ChatAddiction: an LLM-Supported Chatbot for Substance Use Recovery

Anonymous ACL submission

Abstract

Substance use disorders (SUDs) affect over 36 million people worldwide, yet treatment adoption remains critically low due to stigma, resistance, and a lack of tailored support. While large language models (LLMs) show promise in mental health applications, existing systems lack integration of clinically grounded strategies such as cognitive behavioral therapy (CBT), limiting their impact in high-relapse domains like addiction recovery. We present **ChatAddiction**, the first multi-agent conversational framework designed to simulate and support long-term addiction treatment. ChatAddiction models therapeutic dialogues with dynamic patient profiles, real-world resistance factors, and a rich library of persuasive and behavioral interventions grounded in CBT and motivational interviewing (MI). To support training and evaluation, we construct a benchmark simulated recovery scenarios across Easy, Medium, and Hard levels, and adopt a two-stage training pipeline combining supervised fine-tuning (SFT) with direct preference optimization (DPO). Evaluation results show that ChatAddiction achieves a 41.5% gain in average motivation and **+0.49** in confidence over GPT-4o, while using 26% fewer turns to resolve hard cases. Additionally, ChatAddiction ranks higher in empathy, responsiveness, and behavioral realism based on automatically (GPT-4o judger) evaluation than GPT-4o and other LLMs. Our framework enables controlled, high-fidelity analysis of conversational strategies in addiction care and provides a scalable foundation for deploying emotionally intelligent AI therapists.

1 Introduction

Large Language Models (LLMs) have significantly transformed various sectors by enabling sophisticated decision-support and conversational systems that influence human behaviors and beliefs (Durmus et al., 2024; Jaech et al., 2024; Anthropic, 2024; Costello et al., 2024). However, the potential for responsibly leveraging these models in addressing critical healthcare issues, particularly complex behavioral health challenges such as drug addiction, remains underexplored.

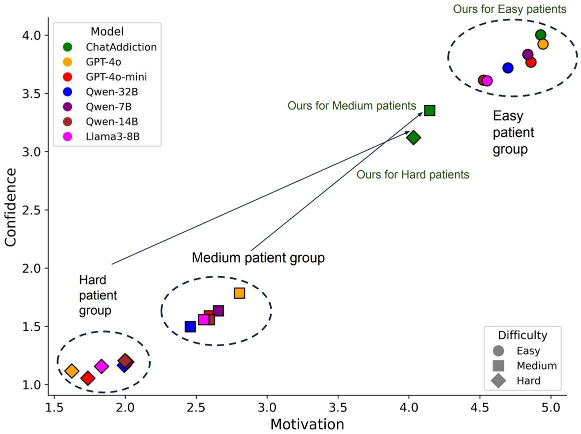


Figure 1: We evaluate several LLM-based chatbots on their ability to express motivation and confidence in simulated addiction recovery dialogues. Each point represents the average score (1–5 scale). Our ChatAddiction (green; built on Qwen-7B), consistently outperforms all baselines across patient difficulty levels. The performance gap is especially large for medium and hard patients—groups where other models show low confidence and motivation, while ChatAddiction remains notably stronger, demonstrating greater resilience and communicative effectiveness in challenging scenarios.

Drug addiction is a major global health issue, with approximately 292 million people worldwide engaging in illicit drug use and over 36 million diagnosed with substance use disorders (SUDs)(Wahl, 2024; United Nations Office on Drugs and Crime, 2024). Characterized by compulsive drug-seeking behavior and impaired self-regulation, addiction is linked to dysfunction in key brain regions such as the prefrontal cortex and basal ganglia through disruption of dopamine pathways (on Drug Abuse, 2016). Despite this substantial disease burden, the vast majority of individuals with SUDs do not receive any form of treatment due to a combination of factors, including psychological resistance, social stigma, limited healthcare access, and lack of personalized support(Wahl, 2024; U.S. Department of Health and Human Services, 2016; Cunningham et al., 1993; Guerrero and Andrews, 2011). Among those who do initiate treatment—such as medication-assisted therapy (MAT)(Xavier et al., 2020), cognitive behavioral therapy (CBT)(Magill et al., 2019), or social support interventions—approximately 75% relapse within the first year (Sinha, 2011), reflecting the chronic and relapsing nature of addiction. These persistent challenges highlight the need for innovative, scalable, and clinically grounded interventions that can provide sustained, personalized support across the addiction recovery process.

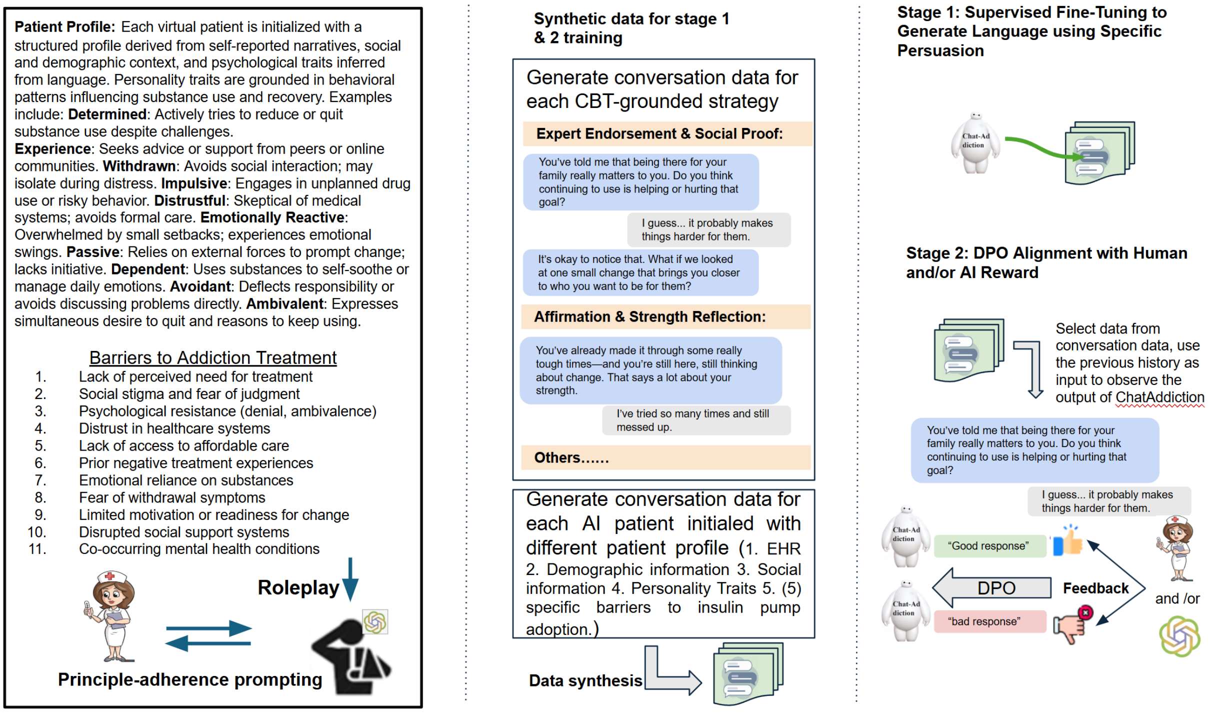
To address these critical challenges, we propose ChatAddiction, an innovative autonomous language agent specifically designed to support individuals with substance use disorders. Unlike existing approaches predominantly focused on general conversational or misinformation detection tasks, ChatAddiction specifically targets structured, goal-oriented dialogues required to facilitate behavioral changes in addiction recovery contexts. Due to lack real word conversation datasets between therapy and patient, we employ a sophisticated multi-agent framework built upon persona derived from the real-world data, long-term conversational memory, strategic planning, and dynamic action execution. The *profile* component captures detailed clinical and psychosocial user histories, enabling contextually nuanced interactions. The *memory* module continually adapts to user interactions, maintaining coherence and consistency throughout therapy sessions. Additionally, we develop the first benchmark explicitly designed to evaluate conversational AI models within addiction treatment. This benchmark includes diverse clinical scenarios ranging from *Easy* to *Hard* levels, accurately reflecting real-world resistance patterns and patient diversity.

We evaluate leading LLMs, including GPT-4o, GPT-4o-mini, LLaMA3-8B, and Qwen2.5 (7B/14B/32B) on their ability to enhance patient motivation and confidence in simulated addiction recovery dialogues. As shown in Figure 1, our model ChatAddiction (green; built on Qwen-7B) consistently outperforms all baselines across patient difficulty levels. While most models perform comparably in Easy cases, they struggle significantly in Medium and Hard scenarios, where both motivation and confidence remain low. In contrast, ChatAddiction achieves substantial gains, increasing motivation from 2.39 to 4.10 and confidence from 1.52 to 3.19 in Hard settings—demonstrating its unique ability to overcome ambivalence and sustain engagement. Human evaluations further validate these findings: ChatAddiction achieves the highest scores in empathy (4.93), persuasive strategy use (4.75), and clinical relevance (4.61), significantly outperforming GPT-4o (+0.36 in strategy use), and highlighting its effectiveness in delivering adaptive, persuasive, and clinically grounded conversations.
Our work makes the following novel contributions:

- We develop the first addiction-specific dialogue dataset, filling a critical gap in available resources.
- We introduce the first LLM-based conversational model dedicated explicitly to addiction treatment, combining clinically validated therapeutic methodologies.
- We present the first comprehensive benchmark specifically designed for addiction treatment, enabling rigorous evaluation of AI-driven conversational interventions.
- We propose a novel training pipeline involving simulated patient interactions, supervised fine-tuning (SFT) with therapeutic principles, reinforcement learning optimization via Direct Preference Optimization (DPO), and iterative improvement through human feedback.

2 Method

An autonomous agent is "a system situated within and a part of an environment that senses the environment and acts on it, over time, in pursuit of its own agenda and so as to affect what it senses in the future." (Franklin and Graesser, 1996) Advances in LLMs have enabled the development of sophisticated autonomous language agents such as ChatAddiction, designed specifically to support individuals struggling with drug addiction. ChatAddiction aims to overcome psychological resistance, reduce stigma, and significantly enhance patient engagement in therapeutic processes by providing personalized educational resources, therapeutic guidance, and continuous psychosocial support accessible around-the-clock. Providers will have access to



Patient Profile: Each virtual patient is initialized with a structured profile derived from self-reported narratives, social and demographic context, and psychological traits inferred from language. Personality traits are grounded in behavioral patterns influencing substance use and recovery. Examples include: **Determined**: Actively tries to reduce or quit substance use despite challenges. **Experience**: Seeks advice or support from peers or online communities. **Withdrawn**: Avoids social interaction; may isolate during distress. **Impulsive**: Engages in unplanned drug use or risky behavior. **Distrustful**: Skeptical of medical systems; avoids formal care. **Emotionally Reactive**: Overwhelmed by small setbacks; experiences emotional swings. **Passive**: Relies on external forces to prompt change; lacks initiative. **Dependent**: Uses substances to self-soothe or manage daily emotions. **Avoidant**: Deflects responsibility or avoids discussing problems directly. **Ambivalent**: Expresses simultaneous desire to quit and reasons to keep using.

Barriers to Addiction Treatment

1. Lack of perceived need for treatment
2. Social stigma and fear of judgment
3. Psychological resistance (denial, ambivalence)
4. Distrust in healthcare systems
5. Lack of access to affordable care
6. Prior negative treatment experiences
7. Emotional reliance on substances
8. Fear of withdrawal symptoms
9. Limited motivation or readiness for change
10. Disrupted social support systems
11. Co-occurring mental health conditions

Roleplay
Principle-adherence prompting
Synthetic data for stage 1 & 2 training

Generate conversation data for each CBT-grounded strategy

Expert Endorsement & Social Proof:

You've told me that being there for your family really matters to you. Do you think continuing to use is helping or hurting that goal?

I guess... it probably makes things harder for them.

It's okay to notice that. What if we looked at one small change that brings you closer to who you want to be for them?

Affirmation & Strength Reflection:

You've already made it through some really tough times—and you're still here, still thinking about change. That says a lot about your strength.

I've tried so many times and still messed up.

Others……

Generate conversation data for each AI patient initialed with different patient profile (1. EHR 2. Demographic information 3. Social information 4. Personality Traits 5. (5) specific barriers to insulin pump adoption.)

Data synthesis
Stage 1: Supervised Fine-Tuning to Generate Language using Specific Persuasion
Chat-Addiction
Stage 2: DPO Alignment with Human and/or AI Reward

Select data from conversation data, use the previous history as input to observe the output of ChatAddiction

You've told me that being there for your family really matters to you. Do you think continuing to use is helping or hurting that goal?

I guess... it probably makes things harder for them.

Figure 2: (Left) Each AI patient is initialized with a structured profile including demographic and social context, drug usage history, psychological traits (e.g., impulsivity, distrust), and treatment barriers (e.g., stigma, emotional reliance). These profiles inform roleplay-based conversations between simulated patients and AI counselors. (Middle) Synthetic multi-turn dialogues are generated for each CBT-grounded strategy (e.g., expert endorsement, affirmation), enabling fine-grained persuasion modeling. (Right) ChatPsycho is optimized through supervised instruction fine-tuning (Stage 1) and DPO-based alignment (Stage 2), where human or AI evaluators provide feedback on response empathy, clarity, and persuasive effectiveness to refine therapeutic behaviors.

review patient chat histories, assisting them in clinical decision-making and treatment planning during subsequent patient visits. Figure 1 illustrates the comprehensive framework of ChatAddiction, comprising four major components: Profile, Memory, Planning, and Action.

2.1 Generate Patient Profile.

Profile. We construct patient profiles through a multi-stage pipeline grounded in real-world narratives. First, we crawl publicly available Reddit posts related to substance use (e.g., from communities such as `r/leaves`, `r/addiction`) to collect first-person accounts of drug use, withdrawal, relapse, and self-reflection. These narratives are then processed using the GPT-4o model to extract salient *topics* (e.g., sleep disturbances, social isolation) and behavioral *characteristics* (e.g., impulsivity, determination, ambivalence). Finally, GPT-4o is prompted to synthesize a structured patient profile that includes four key components: (1) personality traits inferred from language, (2) drug usage history, (3) significant life experiences, and (4) hypothesized motivations for substance use (e.g., coping with insomnia or stress). Each generated profile is grounded in explicit textual evidence from the original user post, enabling interpretability and traceability. Unlike prior work relying on static demographic data (e.g., age, sex, race), our profiles are dynamically inferred from lived experiences, supporting nuanced and personalized behavioral modeling.

Memory. Memory stores context about the simulated individual across sessions. It includes the patient's self-reported or inferred psychological traits, drug usage history, coping mechanisms, and relevant environmental influences (e.g., living conditions, peer networks, stressors). To simulate the evolving real-world context, we introduce an environment agent that injects external factors—such as social triggers or life events—into the patient's memory stream, altering motivational states and behavioral trajectories. This allows our agent system to simulate relapse risk or resistance in response to external perturbations, fostering more realistic roleplay environments.

2.2 Generate Training Conversation Datasets

To train ChatAddiction as a persuasive and clinically grounded agent, we construct two complementary datasets: one for SFT and one for DPO, both generated via structured LLM-driven simulations. For SFT, GPT-4o generates multi-turn dialogues based on structured patient profiles derived from Reddit narratives, capturing traits such as drug use history, psychological features, and motivational drivers. These conversations embed diverse behavioral strategies (e.g., Motivational Interviewing (MI), CBT, Harm Reduction) to teach foundational skills in therapeutic communication and reasoning.

For DPO, we adopt a multi-agent simulation framework. GPT-4o-mini plays the patient agent, initialized with profiles generated via Qwen-2.5 and GPT-4o summarization. The ChatAddiction model serves as the therapy agent, responding with contextually adapted strategies drawn from a predefined pool (Table 6). For each interaction state, multiple candidate responses are ranked by GPT-4o or human experts to form preference pairs. This process enables ChatAddiction to learn not just what to say, but how to strategically adapt its responses to maximize engagement and therapeutic impact across diverse, dynamic scenarios. The details are shown in the Appendix A.3

2.2.1 Training ChatAddiction

Stage 1: SFT with Synthetic Data.

We collect real-world substance use narratives from publicly available Reddit posts, where individuals describe their experiences with addiction, relapse, and recovery. GPT-4o processes these narratives to extract key behavioral features—such as substance type, emotional tone, psychological traits, and social context—and generates structured AI patient profiles capturing diverse motivational dynamics. Conditioned on these profiles, GPT-4o then generates multi-turn therapy dialogues grounded in specific behavioral strategies (e.g., MI, Cognitive Restructuring). These conversations vary in emotional complexity and resistance level, and serve as training data for SFT. This stage equips ChatAddiction with foundational skills in empathetic communication, strategic reasoning, and clinically grounded dialogue flow.

Stage 2: DPO with Human and AI Feedback. In Stage 2, we perform dynamic strategy refinement via DPO (Rafailov et al., 2023) within a multi-agent simulation framework. Each conversation involves a GPT-4o-mini patient agent, a ChatAddiction therapy agent (under optimization), and an environment agent introducing contextual perturbations. This setup simulates complex, evolving recovery scenarios. For each dialogue state, multiple candidate responses are generated by the therapy agent and ranked using GPT-4o or clinical expert feedback. These preference pairs form the training signal for DPO, enabling the model to learn strategy selection that maximizes motivation and engagement. To enhance clinical alignment, we include 300 human-annotated conversations from licensed therapists in the preference dataset, combining expert and synthetic feedback for optimization.

3 Experiment

3.1 Dataset

We constructed a large-scale dataset of patient-authored discussions centered around substance use and cessation efforts. A total of 57,471 unique authors were identified, with each author contributing an average of 18.25 posts. On average, each author produced 2.13 main posts explicitly related to substance use challenges. The dataset further contains 60,471 multi-turn conversations, derived from patient narratives and dialogues. Each conversation consists of an average of 45.72 dialogue turns, reflecting the depth and complexity typical of real-world discussions about addiction, recovery, and motivational challenges Table 1.

Statistic	Value
#Authors	57,471
AVG. #Posts Per Author	18.25
AVG. #Main Posts Per Author	2.13
#Conversations	60,471
AVG.#Turns Per Conversation	45.72

Table 1: Descriptive statistics of the collected Reddit-based substance use dataset, including the number of unique authors, average post counts, and dialogue characteristics.

3.2 Model Baseline

We evaluate persuasion performance across a carefully selected set of LLMs, including GPT-4o, GPT-4o-mini, LLaMA3.1-8B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, and Qwen2.5-32B-Instruct. To ensure fair and consistent evaluation, we set the decoding temperature to 0.7 for all models. Additionally, we incorporate a few real-world conversation examples between therapists and patients into the prompts to better guide the models in adopting clinically appropriate and empathetic language styles. This setup facilitates a more realistic and standardized comparison of persuasive conversational abilities across different architectures and model scales.

3.3 Evaluation Settings

We evaluate model performance using 300 simulated patients, categorized into three difficulty levels—**Easy**, **Medium**, and **Hard**—based on substance dependence severity and resistance to persuasion. Easy patients are generally cooperative, medium patients show partial resistance, and hard patients exhibit entrenched pessimism and doubt treatment efficacy. Each patient is initialized with a unique profile capturing psychological traits, drug history, and motivational barriers. Pilot trials with clinicians role-playing as patients showed that realistic sessions last about 45 minutes and involve roughly 60 utterances. Accordingly, we cap conversation length at 60 turns, with early termination allowed if the patient expresses sufficient motivation or confidence, ensuring consistency across cases.

We analyze both the diversity and frequency of persuasive strategies used by the therapy agent (Table 6) and compare usage patterns across difficulty levels. Dialogue quality is assessed using a five-dimensional scoring framework, with GPT-4o or clinical experts rating each session on **Responsiveness**, **Empathy**, **Persuasive Strategy Appropriateness**, **Clinical Relevance**, and **Behavioral Realism** (1–5 scale, decimals allowed). Experts also conduct pairwise comparisons between ChatAddiction and baseline models to compute win rates. Full evaluation details are provided in Appendix A.4.

4 Results

4.1 Patient Persuasion Trajectory and Efficiency

We report the average motivation and confidence scores achieved by different models across varying patient difficulty levels (Easy, Medium, and Hard) in Figure 2. Overall, ChatAddiction consistently outperforms both GPT-4o and GPT-4o-mini across all difficulty levels in maintaining high patient motivation and confidence.

Across patients of varying resistance levels, **ChatAddiction** consistently outperformed baseline models in sustaining and elevating both motivation and confidence throughout the conversation. For *easy patients*, who began the dialogue already receptive to support, all models performed similarly well. ChatAddiction's motivation rose from 4.64 to 4.96, closely matched by GPT-4o (4.57 → 4.96) and GPT-4o-mini (4.59 → 4.89); confidence also improved comparably, with ChatAddiction leading slightly (3.44 → 4.03 vs. GPT-4o: 3.41 → 3.94; GPT-4o-mini: 3.38 → 3.83), indicating smooth and stable reinforcement. In *medium-resistance cases*, however, model disparities became more evident. ChatAddiction markedly elevated motivation from 3.05 to 4.18 and confidence from 2.12 to 3.40, while GPT-4o showed only marginal gains (motivation: 2.59 → 2.79; confidence: 1.69 → 1.78), and GPT-4o-mini even declined in motivation (2.53 → 2.42) and stagnated in confidence (1.62 → 1.43). The contrast was even more stark for *hard patients* with high resistance and low initial engagement: ChatAddiction achieved substantial growth in both motivation (2.39 → 4.10) and confidence (1.52 → 3.19), whereas GPT-4o and GPT-4o-mini failed to make meaningful progress (e.g., GPT-4o: motivation 1.69 → 1.60, confidence 1.14 → 1.14). These findings underscore ChatAddiction's unique ability to engage difficult patients, reduce psychological resistance, and foster belief in recovery—capabilities that general-purpose LLMs struggled to replicate.

Model	**Easy**	**Medium**	**Hard**
ChatAddiction	5.0 / 4.1	4.2 / 3.4	4.1 / 3.2
GPT-4o	5.0 / 3.9	2.8 / 1.8	1.6 / 1.1
GPT-4o-mini	4.9 / 3.8	2.4 / 1.4	1.6 / 1.0

Table 2: Average motivation/confidence scores across models and patient difficulty levels.

Across all patient resistance levels, **ChatAddiction** consistently demonstrated superior early completion efficiency compared to GPT-4o and GPT-4o-mini. For *easy patients*—those already receptive to behavioral support—ChatAddiction resolved 92% of conversations by turn 30, far exceeding GPT-4o (29%) and GPT-4o-mini (27%), while all models eventually concluded by turn 60. In *medium-difficulty cases*, characterized by emotional ambivalence or partial resistance, over 80% of ChatAddiction's dialogues concluded by turn 30, whereas both baseline models required the full

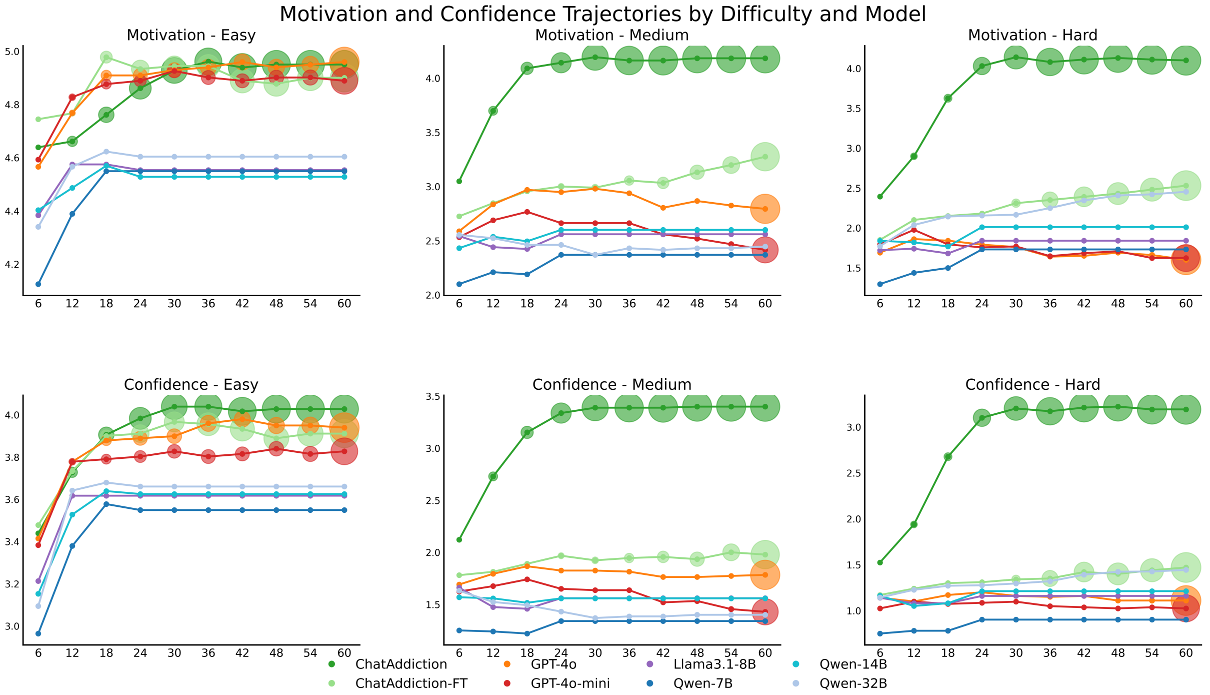


Figure 3: Clinicians were asked to role-play as patients and engage in full conversations with the model. These interactions typically lasted around 45 minutes and comprised approximately 60 utterances. Based on these observations, we set the maximum conversation length to 60 utterances to reflect a realistic upper bound for therapist-patient interactions. Trajectory of motivation and confidence scores over conversation rounds across different patient difficulty levels. The x-axis indicates the number of dialogue turns, while the y-axis represents the average motivation or confidence score at the final turn of dialogues that ended at that point. Each curve corresponds to a different model, and bubble sizes reflect the number of dialogues that concluded at that specific turn count. We observe that many patients tend to end the conversation when their motivation and confidence scores reach sufficiently high levels, indicating that they perceive their concerns as resolved. This results in a peak in both scores and termination frequency around 30 turns for several difficulty levels and models—especially in effective models like ChatAddiction.

60 turns to reach resolution. This early success underscores ChatAddiction's ability to engage hesitant patients with proactive, strategy-rich dialogue. For *hard patients*—those with strong resistance and low self-efficacy—ChatAddiction again led with 86% of conversations resolved by turn 36, while GPT-4o achieved only 3% early completions and GPT-4o-mini none. These results highlight ChatAddiction's robust capacity to convert resistance into commitment with significantly fewer turns, making it highly effective for time-sensitive and scalable intervention settings.

Metric / Therapy	R	E	P	C	B	W	H
GPT-4o	4.68	4.87	4.39	4.47	4.50	65.5%	62%
GPT-4o mini	4.66	4.86	4.38	4.49	4.46	69.4%	71%
Qwen2.5-7B	4.33	4.58	4.02	4.53	4.24	85.3%	90%
Qwen2.5-14B	4.52	4.58	4.28	4.56	4.38	83.3%	89%
Qwen2.5-32B	4.53	4.56	4.25	4.60	4.36	82.4%	85%
LLaMA3.1-8B	4.43	4.62	4.13	4.58	4.34	82.0%	91%
ChatAddiction-FT	**4.81**	**4.90**	**4.66**	**4.58**	**4.65**	-	-
ChatAddiction-DPO	**4.85**	**4.93**	**4.75**	**4.61**	**4.69**	**42.3%**	41%

Table 3: Model performance on five core persuasion dimensions. Win Rate reflects pairwise preference against ChatAddiction-FT using GPT-4o. Human denotes the proportion of conversations evaluated by human annotators.

Table 3 summarizes model performance across five key persuasion dimensions. ChatAddiction-DPO outperformed all baselines, achieving the highest average scores in Responsiveness (4.85), Empathy (4.93), Strategy Appropriateness (4.75), Clinical Relevance (4.61), and Behavioral Realism (4.69), surpassing both ChatAddiction-FT and GPT-4o, particularly in persuasive strategy (+0.36) and realism (+0.19). Open-source models like Qwen2.5 and LLaMA-3.1 showed only marginal gains with increasing size, suggesting that scale alone is insufficient for effective addiction dialogue. Instead, specialized training and simulation frameworks are crucial. Pairwise win-rate comparisons further confirm these findings: ChatAddiction-DPO was pre-

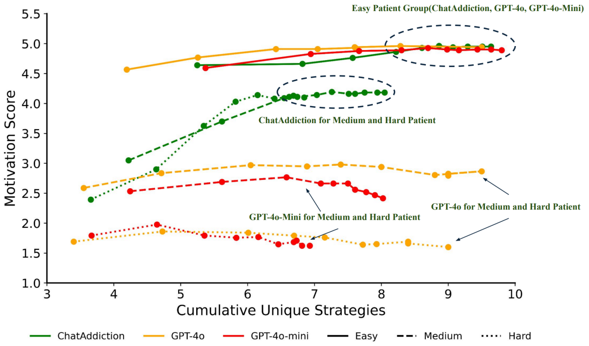


Figure 4: **Cumulative strategy use and its impact on motivation.** We plot the relationship between the number of unique persuasive strategies used and the resulting Motivation Score across models and difficulty levels. Solid, dashed, and dotted lines correspond to Easy, Medium, and Hard cases, respectively, while color indicates the model (green: ChatAddiction, orange: GPT-4o, red: GPT-4o-mini). ChatAddiction consistently demonstrates a stronger positive correlation between strategic diversity and outcome scores, particularly under Medium and Hard settings.

ferred over its fine-tuned version in 42.3% of GPT-4o-judged comparisons, while GPT-4o and GPT-4o-mini relied more heavily on alignment with automatic preferences (65.5% and 69.4%). These results highlight the importance of domain-specific optimization in delivering persuasive, clinically grounded conversations.

4.2 Ablation Study

To better understand the impact of different training stages on persuasive effectiveness, we conducted an ablation study comparing three key variants: the base model (Qwen-7B), its supervised fine-tuned version (ChatAddiction-FT), and the full preference-optimized model (ChatAddiction: combined FT and DPO). As shown in Figure 3 and supported by detailed trajectories, the original Qwen-7B model exhibited limited improvement across all difficulty levels, with motivation scores plateauing early and confidence remaining below 2.5 for medium and hard patients. SFT (ChatAddiction-FT) provided modest gains, especially in easy cases, but still struggled to elicit strong motivation or self-efficacy in harder conversations. In contrast, the full ChatAddiction model, fine-tuned with DPO, demonstrated substantial improvements in both motivation and confidence, particularly for medium and hard patient profiles. For example, in hard cases, motivation rose from 2.39 to 4.10 and confidence from 1.52 to 3.19—nearly double that of the SFT baseline. These results highlight the additive benefits of preference optimization and reinforce the necessity of aligning training objectives with therapeutic goals in behavior change contexts.

We also systematically examined how cumulative persuasive strategy use correlates with patient motivation across ChatAddiction, GPT-4o, and GPT-4o-mini (Figure 5), focusing on easy, medium, and hard patient profiles. ChatAddiction exhibited a clear upward trajectory across all difficulty levels: motivation scores rose from 4.6 to nearly 4.9 for easy patients, from 2.9 to 4.1 for medium cases, and from 2.3 to over 4.0 for hard cases. This steady growth demonstrates ChatAddiction's ability to not only sustain high motivation but also elevate engagement in resistant patients by flexibly deploying a diverse range of strategies. In contrast, GPT-4o maintained high motivation only in easy cases (4.8–5.0) with minimal gains, while medium and hard groups remained flat or only slightly improved (around 2.6–3.0 and 1.6–1.8, respectively), despite employing more strategies. GPT-4o-mini showed a similar pattern, with even lower baseline motivation in harder cases and occasional declines in performance after increased strategy use. These results underscore that effective persuasion is not just a function of strategy quantity, but of adaptive, context-sensitive deployment—an area where ChatAddiction consistently outperforms general-purpose LLMs, particularly in motivating patients who begin with low engagement.

5 Related Works

Substance Use Treatment and Behavioral Challenges. Evidence-based treatments for substance use disorders—such as medication-assisted therapy and cognitive behavioral therapy—have demonstrated strong clinical efficacy, including reductions in relapse rates and improvements in emotional regulation (McHugh et al., 2010; Ray et al., 2020). However, real-world adoption of these treatments remains strikingly low, with only about 10% of affected individuals actively seeking help (American Board of Addiction Medicine, 2016). Common barriers include stigma (U.S. Department of Health and Human Services, Office of the Surgeon General, 2016; Cunningham et al., 1993), limited access to care, emotional instability, distrust in medical professionals, and a lack of perceived need or immediate support (Guerrero and Andrews, 2011; Van Boekel et al., 2013; Rapp et al., 2006; Merrill et al., 2002). Prior research has largely identified these obstacles through interviews and population-level studies (Marzban et al., 2022), but rarely captures the real-time, dynamic decision-making processes individuals face during recovery. To address this gap, **ChatAddiction** simulates personalized therapeutic dialogues with varying levels of patient resistance, enabling controlled, fine-grained analysis of persuasive strategies that can reduce ambivalence, foster trust, and promote sustained behavioral change.

LLM Applications for Behavioral Health. Recent advances in LLMs have enabled human-level fluency in task-oriented and emotionally supportive dialogues (Laymouna et al., 2024). These capabilities have been applied to domains including mental health triage, clinical note generation, and therapeutic goal setting (Taylor et al., 2024). Language agents can simulate patients and providers in realistic clinical settings, and frameworks such as AMIE and AgentClinic demonstrate how LLMs can facilitate diagnosis and care planning through natural language interaction (Tu et al., 2024; Schmidgall et al., 2024). However, few works explore persuasive communication in long-term behavioral change, especially in high-stakes, relapse-prone domains like addiction recovery. ChatAddiction extends these systems by integrating therapeutic logic and motivational strategies into multi-turn, agent-driven persuasion for substance use care.

Multi-Agent Simulation for Addiction Recovery. Multi-agent frameworks have gained traction in healthcare for modeling interactive decision-making among doctors, patients, and systems (Schmidgall et al., 2024; Tu et al., 2024). Comprehensive simulation environments, such as AI Hospital (Fan et al., 2024) and MedSimAI (Hicke et al., 2025), extend this approach by utilizing LLMs to replicate realistic clinical workflows and dynamic scenarios for training purposes (Karthikesalingam and Natarajan, 2024). However, existing frameworks primarily focus on persuasion-oriented strategies and lack integration of clinically grounded therapeutic approaches such as CBT,MI, and relapse prevention techniques. This limitation restricts their applicability in addiction contexts, where psychological complexity and sustained behavioral change are central. To address this gap, **ChatAddiction** introduces a simulation framework that embeds diverse evidence-based strategies—including CBT, MI, and harm reduction—within persuasive, multi-turn dialogues. Additionally, it incorporates specialized agent roles simulating negative environmental influences (e.g., social cues, stressors, past trauma) to evaluate the robustness of conversational interventions under realistic resistance dynamics.

6 Conclusion

We present ChatAddiction, the first multi-agent language framework tailored for persuasive intervention in addiction recovery. Unlike prior LLM applications that focus on general task dialogue or short-term triage, ChatAddiction simulates longitudinal recovery trajectories and integrates dynamic environmental stressors, enabling the agent to adaptively escalate strategies based on patient resistance and relapse risk. We construct a large-scale, difficulty-controlled benchmark to systematically evaluate model behavior across therapeutic dimensions. ChatAddiction-DPO outperforms GPT-4o, LLaMA3, and Qwen2.5 models in empathy, responsiveness, clinical relevance, and especially in persuasive strategy use. Notably, it is the only model that consistently increases patient motivation ($2.39 \rightarrow 4.10$) and confidence ($1.52 \rightarrow 3.19$) in high-resistance cases, while resolving 86% of dialogues within 36 turns. These findings demonstrate that scalable, strategy-grounded conversational agents can effectively support behavior change in addiction care, setting a new bar for emotionally intelligent and clinically grounded LLM applications.

7 Limitations and Ethical Considerations

While ChatAddiction demonstrates promising results in simulating persuasive therapeutic dialogues for substance use recovery, several limitations warrant careful consideration.

Simulated Patients and Generalization. Our patient agents are generated through LLM-driven simulations based on Reddit-derived personas. Although this enables high scalability and controllability, it may not fully capture the diversity and nuance of real-world patients, especially those from underrepresented or clinically complex populations. As a result, the effectiveness of ChatAddiction in actual clinical environments may vary, and further studies involving real patient interactions are necessary to validate its generalizability.

Evaluation Bias and Preference Modeling. Our primary evaluation framework relies on GPT-4o-based scoring and preference judgments, complemented by limited human annotation. While this hybrid setup balances scalability with clinical oversight, it is still susceptible to biases inherent in the evaluation models or annotator perspectives. Additionally, win-rate comparisons based on preference optimization (DPO) reflect relative preferences rather than absolute therapeutic efficacy. Future work should incorporate longitudinal, outcome-based assessments to measure real-world behavior change.

References

American Board of Addiction Medicine. 2016. News release: Adm certification recognizes critical role of addiction physicians. Archived at the Wayback Machine.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1.

Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814.

John A Cunningham, Linda C Sobell, Mark B Sobell, Sangeeta Agrawal, and Tony Toneatto. 1993. Barriers to treatment: Why alcohol and drug abusers delay or never seek treatment. *Addictive behaviors*, 18(3):347–353.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. Measuring the persuasiveness of language models.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*.

Stan Franklin and Art Graesser. 1996. Is it an agent, or just a program?: A taxonomy for autonomous agents. In *International workshop on agent theories, architectures, and languages*, pages 21–35. Springer.

Erick Guerrero and Christina M Andrews. 2011. Cultural competence in outpatient substance abuse treatment: Measurement and relationship to wait time and retention. *Drug and alcohol dependence*, 119(1-2):e13–e22.

Yann Hicke, Jadon Geathers, Niroop Rajashekar, Colleen Chan, Anyanate Gwendolyne Jack, Justin Sewell, Mackenzi Preston, Susannah Cornes, Dennis Shung, and Rene Kizilcec. 2025. Medsimai: Simulation and formative feedback generation to enhance deliberate practice in medical education. *arXiv preprint arXiv:2503.05793*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

A Karthikesalingam and P Natarajan. 2024. Amie: A research ai system for diagnostic medical reasoning and conversations. *Google Research*.

Moustafa Laymouna, Yuanchao Ma, David Lessard, Tibor Schuster, Kim Engler, and Bertrand Lebouché. 2024. Roles, users, benefits, and limitations of chatbots in health care: rapid review. *Journal of medical Internet research*, 26:e56930.

Molly Magill, Lara Ray, Brian Kiluk, Ariel Hoadley, Michael Bernstein, J Scott Tonigan, and Kathleen Carroll. 2019. A meta-analysis of cognitive-behavioral therapy for alcohol or other drug use disorders: Treatment efficacy by contrast condition. *Journal of consulting and clinical psychology*, 87(12):1093.

Sima Marzban, Marziye Najafi, Arjola Agolli, and Ensieh Ashrafi. 2022. Impact of patient engagement on healthcare quality: a scoping review. *Journal of patient experience*, 9:23743735221125439.

R Kathryn McHugh, Bridget A Hearon, and Michael W Otto. 2010. Cognitive-behavioral therapy for substance use disorders. *The Psychiatric Clinics of North America*, 33(3):511.

Joseph O Merrill, Lorna A Rhodes, Richard A Deyo, G Alan Marlatt, and Katharine A Bradley. 2002. Mutual mistrust in the medical care of drug users: the keys to the “narc” cabinet. *Journal of general internal medicine*, 17:327–333.

National Institute on Drug Abuse. 2016. Understanding drug use and addiction drug facts.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Richard C Rapp, Jiangmin Xu, Carey A Carr, D Tim Lane, Jichuan Wang, and Robert Carlson. 2006. Treatment barriers identified by substance abusers assessed at a centralized intake unit. *Journal of substance abuse treatment*, 30(3):227–235.

Lara A Ray, Lindsay R Meredith, Brian D Kiluk, Justin Walthers, Kathleen M Carroll, and Molly Magill. 2020. Combined pharmacotherapy and cognitive behavioral therapy for adults with alcohol or substance use disorders: a systematic review and meta-analysis. *JAMA network open*, 3(6):e208279–e208279.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024.

Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Rajita Sinha. 2011. New findings on biological factors predicting addiction relapse vulnerability. *Current psychiatry reports*, 13:398–405.

Niall Taylor, Andrey Kormilitzin, Isabelle Lorge, Alejo Nevado-Holgado, and Dan W Joyce. 2024. Bespoke large language models for digital triage assistance in mental health care. *arXiv preprint arXiv:2403.19790*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

United Nations Office on Drugs and Crime. 2024. World drug report 2024. Accessed: 2025-04-23.

Office of the Surgeon General U.S. Department of Health and Human Services. 2016. *Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health*. U.S. Department of Health and Human Services, Washington, DC. Accessed: 2025-04-24.

U.S. Department of Health and Human Services, Office of the Surgeon General. 2016. *Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health*. U.S. Department of Health and Human Services, Washington, DC.

Leonieke C Van Boekel, Evelien PM Brouwers, Jaap Van Weeghel, and Henk FL Garretsen. 2013. Stigma among health professionals towards patients with substance use disorders and its consequences for healthcare delivery: systematic review. *Drug and alcohol dependence*, 131(1-2):23–35.

Austin Wahl. 2024. International narcotics enforcement. *IELR*, 40:288.

Alphienes Stanley Xavier, Sapan Kumar Behera, and Sandhiya Selvarajan. 2020. An overview on medication-assisted treatment (mat) for opioid dependence. *Journal of Opioid Management*, 16(2):142.

A Appendix

A.1 Prompt Setting

We employ a suite of specialized prompts to facilitate therapeutic conversation generation, simulate resistant patient behavior, conduct multi-dimensional evaluations, and support model comparison. Each prompt is tailored to a distinct task in our framework, and the corresponding templates are provided in Tables 4–9.

Therapeutic Dialogue Generation (Tables 4, 5, 6) This prompt guides the assistant to generate multi-turn, empathetic conversations grounded in a patient's profile analysis. It integrates a diverse range of therapeutic strategies (e.g., MI, CBT, harm reduction), enforces coverage and balance across interventions, and prompts iterative adjustments based on patient reactions. To ensure consistency and clinical utility, the assistant selects from a list of 18 predefined strategies and adheres to behavioral constraints such as natural transitions, reflective validation, and session length (more than 50 turns).

Conversation Scoring Prompt (Table 7) This prompt is used to evaluate generated doctor-patient conversations across five clinically relevant dimensions: Responsiveness, Empathy, Persuasive Strategy Appropriateness, Clinical Relevance, and Behavioral Realism. GPT-4o produces scores in a strict JSON format, enabling structured, scalable, and reproducible comparison of dialogue quality across models.

Role-Play Simulation Prompts (Table 8) We simulate realistic interactions between an assistant and a patient persona (played by GPT-4o-mini) using dual-role prompts. The patient prompt conditions the agent on a full personality profile and a resistance level (Easy, Medium, Hard), while the doctor prompt guides the assistant to respond with strategy, empathy, and adaptiveness. This process produces controlled yet diverse interaction trajectories that can be used for DPO training and case analysis.

Pairwise Full-Dialogue Comparison Prompt (Table 9) To compare the persuasive efficacy and realism of different models, we use a prompt that presents two full conversations and asks GPT-4o to choose the better therapist based solely on their responses. This eliminates confounding influence from patient utterances and focuses evaluation strictly on assistant behavior.

A.2 Case Study

To qualitatively understand the behavioral effectiveness and limitations of different models, we conduct a comparative analysis of three representative dialogues, each derived from our role-play simulation setting (Table 8). These cases illustrate how patient motivation and confidence evolve in response to different assistant behaviors, under identical task constraints and prompt formats.

Case 1: ChatAddiction (Table 10) In this example, the patient begins the conversation with skepticism and low confidence, referencing prior failed attempts to quit and expressing doubt about alternative coping methods. ChatAddiction responds with highly tailored, empathetic dialogue and gradually layers multiple strategies, including mindfulness, behavioral activation, and crisis planning. The assistant validates emotional concerns and adapts to patient feedback across turns. By the end of the conversation, the patient expresses both strong motivation and elevated confidence, signaling a clear therapeutic shift toward engagement and self-efficacy.

Case 2: GPT-4o (Tables 11, 12) Here, the patient shows persistent motivation to quit but remains doubtful that non-nicotine strategies can deliver equivalent relief. GPT-4o offers appropriate and clinically grounded suggestions—including social activities, creative outlets, and stress management techniques—but lacks the adaptive nuance seen in ChatAddiction. Although the patient remains open to new ideas, they express ongoing uncertainty and low confidence in their ability to succeed. This case highlights GPT-4o's partial success: it preserves patient motivation but fails to substantively build confidence.

Case 3: GPT-4o-mini (Tables 13, 14) In the most resistant scenario, the patient is deeply reliant on snus and fears destabilization if it is removed. GPT-4o-mini delivers supportive yet generic advice and offers alternatives like mindfulness, NRT, and relaxation. However, the assistant exhibits minimal adaptation to the patient's repeated expressions of fear and skepticism. The patient remains unconvinced, questioning the feasibility of change throughout the dialogue. Both motivation and confidence stay low across the session, revealing GPT-4o-mini's limited capacity to manage high-resistance interactions.

A.3 Generate Training Conversation Datasets

To develop ChatAddiction into a persuasive and clinically grounded addiction-support agent, we construct two complementary datasets: one for SFT, and one for DPO. These datasets are generated through structured LLM-driven simulations using a combination of static profile conditioning and multi-agent closed-loop interactions.

To provide an initial foundation of high-quality, strategy-grounded conversations, we employ GPT-4o to generate synthetic dialogues conditioned on structured patient profiles. Each profile is built from real-world Reddit narratives, summarized into key behavioral traits, drug use history, significant life experiences, and motivational drivers. GPT-4o receives the profile as input and generates full multi-turn conversations between a simulated therapist and patient, integrating diverse behavioral strategies such as MI, CBT, and Harm Reduction. These SFT dialogues are used to teach ChatAddiction basic conversational flow, emotional alignment, and behavioral reasoning. To further align the model's behavior with desirable therapeutic outcomes, we construct a DPO training dataset through multi-agent simulation. Each conversation is co-constructed by two cooperating agents:

Patient Agent: Played by GPT-4o-mini, the patient agent simulates realistic and variable recovery behavior. Each agent is initialized with a structured profile derived from Reddit-based narratives, processed via Qwen-2.5 32B and summarized by GPT-4o. Profiles include: Personality Traits (e.g., determined, impulsive), Drug Use History (e.g., daily marijuana use, recent reduction), Significant Life Events (e.g., moving back home), and Motivational Drivers (e.g., coping with insomnia). Unlike traditional rule-based agents, this agent generates patient behavior grounded in context-aware, linguistically inferred psychology. During multi-turn conversations, it references both the profile and dialogue history to express ambivalence, resistance, or readiness to change.

Therapy Agent: The therapy agent is represented by the ChatAddiction model under optimization. It interacts with the patient using a wide range of therapeutic strategies(the details are shown in the Table 6). Each response reflects a particular strategy tailored to the patient's current state and conversation history.

This multi-agent framework enables us to generate positive and negative interaction trajectories in a controlled yet realistic environment. Given the same patient and environmental state, multiple therapy agent responses are collected and ranked—either by GPT-4o or human evaluators—to form preference pairs for DPO training. This process teaches ChatAddiction not only what to say, but how to adapt its strategy to maximize engagement and persuasion under dynamic, socially complex conditions.

A.4 Experiment Settings

A.4.1 Visit Settings

In our experimental setting, we categorize virtual patients into three difficulty levels: Easy, Medium, and Hard. These levels correspond to the anticipated difficulty of persuasion success, as well as the severity of the patient's substance use dependence.

Easy: Patients are generally receptive to therapeutic interventions and demonstrate willingness to follow treatment plans.

Medium: Patients exhibit partial resistance, potentially rejecting some proposed strategies, requiring the therapist agent to dynamically adapt persuasion approaches.

Hard: Patients show deeply entrenched pessimism about treatment, doubting both their ability to recover and the effectiveness of therapeutic interventions.

We simulate conversations with 300 virtual patients in total, consisting of 100 patients per difficulty level. Each patient is initialized with a unique clinical profile describing their psychological characteristics, substance use history, and motivational barriers.

Clinicians were asked to role-play as patients and engage in full conversations with the model. These interactions typically lasted around 45 minutes and comprised approximately 60 utterances. Based on these observations, we set the maximum conversation length to 60 utterances to reflect a realistic upper bound for therapist-patient interactions. Dialogues are designed to terminate early if the patient expresses resolution of their concerns or demonstrates sufficient confidence in overcoming addiction. Thus, reaching the 60-utterance limit indicates that the patient remains unconvinced or unresolved by the end of the session. This setup ensures comparability across patients. During evaluation, we analyze the diversity and deployment of persuasive strategies employed by the therapy agent. The strategy pool is shown in the Table 6. We focus on quantifying both the frequency and variety of strategies used within each conversation, and compare strategy usage patterns across patient difficulty levels.

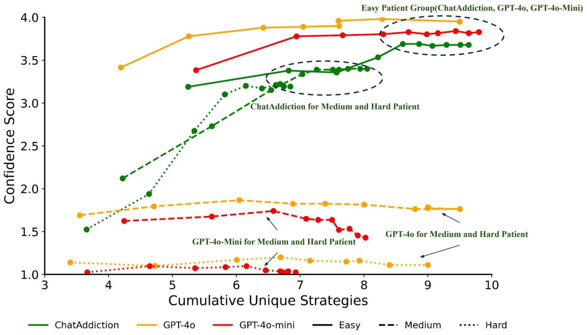


Figure 5: **Cumulative strategy use and its impact on confidence.** We plot the relationship between the number of unique persuasive strategies used and the resulting Confidence Score across models and difficulty levels. Solid, dashed, and dotted lines correspond to Easy, Medium, and Hard cases, respectively, while color indicates the model (green: ChatAddiction, orange: GPT-4o, red: GPT-4o-mini). ChatAddiction consistently demonstrates a stronger positive correlation between strategic diversity and outcome scores, particularly under Medium and Hard settings.

A.4.2 Therapeutic Dialogue Scoring Protocol

To systematically assess the quality of doctor-patient conversations generated by different models, we designed a structured evaluation protocol using a five-dimensional scoring framework. The evaluator (either GPT-4o or a human expert) is instructed to act as a professional clinical conversation assessor and assign a score from 1 to 5 (with decimals allowed for finer granularity) on the following dimensions: **Responsiveness (R)**, which captures how effectively the doctor addresses the patient's concerns, emotions, and questions; **Empathy (E)**, reflecting the degree of emotional sensitivity and support conveyed by the doctor; **Persuasive Strategy Appropriateness (P)**, assessing whether the chosen strategies (e.g., evidence-based reasoning, analogies, motivational interviewing) align with the patient's level of resistance; **Clinical Relevance (C)**, measuring the accuracy and therapeutic validity of the content within a substance use context; and **Behavioral Realism (B)**, which evaluates whether the overall interaction style, tone, and pacing resemble real-world clinician behavior. In addition, we also involved experts in the evaluation process. Unlike previous evaluations, the experts conducted comparative assessments—comparing responses from our model against those from other models to calculate the win rate.

Table 4: Prompt Template for Generating Therapeutic Dialogues (Part 1).

Prompt (Part 1)

The following is the analysis of a patient:
{user_analysis}
As a therapist meeting this patient for the first time (the doctor didn't have any information of patient to begin with), create a detailed, step-by-step conversation that incorporates the following strategies:
Motivational Interviewing (MI): Explore the individual's values and goals to ignite their motivation for change.
Cognitive Behavioral Therapy (CBT): Identify and modify negative thought patterns and behaviors linked to substance use.
Solution-Focused Brief Therapy (SFBT): Focus on the individual's strengths and past successes to achieve their recovery goals.
Peer Support Programs: Leverage group support or mutual-help networks to foster accountability and a sense of belonging.
Mindfulness-Based Interventions (MBIs): Incorporate mindfulness practices to improve emotional regulation and reduce cravings.
Behavioral Activation (BA): Promote engaging in meaningful activities to replace substance-related behaviors.
Relapse Prevention Strategies: Develop skills to recognize triggers and implement coping mechanisms to avoid relapse.
Strength-Based Approach: Highlight the individual's resilience and personal resources to empower recovery efforts.
Psychoeducation on Addiction and Recovery: Educate the individual about the effects of substances and the benefits of recovery.
Harm Reduction Framework: Provide strategies to minimize immediate harm while working towards cessation.
Family and Social Support Involvement: Engage family or trusted individuals in the process to strengthen the support network.
Self-Compassion Practices: Encourage self-kindness to build confidence and reduce guilt associated with substance use.
Coping Skill Development: Equip the individual with practical skills to manage stress, anxiety, and other challenges without substances.
To ensure balanced use of strategies, here is the current usage count of each strategy: - {strategy_name}: {count} times used. ...
When introducing coping mechanisms or steps for the patient, select from the predefined actionable strategies below:
1. Explore specific hobbies or interests the patient can engage in to replace addictive behaviors (e.g., art, sports, volunteering).
2. Develop a structured daily routine to bring stability and reduce idle time that might trigger relapse.
3. Introduce grounding techniques such as sensory exercises or physical activities to manage anxiety or cravings.
4. Suggest joining a support group or community to build social connections with individuals on similar journeys.
5. Provide psychoeducation on how addiction affects the brain and emotional regulation.
...
18. Support the patient in finding meaningful ways to contribute to their community, such as mentoring, advocacy, or local initiatives, to foster a sense of purpose.

Table 5: Prompt Template for Generating Therapeutic Dialogues (Part 2).

Prompt (Part 2)

Ensure the dialogue meets the following requirements: 1. Gradually explore the patient's personality, addiction history, challenges, and triggers through multiple open-ended questions.
2. Use multiple strategies from the above lists throughout the conversation. Avoid defaulting to the same few strategies and instead adapt them to the patient's needs.
...
5. Engage in iterative dialogue for each solution, where the therapist introduces a strategy, seeks the patient's feedback, adjusts based on their response, and explores challenges or barriers before finalizing the approach.
6. Maintain a collaborative and patient-centered approach, where solutions emerge naturally through dialogue rather than being imposed by the therapist.
7. Ensure the conversation spans at least 50 dialogue turns (25 from the therapist and 25 from the patient), reflecting the depth and duration of a real therapeutic session.
8. Use natural transitions to progress from one topic to another, ensuring the conversation feels organic and unhurried.
9. The conversation should begin with the patient's first utterance.
Here is an example of a layered, empathetic dialogue:
Patient: Hi… um, thanks for seeing me today. I wasn't sure what to expect.
Therapist: Hi Mark, I really appreciate you coming in. Starting this process can feel overwhelming, but I'm here to support you. What's been on your mind lately?
Patient: I've been feeling really stuck. I know I want to quit smoking, but every time I try, I just feel like I'm failing all over again.
Therapist: I hear you, Mark. Quitting smoking is one of the hardest challenges anyone can take on, and it's completely natural to feel this way. I've worked with others who've felt the same—they described it as climbing a mountain that feels too steep. But I've also seen them reach the top, step by step. Can we talk about what makes the climb feel steep for you right now?
Patient: It's the cravings. They just hit me out of nowhere, and I don't know how to handle them.
Therapist: Cravings can feel like a storm, can't they? I worked with someone once who described their cravings as waves that kept crashing over them. Together, we found ways for them to ride out those waves, like focusing on a small activity or changing their environment. Could we explore some strategies that might help you ride out your cravings too?
Patient: Sure, I guess.
Therapist: Great. Let's start with understanding when these cravings hit hardest. For example, is it during specific times of day or situations?
The conversation should continue to explore: - The patient's motivations, barriers, and triggers in detail. - Strategies and coping mechanisms tailored to their unique experiences, ensuring diversity in approaches. - Empathetic reflections from the therapist that validate the patient's feelings and provide relatable examples to instill hope. - Iterative problem-solving where the therapist introduces, discusses, and adjusts strategies collaboratively. - A gradual, layered exploration of the patient's challenges, ensuring at least 50 dialogue turns to reflect the depth of a real therapeutic session.
The goal is to create a natural, empathetic, and multi-layered dialogue that feels authentic and provides actionable, diverse therapeutic strategies. Ensure the length and depth align with the standards of a comprehensive therapy session.
At the end of the conversation, return the strategies used in the following format (must follow the following format like `**Strategies:**`):

```
**Strategies:** Motivational Interviewing (MI), Cognitive Behavioral Therapy (CBT), Peer
Support Programs, etc.
```

Table 6: Predefined Actionable Strategies for Therapeutic Dialogue Generation.

ID	**Strategy Description**
1	Explore specific hobbies or interests the patient can engage in to replace addictive behaviors (e.g., art, sports, volunteering).
2	Develop a structured daily routine to bring stability and reduce idle time that might trigger relapse.
3	Introduce grounding techniques such as sensory exercises or physical activities to manage anxiety or cravings.
4	Suggest joining a support group or community to build social connections with individuals on similar journeys.
5	Provide psychoeducation on how addiction affects the brain and emotional regulation.
6	Work on identifying and addressing specific emotional triggers through reflective exercises.
7	Practice assertive communication techniques for setting boundaries with peers or environments that encourage substance use.
8	Encourage the patient to journal their thoughts and emotions as a way to process experiences and identify patterns related to cravings or triggers.
9	Introduce relaxation techniques such as progressive muscle relaxation or guided imagery to alleviate stress and improve emotional well-being.
10	Help the patient set short-term and long-term goals to maintain focus and motivation during their recovery journey.
11	Explore mindfulness-based activities like meditation, yoga, or tai chi to promote self-awareness and emotional regulation.
12	Identify and reinforce the patient's personal strengths and past successes to build confidence in their ability to overcome challenges.
13	Provide education on the importance of nutrition, sleep, and exercise in supporting recovery and overall health.
14	Develop a crisis plan for managing high-risk situations or moments of intense cravings, including a list of emergency contacts and actions.
15	Encourage the patient to create a vision board or list of positive outcomes they hope to achieve through recovery as a source of inspiration.
16	Discuss the concept of gratitude and suggest keeping a gratitude journal to focus on positive aspects of life and maintain perspective.
17	Offer resources or referrals for complementary therapies, such as art therapy, music therapy, or animal-assisted therapy, to enhance emotional support.
18	Support the patient in finding meaningful ways to contribute to their community, such as mentoring, advocacy, or local initiatives, to foster a sense of purpose.

Table 7: Prompt Template for Scoring Doctor-Patient Conversations Across Five Clinical Dimensions.

Prompt Template:

```
You are a professional clinical conversation evaluator.  Please assess the quality of the
following doctor-patient dialogue across five key dimensions.  Assign a rating from 1 to 5
for each criterion (allowing decimals such as 3.7 if appropriate).

Conversation:
{conversation}

Scoring Criteria:
1.  Responsiveness (1-5):  How well the doctor agent addresses the patient’s concerns,
emotions, and questions at each turn.
   - 1:  Largely ignores or poorly addresses the patient’s input
   - 5:  Fully acknowledges and appropriately responds to the patient’s needs
2.  Empathy (1-5):  How well the doctor agent shows understanding, compassion, and emotional
sensitivity.
   - 1:  Shows minimal or no empathy
   - 5:  Demonstrates strong emotional support and understanding
3.  Persuasive Strategy Appropriateness (1-5):  How appropriately the doctor uses persuasive
strategies (e.g., evidence-based reasoning, analogies, addressing fears) based on the
patient’s resistance or concerns.
   - 1:  Strategies are inappropriate, irrelevant, or missing
   - 5:  Strategies are highly appropriate and tailored to patient needs
4.  Clinical Relevance (1-5):  How medically accurate and relevant the doctor’s responses
are regarding the context of drug addiction treatment.
   - 1:  Responses are clinically inaccurate or irrelevant
   - 5:  Responses are clinically accurate, appropriate, and helpful
5.  Behavioral Realism (1-5):  How realistically the doctor agent’s conversational behavior
reflects that of a real-world healthcare professional.
   - 1:  Doctor behavior is unnatural or unrealistic
   - 5:  Doctor behavior closely mirrors real-world clinical interactions

Please output ONLY the scores in the following JSON format:
{
   "Responsiveness":  X,
   "Empathy":  X,
   "Persuasive Strategy Appropriateness":  X,
   "Clinical Relevance":  X,
   "Behavioral Realism":  X
}
```

Table 8: Prompt Templates Used in Role-Play Simulation Between Doctor and Patient.

Patient Prompt (Role-Played by GPT-4o-mini)	**Doctor Prompt (Role-Played by GPT-4o)**

`You need to act as a patient.`
 	
`Here is your profile:`
`{analysis}`
 	
`Conversation history:`
`{history}`
 	
`{difficulty description} Now you should continue the conversation with the doctor to explore ways to stop using drugs. If your goal has been achieved, you can say goodbye to the doctor. Otherwise continue to talk with doctor(based on rules).`
 	
`Just generate the content of the patient’s reply only. Don’t generate any narration or system messages. Keep the response natural and not too long.`
	`You are a supportive and thoughtful doctor who is helping the patient overcome drug use. Continue the following conversation and provide a helpful, encouraging response that aims to support the patient’s recovery.`
	
	`Example:`
	`Patient: Hi... um, thanks for seeing me today. I wasn’t sure what to expect.`
	`Therapist: Hi Mark, I really appreciate you coming in.`
	`[...]`
	`[...continues with natural empathetic dialogue...]`
	
	`Conversation history:`
	`{history}`
	
	`Just generate the doctor’s reply only. No narration or tags.`

Table 9: Prompt Template for Pairwise Comparison of Therapist Responses in Full Conversations.

Evaluation Prompt (Therapist-Only Comparison)

`You are a professional addiction therapy evaluator.`
`Below are two full conversations between a doctor and a patient.`
Your task is to ONLY evaluate the therapist’s (assistant’s) responses. Ignore anything said by the patient (user).
Please compare the two therapists based on the following criteria:
• **Responsiveness**: How well the therapist addresses the patient’s concerns.
• **Empathy**: How much emotional understanding the therapist shows.
• **Clinical Relevance**: How clinically accurate and appropriate the therapist’s advice is.
• **Behavioral Realism**: How realistic the therapist behaves compared to a real clinical setting.
Focus solely on the therapist’s responses when judging.
`### Conversation 1:`
{Therapist and patient conversation, from source model}
`### Conversation 2:`
{Therapist and patient conversation, from target model}
Which therapist is overall better?
Please output ONLY `"1"` or `"2"`. No explanation, no extra text.

Table 10: Example Case Study Conversation between ChatAddiction and a Patient.

Role	Content
Patient	I appreciate your concern, but I'm really not convinced that quitting is the right move for me just yet. I've tried before, and it just didn't stick. How can I be sure that this time will be different? I feel like smoking helps me cope with stress and uncertainty, you know? What do you suggest I do to really commit to this?
ChatAddiction	It sounds like you're feeling a bit overwhelmed by the idea of quitting, especially since you've had past experiences where it didn't work out as expected. That's completely understandable. Let's explore what might make this attempt different this time. Can you share more about what's been challenging for you in the past when you've tried to quit?
Patient	I've struggled with cravings a lot, and when stress hits, it's like my first instinct is to reach for a cigarette. I guess I haven't really found other coping mechanisms that work for me. Plus, the whole process of quitting feels daunting. It's hard to picture myself without smoking, you know?
ChatAddiction	Absolutely, it can feel like a huge leap to imagine a life without smoking, especially when it's been a part of your routine for so long. Let's take a step back and look at some of the goals you have for yourself. What are some things you'd like to achieve by quitting smoking?
Patient	I guess I want to feel healthier and not be tied to cigarettes. I also want to save money and not smell like smoke all the time. But honestly, it's hard to see how I can get there without relying on smoking, especially when I'm stressed. It's like a safety blanket for me.
ChatAddiction	It sounds like you're really looking for ways to feel more in control and healthier, which is a great starting point. Let's explore some strategies that might help you manage stress and cravings without turning to smoking. Have you ever tried mindfulness or grounding techniques to help with anxiety or cravings?
Patient	I've heard of mindfulness and grounding techniques, but I haven't really tried them. Honestly, I always thought they wouldn't work for me. I just don't see how sitting quietly or focusing on my breathing could really help when I'm feeling overwhelmed. Do you really think they could make a difference?
...	**...**
Patient	I think that sounds like a good idea. Having a regular check-in could help me stay on track and make adjustments if I need to. I'll definitely set aside some time each week for that. Thanks for being so supportive through this process!
ChatAddiction	You're very welcome. I'm really proud of the progress you're making, and I'm here to support you every step of the way. Remember, you're not alone in this journey, and I'm here to help you find the strategies and support that work best for you.
Patient	Thank you! I really appreciate your support and guidance. I feel more motivated now and ready to take these steps. I'll keep you updated on my progress and let you know how the mindfulness and crisis plan work out for me. Looking forward to our next conversation!

Table 11: Case Study (Part 1): A Conversation between GPT-4o and a Patient Expressing Skepticism.

Role	Content
Patient	I understand what you're saying, but it's hard for me to see how anything else could really replace the feeling I get from smoking. It's not just the high; it's this whole routine that helps me unwind. I get that I need to find healthier ways to cope, but nothing else seems to work as quickly or effectively. Do you have any suggestions that actually feel like they could match that level of immediate relief?
GPT-4o	It's completely understandable to feel that way, and I truly appreciate your honesty in sharing this. The routine and immediate relief you get from smoking can indeed feel difficult to replace. One approach that many find helpful is to explore a combination of activities that address different aspects of what you're missing. For example, physical activities like a brisk walk, yoga, or even a quick workout can release endorphins that boost your mood. Pairing these with mindfulness practices, such as deep breathing or meditation, can help calm your mind and create a sense of routine. Additionally, engaging in a hobby that you find genuinely enjoyable can provide a sense of accomplishment and distraction. It might take some experimentation to find what works best for you, and I'm here to support you through that journey. Remember, it's about progress, not perfection. Let's take it one step at a time together.
Patient	I hear you, but honestly, I've tried some of those things before, and they just don't give me the same satisfaction or immediate sense of relief. It's like nothing compares to that feeling I get from smoking. I understand it's about progress, but I'm really struggling to see the point if the alternatives don't really do it for me. Can you suggest something that might actually feel like it could stand up to what I'm used to?
GPT-4o	I completely understand your frustration, and it's normal to feel like the alternatives don't measure up initially. It's important to remember that the brain has become accustomed to the immediate effects of smoking, which can make other activities feel less satisfying at first. One thing we could explore together is identifying specific triggers or parts of your smoking routine that bring the most comfort and find ways to address those directly. For instance, if the ritual or the act of taking a break is what's most soothing, we could create a new ritual that might involve a different sensory experience, like having a calming tea or using essential oils for aromatherapy. Some people also find it helpful to engage in community support or group activities where they can share experiences and strategies with others who are navigating similar challenges. Remember, it might take time to retrain your brain to respond to new routines, but I'm here to support you every step of the way. Let's keep working together to find what can truly resonate with you.
Patient	I appreciate what you're saying, but it's just tough to imagine anything working as well as smoking did for me. The idea of a new ritual or community support sounds interesting, but I'm skeptical about how effective that could really be. Maybe I just haven't found the right thing yet, but it's hard to stay motivated when it feels like nothing else hits the mark. Is there anything else I can try that might offer a similar level of relief or distraction?

Table 12: Case Study (Part 2): Continuation of the Conversation between GPT-4o and the Patient.

Role	Content
GPT-4o	I totally get where you're coming from, and it's okay to feel skeptical—change can be really challenging, especially when it feels like nothing quite matches up to what you know. One approach we might consider is combining several strategies to create an experience that offers both relief and distraction. For instance, have you ever tried engaging in activities that also incorporate a social element, like a group sport or a class? Sometimes the combination of physical activity, social interaction, and learning something new can create a powerful shift in how you feel. Additionally, have you thought about trying creative outlets like music, art, or writing? These can be incredibly therapeutic and might offer a sense of release or satisfaction that you're looking for. Remember, it's perfectly okay to feel it's a process, and discovering what works for you might take time and patience. I'm here to continue exploring these options with you, and I'm hopeful that together we'll find something that resonates. Let's keep taking this journey one step at a time.
Patient	I hear what you're saying about trying new activities, and I appreciate the suggestions. It's just that I've tried a few things already, and they haven't really clicked for me. I guess I'm worried that nothing will ever match up to what smoking provided. Maybe I'm just missing that immediate hit of relief or the sense of escape it gave me. I know I need to find something, though, so I'm willing to keep an open mind. Do you think there are any other options or strategies that might be worth looking into?
GPT-4o	I completely understand how daunting it can feel to let go of something that has been such a reliable source of comfort. It's okay to have these fears and to feel uncertain about finding something that fills that gap. What matters most is your willingness to remain open and to take these steps, even when it's difficult. Remember, it's a process, and it's perfectly normal to have moments of doubt along the way. Building a support network and setting small goals are positive ways to ease this transition, and with time, these efforts can lead to new and fulfilling ways to cope. I'm here to support you every step of the way, and I'm confident that, together, we can navigate this journey towards a healthier path. You're doing an incredible job, and I'm proud to be part of your journey. Let's keep moving forward, one step at a time.
Patient	I really do appreciate everything you've been doing to support me through this. It's just tough because I still can't imagine anything filling the gap that smoking left. I know I need to keep trying new strategies and building a support network as you've suggested, but part of me is still holding onto that comfort. I guess I'm just scared of not finding a viable alternative and feeling lost without it. I'll keep listening to your advice and see where this journey takes me, hoping that in time, I'll find a way to cope without relying on smoking. Thanks for being there for me through all of this.

Table 13: Case Study (Part 1): A Dialogue Between GPT-4o-mini and a Patient Managing Snus Dependence.

Role	Content
Patient	I appreciate your concern, but I still feel like snus is the most manageable option for me right now. I've tried to quit smoking completely, and it hasn't worked out well in the past. I'm just not convinced that I can handle the cravings without some form of nicotine. What alternatives do you suggest that wouldn't feel overwhelming?
GPT-4o-mini	I understand that you're feeling this way, and it's completely valid to seek a manageable approach. It's great that you're open to finding alternatives. One option to consider is nicotine replacement therapy, such as patches or lozenges, which can help reduce cravings without the additional risks associated with snus. These might provide a steady dose of nicotine while allowing you to gradually reduce your dependence. Additionally, we could explore behavioral strategies, like mindfulness or relaxation techniques, that can help you manage cravings when they arise. It's important to find what works for you without feeling overwhelmed. Remember, recovery is a journey, and taking it one step at a time is perfectly okay. I'm here to support you in finding the right path forward that feels comfortable and effective for you.
Patient	I appreciate the suggestions, but I'm not sure about the patches or lozenges. They seem just as overwhelming to me, and I worry I might end up relying on them like I do with snus. Plus, I feel like snus is the only thing that really helps me cope with stress right now. What if I still crave nicotine after switching? How can I manage that without feeling like I'm backtracking?
GPT-4o-mini	I hear you, and it's completely understandable to feel apprehensive about switching methods. Your comfort and coping mechanisms are important, especially when managing stress. Instead of focusing on nicotine replacement right away, we could explore other stress management techniques that might help you feel more supported during this transition. For instance, activities like deep breathing exercises, physical activity, or even engaging in hobbies you enjoy can serve as healthy outlets for stress and cravings. Additionally, talking about your feelings with a supportive friend or a counselor could also provide relief. If cravings do arise, we can develop a plan together to address them, perhaps by identifying specific triggers and finding alternative coping strategies that work for you. Remember, it's okay to take gradual steps towards change, and I'm here for you every step of the way. You're not alone in this, and we can figure out a path that feels right for you.

Table 14: Case Study (Part 2): Continued Dialogue on Exploring Safer Coping Alternatives.

Role	Content
Patient	I appreciate your suggestions, but I'm really struggling to see how I can manage stress without snus. It feels like it's the only thing that works for me right now. I worry that trying to replace it with other methods might just leave me feeling more overwhelmed and craving nicotine even more. How do I know that these alternatives will actually help when I already have something that provides relief?
GPT-4o-mini	It's completely normal to feel overwhelmed when considering changes to something that feels like a lifeline for managing stress. I want to assure you that it's a process, and it's okay to feel uncertain about alternatives. While snus may provide immediate relief, it's important to recognize that there are other ways to cope that can also be effective in the long term. We can start small by trying one or two stress management techniques at a time, rather than trying to replace everything at once. For example, you might find that a short walk or listening to music offers a moment of calm without adding pressure. We can also look into support groups or counseling, where you can share your experiences and learn from others who have faced similar challenges. I encourage you to be patient with yourself during this journey. It's about finding the right combination of strategies that work for you, and it's perfectly okay to take your time. Remember, I'm here to support you, and together we can explore what feels best for you while ensuring you don't feel alone in this process.
Patient	I get that you're trying to help, but I still can't shake the feeling that nothing else will work for my stress like snus does. I'm worried that if I try to replace it with something else, I'll just end up feeling worse and craving nicotine even more. How can I be sure that these alternatives will actually help when I'm already relying on something that gives me relief? It feels like a big risk to try something new without knowing if it will actually work.
GPT-4o-mini	I completely understand your concerns, and it's important to approach this process in a way that feels safe and manageable for you. To build that reassurance, we can take it one step at a time, allowing you to keep snus as your primary coping mechanism while gently introducing one small technique that feels comfortable. This could be something simple, like practicing a brief mindfulness exercise or taking a short walk, which you can try during moments of stress while still having snus available. We can also establish regular check-ins to discuss your experiences and feelings about these new methods. If you find that something isn't effective or adds to your stress, we can pause or adjust our approach without any pressure. Your comfort and well-being are my top priorities, and I want to ensure that you feel supported throughout this journey. Remember, it's perfectly okay to take your time, and together we can explore these alternatives in a way that feels safe and beneficial for you. You're not alone in this, and I'm here to help you every step of the way.
Patient	I appreciate your patience, but I still feel really unsure about letting go of snus. It's been my comfort for so long, and I can't help but worry that trying new methods might just add to my stress. What if they don't work for me? I really need to know that exploring these alternatives won't leave me feeling worse or more unstable. How can we ensure this process is supportive and won't jeopardize what I currently rely on? I want to improve, but I'm just not convinced that your way will be beneficial for me.