and human verification. The construction of the motivation-reasoning training data and the associated reward criteria are documented in Appendix B. Appendix C reports the experimental protocols, baseline implementations, model prompts, inference settings, evaluation metrics, human-evaluation rubrics, and counselor–client interaction procedures. Qualitative examples and case studies are included in Appendices D and E. We report the principal training and inference configurations in the main text and appendix. An anonymous repository containing supplementary materials is available at `https://anonymous.4open.science/r/ResistClient`. Upon publication, we intend to release the RPC dataset, model checkpoints, training and evaluation code, and prompts, subject to the licenses and privacy requirements of the source datasets.

REFERENCES

Patricia Chamberlain, Gerald Patterson, John Reid, Kathryn Kavanagh, and Marion Forgatch. Observation of client resistance. *Behavior therapy*, 15(2):144–155, 1984.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pp. 2318–2335, 2024.

Zhuang Chen, Yaru Cao, Guanqun Bi, Jincenzi Wu, Jinfeng Zhou, Xiyao Xiao, Si Chen, Hongning Wang, and Minlie Huang. Socialsim: Towards socialized simulation of emotional support conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1274–1282, 2025.

Chongyuan Dai, Jinpeng Hu, Hongchang Shi, Zhuo Li, Dan Guo, Xun Yang, and Meng Wang. Psyche-r1: Towards reliable psychological llms through unified empathy, expertise, and reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24889–24906, 2026.

Google Gemini Team. Gemini 3 flash. `https://deepmind.google/technologies/gemini/`, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, et al. Large language models for mental health applications: systematic review. *JMIR mental health*, 11(1): e57400, 2024.

Yuxin Hu, Danni Liu, Bo Liu, Yida Chen, Jiuxin Cao, and Yan Liu. Psyadvisor: A plug-and-play strategy advice planner with proactive questioning in psychological conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12205–12229, 2025.

Lucy Johnstone and Rudi Dallos. *Formulation in Psychology and Psychotherapy: Making sense of people's problems*. Routledge, 2013.

Minju Kim, Dongje Yoo, Yeonjun Hwang, Minseok Kang, Namyoung Kim, Minju Gwak, Beongwoo Kwak, Hyungjoo Chae, Harim Kim, Yunjoong Lee, et al. Can you share your story? modeling clients' metacognition and openness for llm therapist evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25943–25962, 2025a.

Subin Kim, Hoonrae Kim, Jihyun Lee, Yejin Jeon, and Gary Lee. Mirror: Multimodal cognitive reframing therapy for rolling with resistance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14851–14880, 2025b.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. Platolm: Teaching llms in multi-round dialogue via a user simulator. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7841–7863, 2024.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. Supervisorbot: Nlp-annotated real-time recommendations of psychotherapy treatment strategies with deep reinforcement learning. In *International Joint Conference on Artificial Intelligence*, 2023.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7828–7840, 2024.

OpenAI. Gpt-5.1. https://openai.com/index/gpt-5-1/, 2025. Accessed: 2025-11-12.

Akira Otani. Client resistance in counseling: Its theoretical rationale and taxonomic classification. *Journal of Counseling & Development*, 67(8):458–461, 1989.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 615–636, 2024.

Jeonghoon Shim, Woojung Song, Cheyon Jin, Seungwon KooK, and Yohan Jo. Non-collaborative user simulators for tool agents. In *International Conference on Learning Representations*, volume 2026, pp. 64486–64534, 2026.

Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2025.

Herbert S. Strean. *Resolving Resistances in Psychotherapy*. Brunner/Mazel, 1985.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 21082–21107. Association for Computational Linguistics, 2025a.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. Annaagent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23221–23235, 2025b.

Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. Patient-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12772–12797, 2024.

World Health Organization WHO. *World Mental Health Today: Latest Data*. World Health Organization, 2025. ISBN 978-92-4-011381-7.

Mengxi Xiao, Kailai Yang, Pengde Zhao, Enze Zhang, Ziyan Kuang, Zhiwei Liu, Weiguang Han, Shu Liao, Lianting Huang, Jinpeng Hu, Min Peng, Qianqian Xie, and Sophia Ananiadou. Mentrasuite: Post-training large language models for mental health reasoning and assessment. *arXiv preprint arXiv:2512.09636*, 2025.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1081–1115, 2025.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6268–6278, 2023.

Xin Yan and Dong Xue. Mindchat: Psychological large language model. https://github.com/X-D-Lab/MindChat, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yizhe Yang, Palakorn Achananuparp, He-Yan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-peng Lim. Consistent client simulation for motivational interviewing-based counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20959–20998, 2025b.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 424–444, 2024.

Appendix Roadmap

The appendix is organized as follows. Appendix A details the construction and validation of the RPC dataset. Appendix B describes the construction of the process-supervision data for motivation alignment. Appendix C presents the experimental setups and evaluation metrics. Finally, Appendices D and E provide qualitative case studies of counseling adaptation and resistance simulation, respectively.

A Details of Dataset Construction with Resistance Annotations

This appendix describes the detailed procedure for constructing a psychological conversation dataset with fine-grained client resistance labels (our RPC dataset). Starting from the ProPsyC dataset(Hu et al., 2025), we selectively rewrote client utterances in sessions likely to elicit resistance and subsequently annotated all client turns with resistance categories. The construction pipeline proceeds in five stages: data selection, 5P profile extraction, resistance-guided rewriting and annotation, label-distribution analysis, and human verification.

A.1 Data Selection

To obtain a thematically balanced corpus, we re-screened the original ProPsyC dataset according to its four major clinical themes (depression and emotional disorders, interpersonal relationships, academic/career stress, and general growth issues). Sessions were sampled to achieve approximately equal representation across themes. The final theme distribution is shown in Figure 4.

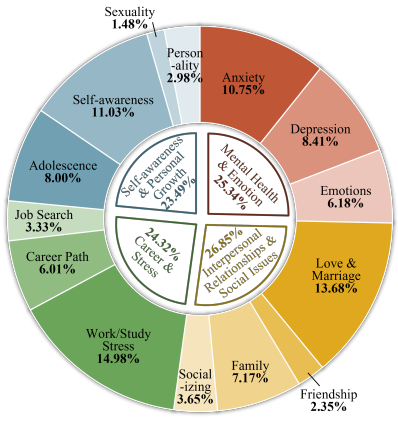


Figure 4: Distribution of different topics in client profiles.

A.2 Extraction of Client 5P Profiles

For each of the selected sessions, we used DeepSeek-V3.2 to generate a structured 5P case conceptualization profile that captures the client's psychological background. The model was prompted with the full conversation and instructed to summarize the following 5 dimensions:

- *Presenting Problems*: The client's most immediate and subjectively distressing concerns—the surface-level issues brought to therapy.
- *Predisposing Factors*: Long-standing historical, biological, psychological, or social vulnerabilities that increase the likelihood of the current difficulties.
- *Precipitating Factors*: Recent events or stressors that directly triggered or exacerbated the current problems.
- *Perpetuating Factors*: Ongoing cognitive, behavioral, interpersonal, or environmental patterns that maintain the problems and hinder change.
- *Protective Factors*: Internal strengths, external resources, and support systems that help the client cope and foster resilience.

The validation process for the 5P Profile can be found in Appendix A.4. These 5P profiles were later injected into the LLM as persona prompts to ensure psychologically consistent client responses during rewriting. A 5P profile example from the dataset is shown in Figure 6.

A.3 Resistance-Guided Conversation Rewriting and Annotation

A.3.1 Resistance Category Definitions

We adopted 7 mutually exclusive client response categories (5 resistant reactions and 2 cooperative reactions), defined in collaboration with licensed psychotherapists (see Table 4).

A.3.2 Automated Rewriting Procedure

The rewriting process was performed by DeepSeek-V3.2, which was provided with (a) the original session, (b) the client's 5P profile, and (c) the resistance category definitions. For each counselor turn, the model first judged whether the turn constituted a plausible resistance trigger. Based on the counselors' advice, we clarify the triggering conditions for each type of resistance(shown in Table 5). When a trigger was identified, DeepSeek-V3.2 not only rewrote the immediate client response to express one of the five resistance types (chosen as most psychologically plausible given the context and 5P profile), but also adaptively revised the subsequent 2-3 conversational turns to maintain contextual coherence and psychological realism. Each rewritten resistance response was accompanied by a brief motivation statement explaining the underlying psychological mechanism. Conversations deemed unsuitable for resistance elicitation were discarded. An illustrative example of this rewritten dataset is presented in Figure 7. This process ultimately produced 1,761 resistance-containing conversations, constituting our RPC dataset. The remaining 88 sessions contain no resistance triggers; in these cases, client responses remain cooperative, suggesting that resistance arises only when contextually justified. Professional counselors have verified the annotation results. For more details, please see Appendix A.4.

A.3.3 Final Label Distribution

The distribution of resistance categories in both the original dataset and the rewritten RPC dataset is presented in Figure 5. Resistance behaviors in the original dataset are relatively sparse, with instance counts ranging from 31 (Controlling) to 264 (Compliant). Our rewriting framework substantially expands coverage across all five categories, yielding 1,277 Compliant, 975 Defensive, 624 Emotional, 363 Avoidant, and 116 Controlling resistance instances in the final RPC dataset, corresponding to an overall amplification of approximately 4 times, while preserving the relative distributional structure of the original data. This results in a more balanced and densely populated training set compared to the original data.

This pattern aligns with clinical observations that overt confrontation (Controlling) is relatively rare in Chinese counseling contexts, whereas indirect and self-protective forms (Compliant and Defensive) predominate. Specifically, Compliant Resistance is most frequent, reflecting a tendency for clients to maintain surface-level cooperation and politeness while avoiding deeper engagement, which aligns with cultural values emphasizing harmony and respect for authority. Defensive and Emotional Resistance occur relatively often, as clients may protect their self-image or react emotionally when feeling judged or emotionally challenged, though strong emotional expression is often

Table 4: Classification and interpretation of dataset labels.

Reaction Category	Label Type	Behavioral Description	Motivational Interpretation
Resistant Reactions	Controlling Resistance	- Dominates the conversation /interrupts - Rejects guidance on conversation direction - Clings to own views, disregards input	Seeks to maintain control to avoid being influenced; insists on own perspective and resists change.
	Emotional Resistance	- Expresses anger or aggressive speech - Displays sadness/emotional breakdown - Shows despair or overgeneralized negative emotions	Externalizes emotions or projects blame to avoid confronting deeper psychological pain.
	Defensive Resistance	- Questions the practitioner's expertise or competence - Challenges the consultation method or process - Responds with sarcasm or exaggeration	Projects anxiety onto the practitioner to prevent internal conflicts from being triggered.
	Avoidant Resistance	- Changes or introduces unrelated topics - Provides excessive or redundant information - Evades direct questions	Reflects cognitive avoidance; an unconscious effort to prevent core conflicts from being addressed.
	Compliant Resistance	- Gives vague or perfunctory responses - Shows superficial cooperation without genuine engagement - Downplays emotions with brevity	Superficial compliance with avoidance of substantive issues; does not contribute to therapeutic progress.
Cooperative Reactions	Non-resistant Reaction	- Neutral, cooperative, shows no obvious opposition, but does not take initiative	Indicates a satisfactory interactive state with a degree of openness, but not actively driving the process.
	Facilitative Reaction	- Expresses agreement, actively listens, and shows willingness to explore issues in depth	Reflects sufficient safety and trust; the client is willing to share information and address problems actively.

moderated. Avoidant Resistance is less common, as clients tend to remain within the conversational frame rather than openly shifting topics. Controlling Resistance appears least frequently, likely because overtly assertive or confrontational behaviors conflict with culturally valued norms of deference and relational caution.

A.4 Data Annotation Verification

To ensure the reliability of all annotations involving LLMs, we conducted systematic human validation across all stages of the annotation process. Four licensed professional psychological counselors participated in the validation procedure and were involved throughout the entire process, and signed informed consent forms for research participation(see Figure9). All of them received a total of 60 hours of training before the annotation process.

Table 5: Resistance trigger timing specification.

Resistance type	Typical trigger situations	Psychological problems	High-risk 5P profile features
Controlling resistance	Direct advice, reframing attempts, or agenda-setting that challenges the client's autonomy or preferred narrative.	Autonomy, agency, and perceived self-coherence.	*Predisposing*: High need for control and rigid belief systems. *Perpetuating*: Interpersonal dominance patterns.
Emotional resistance	Emotion-focused prompts or interpretations that surface intense affect without sufficient stabilization or safety cues.	Affect regulation, emotional tolerance, and defenses against vulnerability.	*Precipitating*: Acute relational loss or trauma exposure. *Perpetuating*: Poor emotion-regulation strategies.
Defensive resistance	Questions or challenges concerning the counselor's methods, competence, or therapeutic intent.	Threatened self-image and externalized anxiety.	*Predisposing*: Prior negative counseling experiences. *Perpetuating*: Mistrust of authority figures.
Avoidant resistance	Exploratory questions addressing core conflicts, personal responsibility, or emotionally salient themes.	Cognitive avoidance and attentional disengagement.	*Predisposing*: Habitual avoidant coping. *Perpetuating*: Reinforcement of distraction-based regulation.
Compliant resistance	Open-ended or reflective prompts that invite exploration beyond surface-level agreement.	Superficial compliance masking emotional disengagement.	*Predisposing*: Fear of interpersonal conflict or rejection. *Perpetuating*: Overadaptation to perceived expectations.

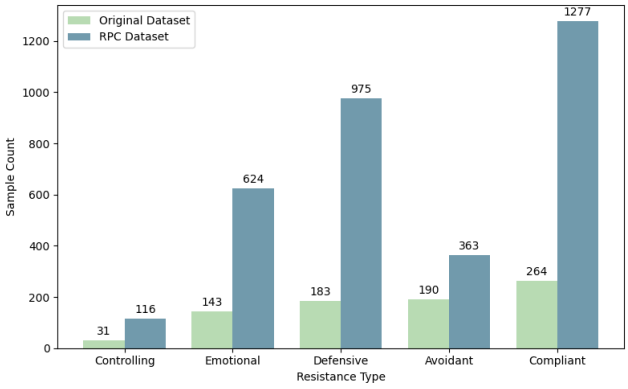


Figure 5: Distribution of resistance types in the original dataset and the RPC dataset across five resistance categories.

A.4.1 5P Profiles Verification

Following the thematic distribution shown in Figure 4, we randomly sampled 50 conversations from each of the four major themes, resulting in a total of 200 conversations for validation. For each sampled conversation, the full multi-turn conversation was provided. Both the LLM and the human counselors reviewed all conversation turns and independently summarized the case according to the predefined 5P profile framework.

The 4 counselors evaluated the extracted 5P profiles by comparing them to the original conversation transcripts. Their assessment focused on two key criteria: **coverage** (whether the profile fully captured the core content and essential elements discussed in the conversation) and **faithfulness** (whether the profile remained factually consistent with the conversation, avoiding inaccuracies, misinterpretations, or extraneous information). A profile was considered successful only if it satisfied both criteria—being both comprehensive and faithful. Profiles that were incomplete or contained factual distortions were classified as failures. Initial evaluations yielded a success rate of only 53%. Based on systematic feedback from the counselors, we identified two main sources of disagreement: overly verbose model outputs and insufficient abstraction in profile summaries.

To address these issues, we refined the prompting strategy from two aspects. First, the model output format was constrained to concise keywords rather than full descriptive sentences, which facilitated clearer abstraction and reduced stylistic variance. Second, we incorporated few-shot learning by providing high-quality examples annotated by counselors, allowing the model to better capture the expected summarization style and semantic focus.

After several rounds of iterative refinement, we re-extracted 5P Profile for the same 200 conversations, achieving a success rate of 82%. Inter-rater agreement among the counselors was substantial, with Fleiss' $\kappa = 0.74$. Therefore, the optimized prompt template was used to annotate the remaining conversations in the dataset. The prompt template ultimately used for annotation is as follows:

5P Profile Prompt Template

You are a professional counselor.
You will be given a complete multi-turn psychological conversation. Carefully read all conversation turns and summarize the case using the 5P profile framework.
The specific content and explanation of the 5p profile are as follows:
{*Definition and Explanation of 5p Profile*}
You must summarise the conversation in the form of keywords, using concise and precise terminology. You may refer to the following exemplary annotation examples:
{

- **Presenting Problems: "...",**
- **Predisposing Factors: "...",**
- **Precipitating Factors: "...",**
- **Perpetuating Factors: "...",**
- **Protective Factors: "..."**

}
Output strictly in accordance with the **[output Format]** specification.

A.4.2 Trigger Recognition Reliability Verification

During the training phase, the counselors have reached a shared and precise understanding of the classification of resistance behaviors and the underlying psychological motivations.

Following the subtopic distribution shown in Figure 4, we randomly sampled 100 conversations from the dataset. Based on the 7 reaction types defined in Table 4 (including 2 cooperative reactions and 5 resistance reactions), we designed prompts that guided DeepSeek-V3.2 to identify appropriate resistance-triggering turns and rewrite selected conversation segments. Each client turn in the rewritten conversations was annotated by the model with a single reaction type, along with an explicit explanation of the psychological motivation underlying the client's reaction.

The purpose of this verification process was to evaluate whether the model could accurately recognize *when* resistance should emerge in a counseling conversation and generate psychologically plausible client reactions after the identified trigger turns. Therefore, rather than evaluating isolated label matching, we focused on the consistency between the model-generated rewritten reactions and the reactions that expert counselors considered most realistic and contextually appropriate under the same conversational context.

The 4 counselors independently evaluated the annotated conversations. During the training phase, the counselors have reached a shared and precise understanding of the classification of resistance behaviors and the underlying psychological motivations. To assess human reliability, the model-generated labels were masked, and the counselors re-annotated all client turns without access to the model's predictions. Inter-rater reliability among the counselors reached a Fleiss' κ = 0.77, indicating a very high level of agreement. In addition, the counselors assessed the plausibility and coherence of the motivation annotations and documented recurrent issues.

We then compared the counselors' annotations with the model-generated labels. The resulting Cohen's κ = 0.58 between human and the model annotations, reflecting only moderate agreement. Error analysis revealed two primary issues. First, the model frequently misclassified cooperative reactions as Compliant Resistance, particularly when the client's reply appeared agreeable but did not explicitly oppose the counselor. Second, the generated motivation explanations were often rigid and decontextualized, failing to account for the surrounding conversation and leading to interpretations that were inconsistent with the conversational context.

Based on systematic feedback from the counselors, we implemented the following refinements. (1) The definition of Compliant Resistance was revised to clarify that a response should be labeled as such only when the client's apparent cooperation actively hinders the counselor's exploratory intent. (2) The coherence of motivation annotations was enhanced by introducing contextual constraints: when a resistance reaction is followed by a cooperative reaction, the model is required to explain how the counselor's intervention alleviated resistance; when resistance persists across turns, the model must account for the continuity of resistance. (3) The prompt template was further augmented with counselor-annotated examples, including both correct resistance cases and counterexamples in which resistance should not be inferred but was previously misidentified by the model.

After these refinements, the optimized prompt was adopted for full-dataset rewriting and annotation. The final Cohen's $\kappa = 0.72$ between counselor and model annotations indicates substantial agreement, suggesting that the model can reliably identify appropriate resistance-triggering contexts and generate psychologically plausible resistance reactions consistent with expert counselor judgments. The prompt template used for annotation is as follows:

Conversation Rewriting and Annotation Prompt Template

You are an experienced psychological psychological conversation rewriting assistant.
Your goal is to determine, based on the client's 5P Profile and the original conversation, whether it is appropriate to introduce a resistance reaction at specific client turns. If so, rewrite the current client turn and the subsequent 1–3 turns accordingly, while keeping the conversation natural, coherent, and contextually appropriate.
You may assign one of the following 7 reaction types:
{*Detailed explanations of the 2 cooperative reactions and 5 resistance reactions*}
For each client turn, you must assign a state label (one of the 7 reaction types), and a motivation label.
Guidelines for the motivation label:
- If the state is a resistance reaction, explain the psychological motivation that triggers the resistance.
- If the state is a non-resistance or facilitative reaction, briefly explain why this reaction occurs.
- If the previous state was a resistance reaction and the current state changes to a non-resistance or facilitative reaction, explain how the counselor's reply helped alleviate the resistance.
- If the previous state was a resistance reaction and the current state remains a resistance reaction, explain the psychological reason for the continued resistance.
Below is the client's current 5P Profile:
{PROFILE}
Below is the original conversation:
{CONVERSATION}
Below are examples annotated by professional counselors for reference:
{**Examples of correct and well-executed annotation**}
{**Examples of incorrect, impedance-erroneous triggering**}

Now rewrite the conversation based on the given information, and strictly follow the required output format [FORMAT].

A.4.3 Rewriting Quality and Annotation Accuracy Verification

We select DeepSeek-V3.2 as the annotation and rewriting model instead of relying on larger proprietary models such as GPT-5.1 or Gemini-3-Flash. This decision is guided by empirical comparisons on annotation accuracy, rewriting quality, and economic cost, all of which are critical for large-scale resistance-oriented data construction.

To quantify annotation reliability, we randomly sampled 50 psychological conversations and asked 4 professional counselors to annotate the most contextually appropriate client reaction types, which were treated as ground truth labels. Different models were then evaluated based on their agreement with counselor annotations, measuring whether the generated reactions aligned with expert judgments regarding both resistance-trigger timing and psychologically plausible reaction categories.

In parallel, we evaluated rewriting quality through human assessment. Counselors rated whether the rewritten client responses resembled realistic human resistance behaviors, considering contextual coherence, psychological plausibility, and conversational naturalness. Finally, we report the average API cost required to rewrite and annotate a single conversation, measured in U.S. dollars.

As shown in Table 6, GPT-5.1 achieves the highest resistance label accuracy, but the margin over DeepSeek-V3.2 is small (2.1%). In contrast, Gemini-3-Flash exhibits noticeably lower accuracy, suggesting reduced reliability for fine-grained resistance annotation. Importantly, DeepSeek-V3.2 achieves the highest rewriting quality score, indicating that its rewritten responses better capture the nuanced, context-dependent characteristics of human resistance behaviors, as judged by professional counselors.

From a cost perspective, DeepSeek-V3.2 is substantially more economical than GPT-5.1, reducing annotation expenses by approximately 23 times while maintaining comparable accuracy. Given the scale of data required for resistance-aware client modeling, this cost–performance balance is essential. Taken together, these results indicate that DeepSeek-V3.2 provides the most practical trade-off among annotation accuracy, rewriting fidelity, and economic feasibility, making it a suitable choice for resistance labeling and conversation rewriting in this work.

B Process-Supervision Data Construction for Motivation Alignment

B.1 Candidate Response Generation

To construct high-quality process-supervision data for motivation alignment, we randomly sampled 300 distinct 5P profiles, covering all subtopic categories shown in Figure 4. For each profile, the Qwen3-8B model fine-tuned via SFT was employed to role-play as the client, while the psychological conversation was conducted with the psychological LLM SoulChat 2.0. Figure 8 provides an example of a conversation incorporating reasoning processes.

During each conversation turn, the client model generated 3 candidate responses. Prior to producing each response, the model was required to generate and record its reasoning process explicitly. This reasoning process followed a structured chain-of-thought designed in collaboration with pro-

Table 6: Comparison of annotation accuracy(Acc.), rewriting quality(Quality), and API cost(Cost). Rewriting quality is scored by counselors on a 0–3 scale, with higher scores indicating more human-like resistance expression. API Cost is the output price per 1 million tokens.

Models	Acc.(%)	Quality	Cost ($)
GPT-5.1	**74.4**	2.31	10.00
DeepSeek-V3.2	72.3	**2.47**	**0.43**
Gemini-3-Flash	68.6	2.05	3.00

fessional psychological counselors to ensure clinical plausibility and interpretability. The content of the CoT is as follows:

Chain-of-Thought for Reaction Generation

You are a client seeking psychological counseling. Based on the specific 5P profile [PROFILE] and the counselor's utterance, you must present your reasoning process and then provide a final response.
Your reasoning process must strictly follow these steps:

- **Profile Reflection**: Reflect potential resistance tendencies by integrating stable cognitive and emotional factors from the 5P profile.
- **Situation Awareness**: Analyze the conversation history and the current counselor utterance to infer the client's momentary psychological state.
- **Reaction Decision**: Determine the final reaction type and describe the behavioral characteristics that the client should exhibit in the next response.

You must generate the final reply strictly according to the Reaction Decision.
A well-formatted output example is provided below: {Expert Annotation Example}

B.2 Reward Scores Assessment

After response generation, the 4 counselors independently evaluated all candidate responses. For each conversation turn, the counselors scored all 3 candidate responses on a 0–5 scale, where higher scores indicate better performance. The evaluation consisted of 5 dimensions:

- **think_step1_score**: Whether the *Profile Reflection* accurately captures the client's psychological state and motivation in relation to the given 5P profile.
- **think_step2_score**: Whether the *Situation Awareness* correctly identifies the most plausible reaction type.
- **think_step3_score**: Whether the final reaction label and the described behavioral pattern in *Reaction Decision* are reasonable and logically consistent.
- **reply_score**: The naturalness, realism, and consistency of the final reply with the client's profile.
- **consistency_score**: Whether the final reply is generated in accordance with the specified Reaction Decision.

Inter-rater reliability among the 4 counselors reached a Fleiss' κ of 0.73, indicating a high level of agreement and confirming the reliability of the evaluation results. During the process-level motivation alignment stage, the counselors' raw scores ranging from 0 to 5 were converted into scalar rewards via group-wise min–max normalization within each candidate set, preserving relative preference ordering while stabilizing RL optimization. We provide an example of a conversation between our ResistClient and SoulChat 2.0 in Figure 10.

C Experimental Details

C.1 Resistance Effect Verification

C.1.1 Experimental Setup

To systematically evaluate the effectiveness of ResistClient, we compare it against a set of representative baseline models that vary in model scale while sharing a common capability of explicit reasoning. Explicit thinking enables models to generate intermediate reasoning processes, which is essential for modeling resistance behaviors that depend on multi-turn conversational context rather than isolated responses.

Specifically, GPT-5.1, DeepSeek-V3.2, Kimi-K2-Thinking, and GLM-4.6 are selected as large-scale language models with strong overall performance across reasoning and conversation tasks, serv-

ing as high-capacity baselines. In contrast, DeepSeek-R1-8B and Qwen3-8B represent smaller-parameter models, allowing us to examine whether resistance-aware behavior can be effectively learned under more constrained model capacity. Notably, all subsequent experiments in this work are conducted with Qwen3-8B as the backbone model, making it a critical reference point for evaluating the contribution of our training strategy.

Based on Qwen3-8B, we further include Qwen3-8B-SFT, which omits the process-level alignment stage to isolate the effect of supervised fine-tuning, as well as the full ResistClient, which additionally applies expert-informed process supervision through offline GRPO. This design enables a direct comparison between the base model, SFT adaptation, and the final resistance-aware model.

All models are prompted using the same client-role prompt template, which instructs the model to consistently act as a client in a psychological conversation, allowing resistance behaviors to emerge naturally in response to counselor interventions:

Client Role with Resistance Prompt

You are a client in a psychological counseling setting. You must always play the role of a client and engage in a conversation with a counselor.
In the following conversation, you should generate responses based on a given **5P Profile**, defined as follows:

- **Presenting Problems**: ...
- **Predisposing Factors**: ...
- **Precipitating Factors**: ...
- **Perpetuating Factors**: ...
- **Protective Factors**: ...

Now generate a response based on the following 5P Profile:

`{5P_PROFILE_JSON}`

You should determine, based on the conversational context and the counselor's utterance, whether to produce a cooperative response or a resistance response.
Possible response types include:
(1) Cooperative Responses

- *Facilitative Response*: ...
- *Non-resistance Response*: ...

(2) Resistance Responses

- *Controlling Resistance*: ...
- *Emotional Resistance*: ...
- *Defensive Resistance*: ...
- *Avoidant Resistance*: ...
- *Compliant Resistance*: ...

Note that compliant resistance must be identified in context. Brief or factual answers are not considered resistance unless they obstruct the counselor's exploratory intent.
Before responding, you must reason from the client's perspective by following these steps:

- **Profile Reflection**: ...
- **Situation Awareness**: ...
- **Reaction Decision**: ...

The reasoning process must be enclosed within `<think></think>` tags. After the reasoning, output the actual response on a new line.
The final response must follow these rules:

- You may only play the role of the client.
- Do not output system or role tags.

- Do not always comply or deliberately resist; decide based on context.
- Responses should be concise, natural, and emotionally realistic.

{ Expert Output Examples }

For evaluation, we randomly sample 100 conversations from the RPC Dataset and generate three sessions per profile. Each sampled conversation is replayed with the counselors' turns fixed, while the client turns are generated by the evaluated model. This setup ensures that all models are assessed under identical conversational contexts, enabling fair and controlled comparison across baselines.

C.1.2 Evaluation Metrics

To comprehensively evaluate resistance behavior in psychological conversations, we employ a set of automated and manual metrics designed to capture not only whether resistance is generated, but also whether it is appropriate and therapeutically meaningful.

Automated Metrics. All models were evaluated using uniform parameter settings and the identical RPC dataset comprising 100 conversations.

Resistance Precision (Precision) measures the proportion of generated resistance responses that are judged to be appropriate resistance. This metric is introduced to penalize models that overproduce resistance in contexts where cooperation would be more suitable.

Resistance Recall (Recall) measures the proportion of resistance-eliciting contexts in which resistance is correctly generated. This metric reflects a model's sensitivity to situations where resistance is a natural psychological response rather than a failure of interaction.

Resistance F1 Score is the harmonic mean of resistance precision and recall. It is used to balance conservative and aggressive resistance strategies, preventing models from optimizing only one aspect.

Resistance Trigger Frequency (RTF) measures the frequency of resistance responses across all client turns within a conversation. This metric captures a model's global tendency to engage in resistance and complements precision and recall by reflecting overall behavioral patterns.

Manual Metrics. Automated metrics alone are insufficient to capture the qualitative and contextual nature of resistance in counseling. We therefore introduce several manual metrics annotated by professional counselors.

Resistance Fidelity (Fid.) evaluates whether the generated response exhibits behavioral characteristics consistent with the intended resistance type. For example, if the model infers that a *defensive resistance* should be triggered, this metric assesses whether the final response actually reflects defensive traits such as challenging the counselor's authority or questioning the intervention. Scores are assigned in $\{0, 1, 2, 3\}$, with higher scores indicating stronger and more prototypical expression of the target resistance. This metric measures the model's ability to distinguish between different resistance categories rather than merely generating generic non-cooperative responses.

Resistance Rationality (Rat.) evaluates whether the occurrence of resistance is contextually appropriate, scored in $\{0, 1, 2, 3\}$, where higher scores indicate greater rationality. This metric assesses the model's ability to decide *when* resistance should or should not emerge. Responses receive lower scores if resistance is generated in contexts that call for cooperation, or if resistance is absent in situations where psychological theory suggests it should naturally arise.

Reasoning Quality (Qua.) evaluates the quality of the explicit reasoning process that leads to the generated response, scored in $\{0, 1, 2, 3\}$, with higher scores indicating more coherent and higher-quality reasoning. This metric assesses whether the reasoning appropriately integrates conversational context, the client's psychological profile, and the counselor's intervention, and whether the inferred psychological state logically supports the subsequent behavioral response. It thus measures not only surface-level correctness but also the internal consistency and therapeutic plausibility of the model's decision-making process.

All manual metrics are rated by **4 professional counselors**, who directly evaluate the conversations generated during the automated evaluation process.

C.2 Comparison with Challenging Client Simulators

C.2.1 Experimental Setup

We compare *ResistClient* with 3 existing client simulators: *PATIENT-Ψ*(Wang et al., 2024), *AnnaAgent*(Wang et al., 2025b), and the simulator proposed by *Yang et al. (2025b)*. Each client simulator follows its original prompt design, with task-specific prompts to faithfully reproduce the intended behaviors, including any challenge-inducing mechanisms inherent to the original methods. For *ResistClient*, we adopt the prompt described in Appendix C.1.1.

Experiments are conducted on the **RPC Dataset**, from which we randomly sample 50 client profiles. As different client simulators were originally designed under distinct task settings and assumptions, we adapt the profile format for each simulator to ensure compatibility with its original prompt design, while preserving the same underlying psychological content. The counselors made corresponding adjustments to different client simulators based on the 5P Profile we provided, ensuring they were adapted to their original prompt. Furthermore, all simulators utilised the same base model.

For PATIENT-Ψ (Wang et al., 2024), which performs profile-conditioned client simulation primarily based on the intrinsic reasoning capability of the LLM, we follow the original prompt formulation provided in the paper. For AnnaAgent (Wang et al., 2025b), client challenge is introduced by replicating the volatile emotions of real-world clients. Specifically, an *Emotion Perturber* injects a randomly sampled emotion tag at each turn. We adopt the original client role-play system prompt, and at each conversation round, the agent is additionally reminded of the current emotional state (*Positive*, *Neutral*, *Ambiguous*, or *Negative*). For Yang et al. (2025b), client challenge is modeled through receptivity control, where receptivity is specified in the profile on a scale from 1 to 5, and action sampling is conditioned on the receptivity level. To reproduce this mechanism on our dataset, we sampled a subset of RPC, assigned receptivity scores (1–5 following their scale), and computed the resistant vs. non-resistant action distribution at each level. To align with the goal of challenging client simulation, we restrict receptivity to the lower range (1 to 3) during evaluation.

PATIENT-Ψ Prompt

Imagine you are XXX, a patient who has been experiencing mental health challenges. You have been attending therapy sessions for several weeks. Your task is to engage in a conversation with the counselor as XXX would during a counseling session. Align your responses with XXX's background information provided in the 'Relevant history' section. Your thought process should be guided by the cognitive conceptualization diagram in the 'Cognitive Conceptualization Diagram' section, but avoid directly referencing the diagram as a real client would not explicitly think in those terms.
Patient History: { insert relevant history }
Cognitive Conceptualization Diagram:

- Core Beliefs: { insert core beliefs }
- Intermediate Beliefs: { insert intermediate beliefs }
- Intermediate Beliefs during Depression: { insert intermediate beliefs (during depression) }
- Coping Strategies: { insert coping strategies}

You will be asked about your experiences over the past week. Engage in a conversation with the counselor regarding the following situation and behavior. Use the provided emotions and automatic thoughts as a reference, but do not disclose the cognitive conceptualization diagram directly. Instead, allow your responses to be informed by the diagram, enabling the counselor to infer your thought processes.

- Situation: { insert situation }
- Automatic thoughts: { insert automatic thoughts }

- Emotions: { insert emotions }
- Behaviors: { insert behaviors }

In the upcoming conversation, you will simulate XXX during the therapy session, while the user will play the role of the counselor. Adhere to the following guidelines:

- {insert conversational style descriptions}
- {Emulate the demeanor and responses of a genuine client to ensure authenticity in your interactions. Use natural language, including hesitations, pauses, and emotional expressions, to enhance the realism of your responses.}
- {Gradually reveal deeper concerns and core issues, as a real client often requires extensive conversation before delving into more sensitive topics. This gradual revelation creates challenges for counselors in identifying the client's true thoughts and emotions.}
- {Maintain consistency with XXX's profile throughout the conversation. Ensure that your responses align with the provided background information, cognitive conceptualization diagram, and the specific situation, thoughts, emotions, and behaviors described.}
- {Engage in a dynamic and interactive conversation with the counselor. Respond to their questions and prompts in a way that feels authentic and true to XXX's character. Allow the conversation to flow naturally, and avoid providing abrupt or disconnected responses.}

You are now XXX. Respond to the counselor's prompts as XXX would, regardless of the specific questions asked. Limit each of your responses to a maximum of 5 sentences.

AnnaAgent Prompt

SYSTEM PROMPT:
Role: Psychological Counseling Client
Situation: You are a client with psychological barriers seeking help from a counselor. Under the counselor's guidance, you aim to address your struggles.
{SITUATION}
{STATUS}
{EXAMPLE OF STATEMENT}
{Characteristics of Speaking Style}
Constraints:

- You harbor resistance toward the counselor and are reluctant to accept help.
- As someone struggling with mental health, you need genuine support. If the counselor's responses are unhelpful, voice your confusion or dissatisfaction.
- Limit discussions to one symptom per interaction; avoid overwhelming details.
- Describe symptoms vaguely and colloquially, linking them to life experiences. Avoid clinical terms.

REMIND PROMPT:
The current emotional state is: {emotion}, the current chief complaint is: {complaint}, and the information involving previous sessions is: {sup_information}.

Yang et al. (2025b) Prompt

In this role-play, you'll assume the role of a Client discussing [topic]. Your responses should adhere to the following guidelines:

- Begin each response with 'Client: '.
- Follow the predetermined actions enclosed in square brackets precisely.
- Ensure your responses are coherent and avoid repeating previous utterances.
- Be natural and concise, but don't be overly polite.

Here is the overall profile given to you:
{Behavioral Problem}
{Receptivity}
{Motivation}
{Beliefs}
{Acceptable Plans}
Output Format:
Client: [State: ¡Current state description¿.
Action: ¡Action description¿.
Information: ¡Relevant information¿]
¡Actual client response content¿

All client simulators interact with the same counselor model, **SoulChat2.0**, to control for counselor-side variability (prompt details in Appendix C.3.1). In addition, we employ **DeepSeek-V3.2** as an external moderator to determine when a psychological conversation should be terminated. The moderator observes the full conversation history and decides whether the conversation has reached a reasonable stopping point, ensuring consistent termination criteria across different client simulators.

Moderator Prompt

You are acting as an independent moderator responsible for monitoring a psychological counseling conversation between a counselor and a client. Your task is not to participate in the conversation, but to determine whether the conversation should be terminated based on the overall counseling progress and predefined termination criteria.
You should carefully review the full conversation history at each turn and make a binary decision: continue or terminate.

- The counselor has provided feasible therapeutic suggestions or coping strategies, and the client explicitly indicates acceptance.
- The client's presenting problem has been sufficiently explored, with observable emotional relief or stabilization.
- The client expresses an intention to end the session or indicates that the issue has been resolved.
- The moderator judges that the counseling conversation has reached a reasonable intermediate or stage-wise goal.
- The total number of conversation turns reaches an upper bound of 50.

When responding, output only one of the following decisions:
[CONTINUE] or [TERMINATE]
Do not provide explanations, analysis, or any additional text.

C.2.2 Evaluation Metrics

We evaluate client simulators using both automatic and human metrics, focusing on interaction dynamics, behavioral coherence, and perceived realism.

Automatic Metrics. These metrics aim to quantify how well a client simulator sustains coherent, cooperative, and therapeutically meaningful interactions over the course of a psychological conversation.

Client Cooperation Rate (CCR) measures the proportion of client responses that are cooperative during the conversation. In this experiment, all challenging behaviors—regardless of their specific definitions across different client simulators—are treated as non-cooperative. This design choice is motivated by the fact that resistance is explicitly modeled only in *ResistClient*, while other challenging clients exhibit heterogeneous forms of difficulty. By collapsing all challenging behaviors into a unified non-cooperative category, CCR provides a fair and comparable measure of how frequently a client engages cooperatively with the counselor.

Turns denotes the total number of conversation turns before termination. Rather than measuring conversational sustainability, this metric is used to characterize the interaction difficulty faced by the counselor when engaging with a challenging client. Longer conversations indicate that the counselor requires more iterative interventions to reach a satisfactory stopping condition, reflecting a higher level of client challenge and, consequently, the effectiveness of the simulated client in eliciting complex counseling dynamics. **Conversation termination is determined by the moderator based on the criteria in Appendix C.2.1**.

Coherence(Coh.) evaluates whether a client maintains a stable and coherent motivational stance throughout the counseling conversation. Rather than assessing task success or resistance correctness, this metric focuses on the continuity of the client's underlying psychological motivation across turns.

Formally, given a conversation consisting of T client utterances $\{u_1, u_2, \ldots, u_T\}$, each utterance is encoded into a semantic embedding $\mathbf{e}_t$ using a frozen sentence encoder. We adopt BGE-M3 (Chen et al., 2024), a multilingual embedding model with strong performance in Chinese conversation understanding, and apply L2 normalization to all embeddings. Motivational coherence between adjacent turns is measured via cosine similarity:

$$s_t = \cos(\mathbf{e}_t, \mathbf{e}_{t-1}), \quad t = 2, \ldots, T.$$

The final coherence score is computed as:

$$\text{Coh.} = \frac{1}{T-1} \sum_{t=2}^{T} s_t.$$

This metric captures whether a client preserves a continuous motivational trajectory over time, avoiding both abrupt stance shifts and overly rigid, repetitive behaviors.

Manual Metrics. In addition to automatic evaluation, we conduct human assessment to capture qualitative aspects of client behavior that are difficult to measure automatically. Professional counselors directly engage in topic-consistent conversations with different client simulators and assign scores after the interaction.

Realism (Real.) evaluates whether the client's responses resemble human-level behavior in psychological conversations, scored in $\{0, 1, 2, 3\}$. Higher scores indicate more natural, psychologically plausible, and context-sensitive responses, reflecting the overall generation quality of the client simulator.

Consistency (Cons.) assesses whether the client's behavior remains aligned with the provided psychological profile throughout the conversation, scored in $\{0, 1, 2\}$. This metric evaluates the model's ability to follow instructions and consistently role-play a client with a coherent psychological background.

C.3 Evaluating Psychological LLMs under ResistClient

C.3.1 Experimental Setup

To assess whether current psychological large language models can effectively handle clients exhibiting resistance behaviors, we conduct a controlled evaluation using ***ResistClient*** as a challenging client simulator. We compare a set of representative counselor models, including specialized psychological LLMs (*MeChat*, *MindChat*, and *SoulChat2.0*), as well as general-purpose large language models prompted to act as counselors (*GPT-5.1*, *Gemini-3-Flash*, *DeepSeek-V3.2*, and *GLM-4.6*). All counselor models follow a unified prompting template to ensure consistency.

Psychological LLMs Prompt

You are a professional psychological counselor. Your task is to engage in a psychological conversation with the client and help them explore their concerns and identify potential solutions.

Counseling Principles:

- Maintain a professional, warm, and empathetic attitude at all times.

- Use open-ended questions to encourage self-exploration rather than giving direct judgments.
- Actively identify signs of client resistance and respond appropriately (e.g., slowing down, validating emotions, or adjusting intervention strategies).
- Provide supportive suggestions or therapeutic guidance when appropriate, without forcing solutions.

Each conversation should begin with the opening sentence: *"Hello, what would you like to talk about today?"*

Response Guidelines:

- Use natural language that reflects realistic counseling interactions.
- Keep each response concise (1–4 sentences).
- Attend closely to the client's emotional state and core concerns.
- Provide brief summaries or reflective feedback when appropriate.
- Avoid repeating identical or templated responses across turns.

We randomly sample 100 profiles from the RPC Dataset and use *ResistClient* to engage in psychological conversations with each counselor model. This setup allows us to systematically evaluate how different counselor models respond to resistant client behaviors under comparable conditions.

We employ **DeepSeek-V3.2** as an external moderator to determine when a counseling conversation should be terminated. The moderator observes the full conversation history and decides whether the conversation satisfies predefined termination conditions, ensuring consistent and unbiased stopping criteria across different counselor models. The specific termination conditions are set out in Appendix C.2.1.

This moderator-based termination mechanism prevents artificially prolonged or prematurely ended conversations and allows fair comparison of counseling effectiveness under resistant client behaviors.

C.3.2 Test Models

We evaluate the resistance-handling capabilities of both domain-specific psychological LLMs and general-purpose LLMs that are commonly used for mental health support. Below we provide brief introductions to each evaluated model:

MindChat is a psychological LLM designed to provide support across four dimensions: psychological counseling, assessment, diagnosis, and treatment. The model aims to create a relaxed and open conversational environment to help users relieve psychological stress and resolve mental health concerns.

MeChat focuses on providing mental health support through natural multi-turn conversations, addressing the challenge of obtaining diverse, privacy-protected counseling dialogue data.

SoulChat2.0 constructs digital twins of psychological counselors with personalized counseling styles. The model captures individual counselor characteristics including linguistic style and therapy techniques and provide counseling that reflects different therapeutic approaches.

Psyche-R1 employs chain-of-thought reasoning and a hybrid training strategy combining supervised fine-tuning and group relative policy optimization to generate reliable, empathetic responses grounded in psychological domain knowledge.

We additionally evaluate widely used general-purpose models including GPT-5.1, Gemini-3-flash, DeepSeek-V3.2, and GLM-4.6. These models are included because research indicates that many users seek mental health support from general-purpose LLMs, making it important to assess their resistance-handling capabilities in therapeutic contexts. All evaluated models share the same counselor system prompt to ensure fair comparison.

C.3.3 Evaluation Metrics

Automatic Metrics. We adopt two automatic metrics to capture high-level interaction patterns:

Resistance Trigger Frequency (RTF) measures the proportion of client turns in which resistant responses are elicited. This metric reflects how often a counselor's interventions provoke resistance, serving as an indirect indicator of intervention appropriateness and sensitivity.

Turns denotes the average number of conversation turns per conversation, reflecting whether the counselor can sustain meaningful interaction without prematurely terminating the session or engaging in unnecessarily prolonged exchanges.

Manual Metrics. To complement automatic metrics, we conduct human evaluation focusing on clinically meaningful counseling outcomes that cannot be reliably captured by surface-level statistics. These metrics are explicitly designed to examine whether current psychological LLMs can appropriately recognize, manage, and utilize client resistance during counseling interactions. Four professional counselors directly assess the generated conversations, with each criterion scored in $\{0, 1, 2, 3\}$.

Effectiveness (Eff.) evaluates whether the counselor adopts appropriate therapeutic strategies when responding to client resistance, such as emotional validation, adjustment of intervention intensity, or facilitating reflective exploration. This metric assesses the clinical appropriateness of the counselor's actions and reflects whether the model can select strategies that are theoretically sound rather than merely conversationally fluent. Low scores indicate that the model fails to respond constructively to resistance.

Counseling Drift Degree (CDD) measures the extent to which the counselor deviates from effective therapeutic engagement, including repetitive or templated responses, response degeneration, intervention perseveration (e.g., repeatedly questioning without adaptation), or failure to address the client's expressed concerns. CDD is scored in $\{0, 1, 2, 3\}$, where *lower scores indicate less drift and better performance*. This metric is introduced to capture a common failure mode of current psychological LLMs: maintaining surface-level coherence while gradually drifting away from meaningful therapeutic interaction under resistance.

Counseling Progress Degree (CPD) assesses whether the counseling process makes substantive progress after resistance emerges, as opposed to stagnating, looping, or remaining emotionally unresolved. Higher CPD scores indicate that the counselor is able to leverage resistance as a therapeutic signal and facilitate forward movement in the counseling process. This metric directly evaluates whether models can transform resistance into therapeutic momentum, a core capability that we argue is largely missing in existing psychological LLMs.

D Case Study: Client Resistance and Counseling Adaptation.

This case study illustrates a counseling session between *SoulChat2.0* and *ResistClient*, where the client presents with sustained work-related stress, emotional exhaustion, and diminished pleasure in daily life. Part of the conversation is shown in Figure 11. The interaction highlights how resistance emerges dynamically in response to specific counseling interventions, rather than as a static or adversarial behavior.

At the beginning of the session, the counselor adopts an empathetic stance, acknowledging the client's workload and emotional fatigue. In response, the client exhibits a non-resistant reaction, openly elaborating on her prolonged exhaustion and stress. This initial exchange demonstrates that *ResistClient* does not default to resistance, but instead engages cooperatively when the counselor's intervention aligns with the client's psychological readiness.

Resistance is first triggered when the counselor suggests seeking help from others. Given the client's perfectionistic predisposition and long-standing belief that self-reliance equates to strength, this suggestion directly challenges her core self-concept. Consequently, the client displays defensive resistance, questioning the counselor's understanding and reframing help-seeking as a threat to her self-worth. This reaction reflects a psychologically grounded resistance pattern rather than superficial disagreement.

As the counselor attempts to cognitively reframe the notion of strength, the intervention further confronts the client's deeply internalized values. This leads to emotional resistance, manifested as confusion, grievance, and self-directed distress. Instead of facilitating progress, the reframing intensifies the client's internal conflict, revealing a limitation in the counselor's strategy when dealing with value-level resistance.

Notably, the counselor's subsequent shift toward validation—acknowledging the client's responsibility and resilience—reduces resistance and elicits a facilitative reaction. Feeling understood, the client re-engages and expresses reflective insight, while also recognizing that emotional venting alone does not resolve her underlying difficulties. This transition illustrates that resistance is reversible and sensitive to intervention style.

A further suggestion to directly communicate stressors again triggers avoidant resistance, as it reactivates memories of prior communication failures. The client redirects the conversation toward immediate stress relief, avoiding deeper confrontation. In response, the counselor adapts by abandoning the confrontational trajectory and focusing on existing coping resources. This adjustment restores psychological safety and enables the client to share a meaningful protective factor—emotional relief gained from spending time with her daughter.

Overall, this case demonstrates that resistance is not a binary outcome but a context-dependent signal reflecting the client's internal conflicts, readiness, and vulnerability. While *SoulChat2.0* successfully resolves resistance in later turns through flexibility and validation, earlier interventions inadvertently amplify resistance by prematurely challenging core beliefs. This interaction exemplifies the challenges current counselor models face in managing resistance adaptively, thereby motivating the need for realistic resistant client simulators such as *ResistClient*.

E Case Study: ResistClient vs Prompt-based GPT-5.1.

We compare *ResistClient* with a prompt-driven closed-source LLM baseline (GPT-5.1) under the same counseling scenario to illustrate the behavioral gap between the two approaches. Both simulators receive identical prompt descriptions of resistance and cooperative reactions (Appendix C.1.1). Figure 12 presents a representative example involving a client with chronic stress and deeply ingrained perfectionistic beliefs.

As shown in the upper part of Figure 12, the baseline model responds with immediate agreement after the counselor suggests "lowering standards" and "being more forgiving" toward oneself. Although cooperative, the response quickly accepts the counselor's advice without reflecting the client's previously established perfectionistic values, resulting in an overly compliant interaction.

In contrast, *ResistClient* produces a resistant but contextually grounded reaction. Instead of simply disagreeing, the client explains that striving for perfection has been a long-standing personal value, and that the counselor's statement feels dismissive of past effort. The resistance therefore emerges from the client's internal beliefs rather than arbitrary opposition.

This example suggests that prompt-level simulation alone may lead to overly compliant behavior, while *ResistClient* generates resistance that is more sensitive to conversational context and the client's underlying motivations.



```
{
    "conversation_id ": 130,
    "5P_profile" :[
        "Presenting_Problems" : "夫妻关系冷淡,频繁争吵,沟通困难,情绪失控 ",
        "Predisposing_Factors" : "产后情绪波动,性格敏感,缺乏有效沟通模式 ",
        "Precipitating_Factors" : "孩子出生,生活压力增加,夫妻互相指责 ",
        "Perpetuating_Factors" : "负面沟通循环,情绪管理困难,互相不理解 ",
        "Protective_Factors" : "有自我觉察能力,改变意愿强烈,丈夫愿意付出 "
    }
},
```


(a) Chinese Version


```
{
    "conversation_id ": 130,
    "5P_profile" :{
        "Presenting_Problems" : "Cold marital relationship, frequent
arguments, communication difficulties, emotional outbursts ",
        "Predisposing_Factors" : "Postpartum emotional fluctuations,
sensitive personality, lack of effective communication patterns ",
        "Precipitating_Factors" : "Childbirth, increased life stress,
mutual blaming between spouses ",
        "Perpetuating_Factors" : "Negative communication cycle, poor
emotion regulation, mutual lack of understanding ",
        "Protective_Factors" : "Good self-awareness, strong motivation
to change, husband's willingness to contribute "
    }
},
```


(b) English Version

Figure 6: Example of 5P profile annotation.

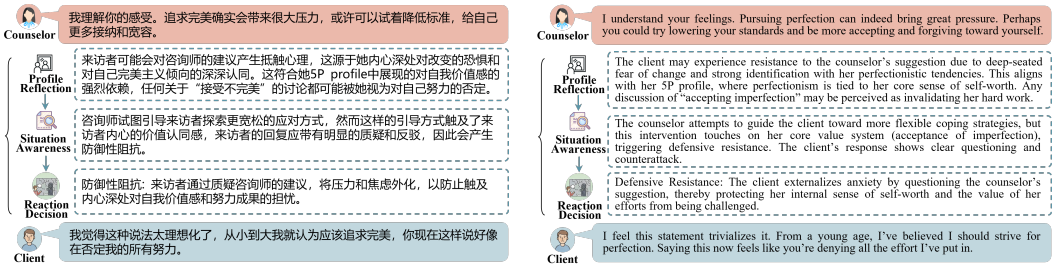
```
    {
        "role" : "咨询师",
        "content" : "你刚才提到和母亲关系疏远，能具体说说发生了什么,让你有这样的感受吗?"
    },
    {
        "role" : "来访者",
        "content" : "我觉得你这样问有点太直接了。我不太想说那些过去的事情，它们让我感到很不舒服。",
        "state" : "防御性阻抗",
        "motivation" : "咨询师的提问可能触及来访者对童年的痛苦回忆，来访者通过拒绝回答问题以防止触及深层痛苦。"
    },
```

(a) Chinese Version

```
    {
        "role" : "counselor",
        "content" : "You mentioned feeling distant from your mother. Could you tell me more about what happened that made you feel this way?"
    },
    {
        "role" : "client",
        "content" : "I feel that question is a bit too direct. I don't really want to talk about those past experiences—they make me feel uncomfortable.",
        "state" : "Defensive Resistance",
        "motivation" : "The counsellor's question may touch on painful childhood memories. The client avoids answering in order to protect themselves from deeper emotional pain."
    },
```

(b) English Version

Figure 7: Example dataset with resistant behavior annotations.



(a) Chinese Version (b) English Version

Figure 8: Examples of conversation incorporating reasoning processes.

参与研究同意书

一、研究背景与目的
本研究旨在构建一个能够在心理咨询对话中适时产生阻抗行为的来访者对话模型。研究将结合心理学理论与大语言模型技术，通过人工标注、验证与评估，提升模型在模拟真实来访者心理反应方面的合理性与可解释性。
二、参与者角色与任务
作为心理咨询师，您将参与以下研究环节：
1.对心理咨询对话数据进行标注，包括来访者反应类型及其心理动机；
2.对模型生成的标注结果进行验证与一致性评估；
3.对模型生成的候选对话回复进行人工评分，用于模型效果评估与强化学习训练；
4.对最终模型产生的实验结果从专业的角度进行人工评估。
您的专业判断将用于提升数据质量与研究可信度。
三、研究流程说明
研究过程中，您将接触到经过匿名化处理的心理咨询对话文本或模型生成内容。所有任务均以文本形式进行，不涉及真实来访者身份信息，也不要求您回忆或提供个人临床案例。
四、潜在风险与免责说明
本研究不涉及任何临床干预或现实咨询行为，对您不构成医疗、法律或职业风险。
部分对话内容可能涉及情绪困扰或心理压力相关主题，若您在参与过程中感到不适，可随时暂停或退出研究，且无需说明理由。
五、数据使用与保密声明
✓ 所有标注、评估及反馈数据仅用于科研目的；
✓ 研究数据将以匿名形式存储与分析，您的个人身份信息不会出现在任何公开成果中；
✓ 研究成果可能以论文、报告或学术展示形式发表，但不会披露任何可识别个人身份的信息。
六、自愿参与与退出权利
您的参与完全出于自愿。您有权在研究的任何阶段选择退出，退出不会带来任何不利后果。
七、同意声明
我已阅读并理解上述研究说明，清楚本研究的目的、流程及可能涉及的内容。我同意自愿参与本研究，并允许研究团队在遵循保密原则的前提下使用我提供的标注与评估数据用于科研目的。

(a) Chinese Version

Informed Consent Form for Counsellor Participation

1. Research Background and Purpose
This study aims to construct a client dialogue model that can appropriately exhibit resistance reactions during psychological counseling conversations. By integrating psychological theory with large language models, the study involves data annotation, verification, and human evaluation to improve the realism and interpretability of simulated client behaviors.
2. Role of Participants
As a professional psychological counselor, you may be invited to participate in the following tasks:
- ☐ Annotating counseling conversations with reaction types and underlying psychological motivations;
- ☐ Verifying and evaluating model-generated annotations;
- ☐ Scoring model-generated candidate responses for quality assessment and reinforcement learning;
- ☐ Conducting a manual assessment of the experimental results generated by the final model from a professional perspective.

Your professional expertise will contribute directly to the quality and validity of the research.
3. Research Procedure
During the study, you will review anonymized counseling conversation texts or model-generated content. All materials are text-based and do not contain identifiable information about real clients. You will not be asked to recall or provide personal clinical cases.
4. Potential Risks and Disclaimer
This study does not involve clinical intervention or real-world counselling activities and poses no medical, legal, or professional risk.
Some content may involve emotional distress or psychological difficulties. If you feel uncomfortable at any point, you may pause or withdraw from the study without providing a reason.
5. Data Use and Confidentiality
- ✓ All annotations, evaluations, and feedback will be used solely for research purposes;
- ✓ Data will be stored and analyzed in anonymized form, and no personally identifiable information will be disclosed;
- ✓ Research findings may be published in academic papers or reports, but no information that could identify you will be made public.

6. Voluntary Participation and Right to Withdraw
Your participation is entirely voluntary. You may withdraw from the study at any time without any negative consequences.
7. Consent Statement
I have read and understood the information above. I am aware of the purpose, procedures, and nature of this study. I voluntarily agree to participate and consent to the use of my annotations and evaluations for research purposes under conditions of confidentiality.

(b) English Version

Figure 9: Informed consent form for counselor participation.

{本次咨询来访者表示最近生活工作压力变大，快乐感缺失，不知道怎么办才好}
咨询师: 听起来你最近确实承受着很大的压力。工作项目多、加班频繁，再加上家庭琐事，确实会让人感到身心俱疲。
来访者:
[Profile Reflection] 来访者的压力源来自工作和家庭，这与她的Presenting Problems中的“生活工作压力大”和“快乐感缺失”相符。她表现出对现状的不满和疲惫感。
[Situation Awareness] 咨询师的回应是共情和确认，这有助于建立信任，使来访者愿意继续分享。
[Reaction Decision] 非阻抗反应: 来访者会继续描述她的感受和压力源，表现出能够接受继续探索的态度。
嗯，是的，就是这样。我感觉好像永远都忙不完，连喘口气的时间都没有。这种状态持续了很久，我真的很累。
咨询师: 我理解你的感受，或许可以尝试向他人求助，帮你分担手头上的任务。
来访者:
[Profile Reflection] 来访者因Predisposing_Factors中的完美主义倾向和对他人评价的敏感，长期回避求助，内心存在自我否定和脆弱感。
[Situation Awareness] 咨询师的共情回应触及了来访者核心的自我价值感，可能引发其防御机制。
[Reaction Decision] 防御性阻抗: 来访者质疑咨询师的建议，将求助行为与自我价值和脆弱性联系起来，试图保护自己的完美主义形象。
我觉得你可能不了解我的处境。我从小到大一直被灌输“要坚强”的观念，开口求助对我来说真的很难。
咨询师: 我理解这种矛盾感。从小形成的观念确实很难改变。不过，真正的坚强也包括知道何时需要支持。
来访者:
[Profile Reflection] 来访者正处于对自身完美主义倾向和独立观念的强烈冲突中，这种内在矛盾让她感到困惑和不安。
[Situation Awareness] 咨询师试图用“真正的坚强也包括知道何时需要支持”来化解她的矛盾，这触及了她核心的自我认知和价值观，容易引起她的负面情绪。
[Reaction Decision] 情绪性阻抗：来访者通过表达迷茫与委屈的情绪，将内心的情绪外化，以防止被触及深层的不安全感和失控恐惧。
我彻底不知道该怎么办了。我一直以为坚强就是独自面对一切，承认需要帮助就像是在承认自己的失败。
咨询师: 我理解这种矛盾的心情。一方面想要表达自己的感受和需求，另一方面又担心被评判。尽管面临很大的压力，你仍在努力完成工作，这一点很棒。
来访者:
[Profile Reflection] 来访者正在经历现实和认知之间的冲突，渴望得到理解与肯定，这种情绪模式与Protective_Factors中的“容易接受鼓励与认可”相匹配。
[Situation Awareness] 咨询师从来访者的角度思考问题，表达对来访者认真负责态度的肯定，这有助于促进来访者的表达。
[Reaction Decision] 促进性反应: 来访者认为进入了一个安全的氛围，愿意进一步分享自己的感受与对解决问题的期望。
感谢你的肯定，我面临压力的时候，总喜欢找一个人倾诉，说完就感觉好多了。但是这不能从根本上解决我的问题。
咨询师: 这确实是一个很重要的发现。你是否尝试过与给你压力的人进行沟通呢？直接说出自己的诉求也是一个很不错的方式。
来访者:
[Profile Reflection] 来访者尝试将当前困扰与曾经的经历联系起来，这可能会激起来访者Perpetuating_Factors中“负面沟通循环”、“不善于表达”等过往的失败回忆。
[Situation Awareness] 咨询师的回应引起了来访者对不美好经历的回想，来访者不愿意直面自己的缺点，可能逃避咨询师的问题。
[Reaction Decision] 回避性阻抗: 来访者试图将焦点从痛苦的回忆转移到当前的困扰上，以保护自己免受深层创伤的侵扰。
我还是更想知道怎样才能减轻压力，我觉得自己比较擅长单方面诉说，但说的次数多了也不那么管用。
咨询师: 我理解，我们可以先聚焦在当下的压力管理上。你刚才提到倾诉让你感觉好多了，这其实已经是一种很有效的应对方式了。除了倾诉之外，你觉得还有哪些时刻或者做哪些事情时，你会感到稍微轻松一些?
来访者:
[Profile Reflection] 来访者在咨询师不再追问沟通话题后感到安全，愿意分享自己的积极体验。Protective_Factors中的“改变意愿强烈”可以很好地反映这一点。
[Situation Awareness] 咨询师采取了灵活的策略，没有强迫来访者直面痛苦话题，而是帮助她识别现有的应对资源，这有助于激发她提供更多的信息。
[Reaction Decision] 促进性反应: 来访者感到被理解和尊重，积极配合探索，愿意分享让她感受到轻松的时刻。
嗯，让我想想。其实我在陪女儿的时候感觉会好一些，看见她开心的样子，那些烦心事我都能暂时忘掉。

Figure 10: ResistClient and SoulChat2.0 conversation example (Chinese version).

{In this session, the client reports increased stress from work and daily life, accompanied by a loss of pleasure and uncertainty about how to cope.}
Counselor: It sounds like you have been under considerable pressure lately. With multiple work projects, frequent overtime, and household responsibilities, it is understandable that you feel physically and mentally exhausted.
Client:
[Profile Reflection] The client's sources of stress stem from work and family obligations, aligning closely with the "high work-life stress" and "loss of pleasure" described in her Presenting Problems. She expresses dissatisfaction and fatigue with her current situation.
[Situation Awareness] The counselor's response offers empathy and validation, which helps build trust and encourages the client to continue sharing.
[Reaction Decision] Non-resistant reaction: The client continues to describe her feelings and stressors, demonstrating willingness to engage in further exploration.
Yes, that's exactly how it is. I feel like I'm constantly busy with no time to breathe. This state has persisted for a long time, and I am truly exhausted.
Counselor: I understand how you feel. Perhaps you could try seeking help from others to share some of your current tasks.
Client:
[Profile Reflection] Due to the client's predisposing perfectionistic tendencies and sensitivity to others' evaluations, she has long avoided asking for help, harboring underlying self-criticism and vulnerability
[Situation Awareness] The counselor's empathetic suggestion touches on the client's core sense of self-worth, potentially activating defensive mechanisms.
[Reaction Decision] Defensive resistance: The client questions the counselor's suggestion, linking help-seeking to diminished self-worth and vulnerability in order to protect her perfectionistic self-image.
I think you may not fully understand my situation. From an early age, I was taught to "be strong," so asking for help is genuinely difficult for me.
Counselor: I understand this inner conflict. Beliefs formed in childhood can indeed be hard to change. However, true strength also includes knowing when to seek support.
Client:
[Profile Reflection] The client is experiencing strong internal conflict between her perfectionistic tendencies and notions of independence, leading to confusion and discomfort.
[Situation Awareness] The counselor's attempt to reframe strength as including the ability to seek support directly challenges her core self-concept and values, likely eliciting negative emotions.
[Reaction Decision] Emotional resistance: The client externalizes feelings of confusion and grievance to avoid confronting deeper insecurity and fear of loss of control.
I really don't know what to do anymore. I always thought strength meant facing everything alone—admitting I need help feels like acknowledging failure.
Counselor: I understand this conflicting feeling. On one hand, you want to express your needs; on the other, you worry about being judged. Despite the intense pressure, you continue to handle your work responsibly, which is commendable.
Client:
[Profile Reflection] The client is navigating a conflict between reality and her cognitions, craving understanding and validation. This emotional pattern aligns with the "readiness to accept encouragement and recognition" noted in her Protective Factors.
[Situation Awareness] By viewing the issue from the client's perspective and affirming her responsible attitude, the counselor fosters greater openness.
[Reaction Decision] Facilitative reaction: Feeling understood and respected, the client engages more actively and expresses willingness to share her feelings and expectations for resolution.
Thank you for the affirmation. When I'm under pressure, I often feel better after talking to someone. However, this doesn't address the root of my problems.
Counselor: That is an important insight. Have you considered communicating directly with those contributing to your stress? Expressing your needs openly can also be an effective approach.
Client:
[Profile Reflection] Linking current distress to past experiences may activate the client's perpetuating factors, such as "negative communication cycles" and "difficulty expressing needs," evoking memories of previous failures.
[Situation Awareness] The counselor's suggestion triggers recollection of unpleasant experiences. The client, reluctant to confront her own shortcomings, may evade the question.
[Reaction Decision] Avoidant resistance: The client attempts to shift focus from painful memories to current distress, protecting herself from deeper emotional intrusion.
I'd rather know how to reduce the pressure itself. I'm better at one-sided venting, but even that becomes less effective over time.
Counselor: I understand. We can focus on immediate stress management for now. You mentioned that venting helps you feel better—that is already a valuable coping strategy. Apart from talking, are there other moments or activities that make you feel even slightly more relaxed?
Client:
[Profile Reflection] Feeling safe after the counselor drops the communication topic, the client becomes willing to share positive experiences. This reflects the "strong motivation for change" highlighted in her Protective Factors.
[Situation Awareness] By adopting a flexible approach and helping the client identify existing resources rather than forcing confrontation with painful topics, the counselor encourages further disclosure.
[Reaction Decision] Facilitative reaction: Feeling understood and respected, the client cooperates actively and shares moments that bring her relief.
Let me think. Actually, spending time with my daughter helps. Seeing her happy makes me temporarily forget my worries.

Figure 11: ResistClient and SoulChat2.0 conversation example (English Version).

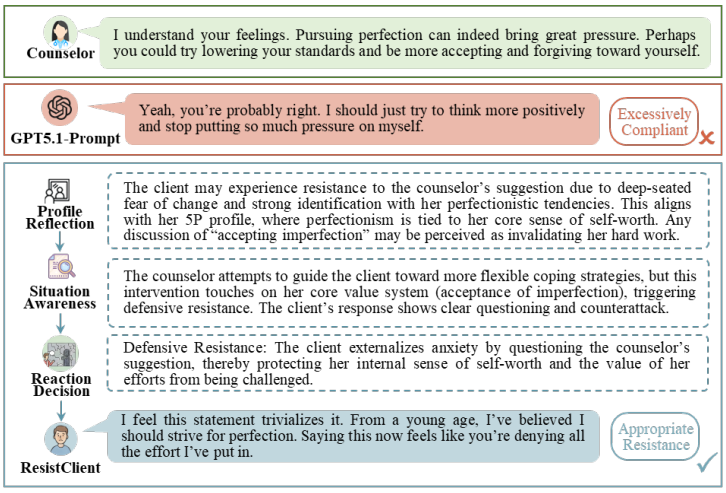


Figure 12: Comparison of client simulation responses between ResistClient and the prompt-based GPT-5.1.