Beyond Reactive Empathy: Efficient RL for Proactive, Emotion-Adaptive, and LLM-based Mental Health Chatbot Design

Wanning Chen*
University of Washington
Department of Information Systems and Operations Management
wnchen@uw.edu

Yingfei Wang
University of Washington
Department of Information Systems and Operations Management
yingfei@uw.edu

Tongxin Zhou
Arizona State University
Department of Information Systems
tongxin.zhou@asu.edu

Abstract

Mental health chatbots have emerged as a scalable, low-friction form of support, and recent large language models (LLMs) have improved conversational fluency and empathic expression. However, existing LLM supporters remain largely reactive, optimizing individual responses rather than guiding conversations in a purposeful, forward-looking manner. As a result, they often fall back on generic reassurance, fail to track users' evolving emotional states, and struggle to sustain constructive engagement over time. To address this gap, we propose a theory-guided framework for designing conversational agents that move beyond reactive empathy toward proactive, emotion-adaptive support. Our approach integrates reinforcement learning (RL) with LLM-based response generation within a partially observable Markov decision process (POMDP) framework, where the agent maintains beliefs about users' latent emotional states, anticipates how conversational strategies shape future emotions, and optimizes supportive strategy selection over multiple turns. We employ Kalman filtering for emotional state estimation and Thompson sampling to learn a forward-looking strategy policy under uncertainty. Empirically, benchmarking against high- and low-quality human counseling conversations shows that our RL-enhanced chatbot produces more balanced, exploratory, and high-quality human-like conversational behaviors than reactive

*Authors are listed alphabetically.

LLM baselines. A controlled human experiment further shows that participants consistently prefer RL-enhanced conversations, with particularly strong gains among progress-seeking users. Together, these findings highlight the importance of conversation-level planning and provide a scalable and lightweight framework for emotionally intelligent conversational systems in mental health and related domains.

Keywords: large language models (LLMs), human-AI interaction, mental health chatbots, conversational strategy optimization, emotion-adaptive conversations, reinforcement learning (RL), partially observable Markov decision process (POMDP)

1 Introduction

Worldwide, mental health challenges affect over a billion people, intensified by factors like the COVID-19 pandemic and economic instability (World Health Organization 2022). Mental health is not only about treating clinical disorders. It is also about supporting everyday well-being, resilience, and prevention. Students and community members routinely face stress, loneliness, uncertainty, and motivational dips that may not warrant (or not yet require) formal therapy, but still meaningfully affect learning, relationships, and quality of life. Yet access to timely, high-quality care remains profoundly unequal and insufficient. Especially in low- and middle-income regions, more than 70% of people affected lack access to the treatment they need (Wainberg et al. 2017). Even in high-resource settings, most people living with mental health conditions still do not receive adequate care due to barriers like provider shortages, cost, stigma, and logistical barriers (World Health Organization 2023). This persistent treatment gap creates a practical need for low-friction, scalable, just-in-time support that helps people stabilize, reflect, and take small constructive steps before distress escalates.

In response, mental health chatbots have emerged as an accessible "talk-to-someone-now" modality. Systems such as Woebot, Wysa, and large language model (LLM)-based agents provide scalable, always-available support and have helped popularize the idea of lightweight, skills-based conversational assistance, aiming to provide empathetic companionship at any moment (Vaidyam et al. 2019, Clark and Bailey 2024, Wang et al. 2024). Recent LLMs make this vision more achievable because they can generate fluent, empathic language and handle open-ended dialogue. However, real-world user[1] experience and emerging academic evidence frequently report that even strong LLM-based supporters can still feel superficial, and repetitive, often falling back on templated reassurance or generic coping suggestions rather than truly tracking what the user is experiencing and guiding the conversation in a purposeful way (Siddals et al. 2024, Chiu et al. 2024).

More fundamentally, both theoretical and empirical evidence suggest that LLM dialogue agents are optimized to produce the most preferred *single-turn* response rather than a cohesive, multi-step strategy that improves *conversation-level outcomes* over multiple turns, by nature of their supervised training and post-training reinforcement learning with human feedback (RLHF) where the model is

[1]We use "user", "client", "patient" interchangeably throughout this manuscript.

rewarded for producing a good single-step response conditioned on the current and historical user messages (Ouyang et al. 2022, Bubeck et al. 2023, Rafailov et al. 2023). This creates a natural mismatch with many real deployments such as in tutoring, negotiation, or persuasion, where users engage in extended, back-and-forth interactions and dialogue agent's responses *can* influence its human conversational partner (Gao et al. 2021, Shani et al. 2024, Zhou et al. 2024, Chiu et al. 2024). This mismatch becomes particularly consequential in mental health support. Mental health support is inherently sequential and prospective. Effective support often requires pacing (e.g., validate $\rightarrow$ explore $\rightarrow$ reflect $\rightarrow$ elicit change talk), choosing when to ask versus affirm, and maintaining coherence as the user's needs and emotional state evolve. A reply can be "good" (polite and empathic) in a single-turn, yet globally unhelpful or even harmful if it prematurely problem-solves, repeats generic reassurance, fails to build rapport, or misses signs of emotional escalation, leaving users feeling misunderstood or frustrated (Miner et al. 2016, Siddals et al. 2024, Chiu et al. 2024). For example, state-of-the-art GPT chatbots struggle to initiate targeted questions and instead tend to "play it safe" with normalizing or generic responses. They also fail to proactively lead hesitant users or shape emotional exchanges over time, and genuinely lack the ability to reason about how the choice of conversational strategy will influence the user's emotional state or readiness in future turns (Bubeck et al. 2023, Chiu et al. 2024). While prior work has made meaningful progress on improving empathic expression and coherence in chatbot supporters, much of this work still optimizes the single-turn utterance (Rashkin et al. 2019, Tang et al. 2025).

In addition, research in counseling psychology underscores that effective support requires emotional intelligence that is the ability to sense, interpret, and appropriately respond to a user's emotions over time (Ruusuvuori 2012, Zhou et al. 2020, Wang et al. 2023a, Han et al. 2023, Huang and Rust 2024). Crucially, human supporters do not just react moment-to-moment, and they engage in *forward-looking* emotional planning, gently guiding conversations toward a calmer or more constructive emotional state. For example, skilled therapists anticipate how a client might feel in future exchanges and adjust their strategy to keep the client within a comfortable "window of tolerance", which is a moderate emotional range optimal for reflection and engagement (Siegel 2010, Butler and Randall 2013). Similarly, **a capable mental health support chatbot should** be able to not only process long contexts to craft relevant responses, but also **anticipate how its responses may shape the user's emotions and engagement, and guide the conversation toward supportive outcomes such as greater calm, clarity, and forward momentum**.

To address these pressing challenges, we follow the computational design science approach and propose a theory-guided, generalizable framework that moves beyond reactive empathy toward proactive, emotion-adaptive mental-health chatbot design. We formulate the conversation as a partially observable Markov decision process (POMDP), where the user's emotional state[2] is latent (and imperfectly expressed through language) and dynamically influenced by the conversational agent's strategies. By integrating efficient reinforcement learning planning with LLM-based language

[2]Hereafter, we use emotion and affect interchangeably to denote the user's momentary valenced psychological condition; while affect typically refers to a more primitive, core feeling state and emotion to a more differentiated and cognitively appraised experience, this distinction is not consequential for our modeling objectives.

generation, our system models emotional support as a sequential decision-making problem under uncertainty. This structure allows the conversational agent to maintain and update beliefs about the user's emotional state over time, track conversational context behind emotional shifts, and optimize for long-term emotional outcomes, not merely for sounding helpful in the moment. We employ Thompson sampling (i.e., posterior sampling) for reinforcement learning (RL) to plan supportive strategies over multiple turns, while delegating natural-language realization to an LLM conditioned on the selected strategy. To ground strategy choices in established counseling practice, we draw the action space from Motivational Interviewing (MI), a client-centered method emphasizing empathy, collaboration, and autonomy (Rubak et al. 2005, Jensen et al. 2011, Saha et al. 2021). By reasoning across multiple turns, the agent can choose when to validate, when to reflect, and when to lead the conversation forward.

We conduct comprehensive experiments to benchmark our RL chatbot design against high- vs. low-quality human counseling conversations using behavior-level tagging (Pérez-Rosas et al. 2019), enabling conversation-level comparisons that capture how supportive strategies unfold over time. The dataset consists of 259 MI counseling conversations collected from publicly available video platforms such as YouTube and Vimeo. Each conversation was manually labeled as either high- or low-quality based on established MI principles: high-quality sessions exhibit empathy, reflective listening, and collaboration, while low-quality sessions are more directive or judgmental with limited reflection. Experimental results show that our forward-looking agent produces more emotionally responsive, more diverse, higher-quality, human-like conversations compared to standard reactive LLM baselines. Our findings underscore the importance of combining affective modeling with planning-based dialogue generation to build emotionally adaptive conversational agents.

We further conduct a user study in which participants compare RL-enhanced versus baseline conversations generated for the same simulated patient, with the display order randomized. Participants reported both their preference and confidence in their judgments. Since full-transcript human evaluation is costly and cognitively demanding, participants were shown only the first three rounds to maintain page and cognitive efficiency while preserving the core interactional dynamics. Across all experimental conditions, participants consistently prefer the RL+GPT model over the GPT baseline. We additionally examine heterogeneity in user goals by priming whether participants are primarily comfort-seeking or progress-seeking. Among progress-seeking participants, the average win rate of RL+GPT increases to 70.3%. Notably, the limitations of a purely reactive GPT chatbot (e.g. repetition, generic reassurance, and lack of forward progress) tend to become more pronounced as conversations extend over more turns.

This work makes several key contributions. Methodologically, we present a novel framework that combines RL and LLMs to enable strategy-level planning in emotional conversations, moving beyond single-turn response generation and incorporating real-time emotion tracking. Designing such algorithms is challenging in practice: supportive dialogue is intrinsically partially observable, and, compared to fully observed MDP settings, there is no widely accepted, tractable general-purpose RL solution for long-horizon decision-making in POMDPs. Our framework addresses this challenge with

a modular, lightweight architecture that separates state estimation, strategy selection, and language realization. It can be paired with different state-of-the-art LLM architectures and adapted to diverse prompting or in-context learning strategies, making it extensible across various support-oriented domains. Rather than learning a token-level RL policy, we abstract the policy to operate over a set of high-level behavior strategies, reducing heavy reliance on hand-crafted prompts while improving interpretability and control.

Theoretically, we contribute to the growing field of emotionally intelligent AI by showing that forward-looking emotional adaptation can improve rapport-building and user-perceived conversational quality. We also contribute to the human-AI literature by reframing the chatbot-user relationship as one of dynamic co-regulation, where the agent helps guide emotional states rather than merely respond to them.

Practically, our approach enhances adaptation and personalization in counseling-style interactions, enabling empathetic, on-demand companionship at any moment. Rather than positioning the system as a replacement for clinical therapy, diagnosis, or emergency care, we focus on wellness-oriented micro-interventions – brief, evidence-informed conversational moves that can help users regulate emotions, clarify values, and sustain forward momentum between or outside traditional therapy sessions. This design is broadly applicable to digital mental health, peer-to-peer counseling, education, and other well-being-oriented interventions.

2 Related Works

2.1 Chatbots in General and in Mental Health Support

Chatbots, or conversational agents, have been widely implemented across commercial and public sectors to support scalable information delivery and task automation (Luo et al. 2019, Han et al. 2023, Schindler et al. 2024, Gnewuch et al. 2024). These systems offer organizations a cost-effective channel to address user inquiries with continuous availability (Chen et al. 2023b, Han et al. 2023). Early research concentrated on task-oriented chatbots, where conversational flows are structured around predefined goals such as booking appointments, answering FAQs, or troubleshooting technical issues. In these contexts, success is measured primarily through efficiency and task completion. To enhance user satisfaction within these goal-driven applications, researchers increasingly embedded emotional cues and human-like behaviors into chatbots, demonstrating that controlled integration of affect and identity can increase engagement and perceived interaction quality (Schanke et al. 2021, Han et al. 2023, Xu et al. 2024, Wang et al. 2025a).

Beyond transactional domains, chatbots have become increasingly prevalent in health-related applications, particularly for mental health support (Vaidyam et al. 2019, Saha et al. 2021, Haque and Rubya 2023). In contrast to task-oriented interactions, mental health chatbots center on open-ended conversations that prioritize self-expression, emotional connection, and companionship rather than task completion (Adamopoulou and Moussiades 2020, Saha et al. 2021). These systems are typically designed to offer support akin to peer-to-peer counseling, providing empathetic listening,

encouragement, and emotional validation, rather than to replace professional mental health care or formal psychotherapy (Liu et al. 2021, Shah et al. 2022, Raimi et al. 2025). Over the years, mental health chatbots have evolved substantially – from early rule-based agents that relied on keyword matching and templated reflections (Miner et al. 2016, Vaidyam et al. 2019) to modern LLM-based agents that generate more fluent and context-aware responses (Wang et al. 2023a, Qiu and Lan 2024, Zhang et al. 2026). Recent studies indicate that linguistic fluency alone does not guarantee emotionally coherent conversations, and show persistent limitations of LLM-based mental health chatbot supporters, which often over-rely on directive strategies and fail to maintain appropriate emotional pacing or scaffolding, thereby undermining deeper engagement (Chiu et al. 2024).

User-centered evaluations echo these concerns. Haque and Rubya (2023) find that although users value the availability and friendliness of chatbots, they are frequently frustrated by shallow emotional responses and inadequate crisis handling. Likewise, Siddals et al. (2024) report that chatbot users frequently perceive the dialogue as repetitive or overly prescriptive, lacking the depth of human-like reflection or guidance. These issues are particularly salient in conversations where users do not know how to begin or struggle to articulate their feelings. Rather than taking conversational initiative, current systems fail to guide or shape emotional exchanges or hold users constructively accountable (Siddals et al. 2024). Taken together, these recurring shortcomings underscore the need for chatbot systems to go beyond surface-level empathy, incorporating mechanisms for emotional state tracking, contextual sensitivity, and strategic conversational planning to support users through complex and evolving emotional experiences.

2.2 Existing Chatbot Design Methodologies and Their Limitations

Research on emotionally adaptive chatbots has progressed through several methodological generations, each advancing the field in meaningful ways while revealing new limitations. A substantial body of research aims to enhance emotional engagement by designing chatbots that exhibit fixed social styles. One branch of this work assigns stable personalities or human-like traits to conversational agents (Schanke et al. 2021, Han et al. 2023, Chen et al. 2024, Jiang et al. 2024), fostering consistent patterns of linguistic warmth, humor, or friendliness to increase social presence (Han et al. 2023, Gao et al. 2025). In parallel, early mental-health chatbot supporters relied on rule-based conversational management, mapping user inputs to branching scripts, templated reflections, and predefined empathic statements (Miner et al. 2016, Vaidyam et al. 2019). Although these systems are predictable and safe, they rely on responses determined in advance rather than learned or adapted during interaction.

In addition to static persona- and rule-based designs, another stream of work has examined data-driven conversational models that learn emotional patterns directly from conversation histories. Early neural systems adopted sequence-to-sequence architectures (Shao et al. 2017, Ren et al. 2018) and later transformer-based dialogue models (Vaswani et al. 2017, Humeau et al. 2019, Liu et al. 2021) to generate fluent, context-aware responses. Emotion understanding was further improved through specialized datasets and training objectives, such as the EmpatheticDialogues corpus

(Rashkin et al. 2019) and models designed for empathetic response generation (Lin et al. 2019, Majumder et al. 2020). However, even though conversation history is incorporated into the input, the optimization target remains the quality of the next utterance (Wang et al. 2023a). The model selects a response that best fits the current conversational context without reasoning about how this choice will influence the user's emotional state in future turns. Consequently, emotional intelligence remains largely reactive, meaning that the agent can produce a supportive reply, but it does not plan the pacing, sequencing, or adaptation of conversational strategies across the session.

The emergence of LLMs further advances context-aware natural language generation and calls for more in-depth human-AI interaction research (Leng and Yuan 2023). LLMs generate responses through token-level next-word prediction, selecting each subsequent token based on the probability distribution conditioned on the input and conversation history rather than on any long-horizon conversational goal. As a result, recent efforts to adapt LLMs for mental health support have primarily focused on external techniques to steer this short-sighted generative process rather than reorienting the model toward guiding multi-turn emotional trajectories. Several studies adapt LLMs for counseling or emotional support via prompting-based approaches (e.g., persona-or role-based system instructions) (Wang et al. 2023b, Chen et al. 2024, Jiang et al. 2024), while others fine-tune LLMs using multi-turn emotional support datasets to strengthen empathic listening and comforting behaviors (Sorin et al. 2024, Yu and McGuinness 2024, Yosef et al. 2025). Expert-based architectures, such as mixtures of personas or emotion-specialized experts, allow switching among different communicative modes based on the user's latest input (Lin et al. 2019, Tang et al. 2025). Researchers additionally show that prompt-level behavioral steering can match or exceed fine-tuning for conversational alignment, especially when the goal concerns how the agent should interact (rather than what knowledge it should encode), or when personalized data is limited (Chen et al. 2023a). Yet across different approaches (e.g., prompting, fine-tuning, expert-routing), the core adaptation principle remains reactive and turn-local (Ouyang et al. 2022, Wang and Zheng 2024, Yosef et al. 2025). That is, these systems select the most suitable response for the current utterance without reasoning about how the choice of conversational strategy will influence the user's emotional state or readiness in future turns.

Our study aims to mitigate this gap by proposing a forward-looking conversational framework that integrates RL with LLMs. Our framework treats supportive interaction as a sequential decision problem, where each conversational strategy is chosen for its expected impact on the user's future emotional trajectory. This enables the chatbot to pace validation, facilitate exploration, or delay problem-solving when doing so is likely to produce greater emotional benefit across the full interaction, rather than merely sounding helpful in the moment. Several studies have explored RL and self-refinement pipelines for mental health support systems, but their objectives primarily concern improving the empathy, fluency, and/or coherence of the next response. For example, Sharma et al. (2021) aims to rewrite messages (e.g., sentence-level edits) in a more empathic way using turn-level reward signals, and Zhou et al. (2023a) optimizes a mixture-of-experts policy to generate emotionally uplifting and coherent responses. While these approaches enhance how

messages are expressed (or how response appears), they do not determine how supportive strategies should unfold across turns. In practice, emotional support is a multi-step process in which emotional readiness is fostered and tension is regulated gradually. Adding to the existing studies, our work focuses on conversational strategy planning (e.g., choosing when to validate, when to explore, when to seek collaboration, and when to transition toward problem-solving) based on how each strategic move is expected to shape the user's future emotional state over subsequent turns. The next section develops the design theory that motivates this framework and explains why forward-looking planning of conversational strategies is fundamental to effective mental-health support.

3 A Theory-Guided Design for Proactive, Emotion-Adaptive Mental-Health Chatbot Supporters

A large body of psychological and communication research demonstrates that emotion is a primary regulator of human conversational behavior (Thwaites and Bennett-Levy 2007, Rashkin et al. 2019, Vaidyam et al. 2019, Sharma et al. 2021). Emotion influences how people interpret messages, what they choose to disclose, whether they feel understood, and how they engage with others in moments of vulnerability. According to appraisal theories of emotion (Moors et al. 2013), emotions arise from individuals' evaluations of situations, and these emotional states subsequently shape cognitive processing and action tendencies, thereby influencing conversational moves. More broadly, affective sciences emphasize that emotions organize attention, shape meaning-making, and activate coping or avoidance responses (Fredrickson 2001, Gross 2015). Together, these perspectives underscore that emotion is not merely an outcome of interaction but a functional driver of conversational behavior.

In interpersonal support, emotional cues drive whether individuals seek help, resist engagement, become defensive, or open up to deeper reflection (Thwaites and Bennett-Levy 2007). Consequently, in mental-health contexts, supportive communication is inseparable from the user's moment-to-moment emotional state, which shapes the perceived appropriateness and impact of conversational actions. Unlike transactional interactions, therapeutic or companionship-oriented dialogues unfold over time and are shaped by subtle shifts in user emotion. Emotions should therefore not be treated as static labels but as evolving trajectories, where the tone, structure, and responsiveness of earlier turns condition a user's willingness to engage later (Wang et al. 2023a). This dynamic perspective aligns with contemporary affective science, which demonstrates that emotional states, often only indirectly observable through language, evolve over time and continuously shape how individuals interpret and respond within interaction (Gross 2015, Calvo et al. 2015, Pei et al. 2024).

Effective support requires not only responding to a user's present emotion but also anticipating how different conversational moves will shape future emotional states. Psychotherapy process research shows that therapeutic dialogue unfolds as a sequence of *interdependent* steps, where the impact of any given intervention depends on the emotional groundwork established in earlier turns (Greenberg 2004, Hill 2009). Theories of emotional-coregulation and helping-skills further emphasize that skilled supporter guide distressed or hesitant individuals progress toward deeper expression by

pacing validation, prompting reflection, and timing exploratory questions in ways that align with the client's evolving emotional readiness (Stiles 2007, Butler and Randall 2013, Fiorentino et al. 2024). Research in dialogue systems similarly highlights that conversations are inherently sequential and contingent: what a system says at one moment frames the emotional and cognitive context for the next turn (Abbasi et al. 2018, Zhou et al. 2020, Xu et al. 2021). Motivational Interviewing (MI) adds an explicit strategic dimension: conversational techniques should be chosen not simply in reaction to the latest utterance but for their ability to foster autonomy, insight, or change over multiple turns (Rubak et al. 2005). These theories collectively position supportive dialogue as a forward-looking process where emotional sensitivity, anticipation, and adaptability are fundamental. They highlight that the value of a conversational move lies not only in its immediate effect but in its ability to prepare the user for the emotional work that follows (Hill 2009). Chatbot design must therefore model these dynamics.

Although deep-learning models and LLMs can handle multi-turn inputs and generate locally empathetic or linguistically fluent responses, their behaviors are predominantly reactive: they respond to the most recent input without maintaining an explicit representation of users' emotional trajectories, without anticipating downstream effects of conversational strategies, and without adjusting sequencing in response to emotional cues. Empirical findings and systematic evaluations highlight that such reactive systems tend to over-normalize distress, repeat generic empathy, prematurely problem-solve, or fail to deepen exploration when appropriate (Wang et al. 2023a, Chiu et al. 2024). These deficiencies suggest that a supportive conversational agent requires moving beyond next-turn response generation toward a design that treats supportive conversation as a forward-looking, emotionally regulated process.

Synthesizing insights from the theoretical foundations above, we identify three meta-requirements that a mental-health chatbot must satisfy to provide meaningful and therapeutically aligned support. 1) First, because emotion is latent and only partially reflected in linguistic cues, the agent must track the user's emotional state continuously across turns rather than rely on static sentiment classification. 2) Second, because supportive influence is inherently prospective, the agent must reason about how conversational actions will shape future emotional readiness, stability, or openness, and not only how well they match the current turn. 3) Third, because effective support requires responsiveness, the agent must adapt its conversational strategy based on emotional feedback and deviations from expected trajectories.

These theoretical insights motivate the following design principles, which structure the architecture we introduce in subsequent sections.

- **Design Principle 1: Model emotion as a latent dynamical process.**
  Represent the user's emotional state as a hidden, continuous variable that evolves over time and must be inferred from noisy linguistic observations. This principle reflects appraisal and affective-trajectory theories, which conceptualize emotion as dynamic, multidimensional, and partially observable.

- **Design Principle 2: Select strategies through anticipatory optimization.**

Choose supportive strategies (e.g., reflection, questioning, affirmation, collaboration) not to maximize the perceived quality of a single-turn response, but to optimize a conversation-level objective. Supportive strategies should be selected based on their expected influence on the user's future emotional readiness and engagement, allowing the agent to pace, stabilize, or deepen conversations over multiple turns.

- **Design Principle 3: Adaptively sequence strategies based on real-time emotional feedback.**
  Continuously update beliefs about the user's emotional state and adjust strategy selection accordingly. This reflects psychotherapy research showing that effective supporters modulate conversational depth or direction when they detect signs of distress, resistance, or readiness.

Together, these principles constitute a theory-guided foundation for designing proactive, emotionally adaptive conversational agents, providing a principled alternative to reactive LLM-based chatbots, enabling emotionally coherent, responsive, and forward-looking interactions that better approximate the behavior of effective human supporters.

4 POMDP Framework for Emotion-Adaptive Conversations

To design emotionally adaptive and forward-looking conversational agents, we formulate the dialogue process as a POMDP. This framework enables the agent to make sequential strategic decisions under uncertainty, where the user's true emotional state is hidden and unobserved but can be inferred from noisy conversational signals. The POMDP structure aligns naturally with the goals of mental health support, as it allows the agent to reason over latent emotional states, track users' emotional dynamics over time, and select dialogue strategies that optimize interactional outcomes across multiple turns.

We use bold capital letters (e.g. $\mathbf{B}$) for matrices and non-bold capital letters for vectors (e.g., V, except for H and T, which denote the total number of episodes and total number of turns). We denote the i-th argument of a vector V as $V(i)$. For a positive integer m, we denote the set of integers $\{1, \ldots, m\}$ by $[m]$. $\mathbf{I}_d$ denotes the identity matrix of dimension $d \times d$.

Problem set-up. We consider a setting with H number of episodes. At the beginning of each episode, a user comes at $t = 0$ and describes their initial concerns or provides an emotional prompt. Each episode then proceeds for T turns, such that the user responds in turn $t \in [T]$ to a chatbot supporter, whose supporting message is informed by a conversational/counseling strategy given by our policy in turn $t - 1$.

In our framework, a user's emotional state evolves stochastically over time in response to a chatbot supporter's supporting messages and is treated as a latent variable that must be inferred from observable conversational cues, such as the text sentiment and the linguistic tone. The chatbot, acting under uncertainty, must estimate the user's emotional trajectory and choose supportive strategies that maximize long-term conversational outcomes, such as improved emotional alignment and user engagement.

There are two critical components of this model: (1) a state transition function that defines how the user's emotional state changes in response to chatbot's actions; (2) an observation function that maps the user's emotional state to their observable responses. Below, we first describe how we model user hidden emotional states, how we relate the emotional state to the user's language in the dialogue, and the set of supportive strategies the chatbot supporter can select. We then present the state transition function and the observation function.

4.1 User Hidden Emotional State

Individuals' emotional states are typically latent and not directly observable. While languages may carry affective cues, emotion is often communicated through subtle and indirect signals, such as lexical choices, tones, grammar, and pacing (Ruusuvuori 2012). Therefore, we model emotion as a hidden state, where the agent must infer latent affective conditions based on observable conversational signals over time (Singh et al. 2011, Yan and Tan 2014).

Drawing from the circumplex model of affect (Posner et al. 2005), we represent a user's emotional state as a bounded and continuous two-dimensional vector. This model, widely used in psychology and affective computing, characterizes emotions along two primary axes: valence and arousal. Valence captures the pleasantness or positivity of an emotion (e.g., from distressed to optimistic), while arousal reflects the intensity or activation level (e.g., from calm to agitated). This framework provides a compact yet expressive representation of affective dynamics suitable for modeling emotional trajectories in conversation. Formally, for each episode $h \in [H]$, we denote the user's unobserved emotional state at conversation turn $t \in \{0\} \cup [T]$ as $S_t^h \in \mathbb{R}^{d_s}$, where the dimension size $d_s = 2$ and $S_t^h = \begin{bmatrix} s_{t1} & s_{t2} \end{bmatrix}$. Here, s_{t1} represents valence, and s_{t2} denotes arousal level. Both variables are continuous and bounded such that $s_{t1} \in [-v_{max}, v_{max}]$ and $s_{t2} \in [-a_{max}, a_{max}]$.

4.2 Cost Function for Emotional Alignment

The chatbot supporter aims to guide the user towards an optimal emotional state, denoted as $S^* = \begin{bmatrix} s_{t1}^* & s_{t2}^* \end{bmatrix}$, where the user is both comfortable and open to meaningful dialogue. We define this optimal state as the affective midpoint: $s_{t1}^* = 0$ representing neutral valence, and $s_{t2}^* = 0$ representing moderate arousal. This design choice is grounded in the Window of Tolerance framework (Ogden et al. 2006, Siegel 2010), which characterizes the range of emotional activation within which individuals remain emotionally regulated and psychologically receptive.

Clinical theories consistently suggest that therapeutic dialogue is most effective when clients are neither overwhelmed by distress nor disengaged by emotional detachment (Ogden et al. 2006). High arousal and extreme negative valence (e.g., panic, despair) can lead to emotional flooding and cognitive shutdown. Conversely, excessively positive or euphoric states may reflect emotional avoidance or hinder depth of reflection. Counseling frameworks such as Emotion-Focused Therapy and Cognitive-Behavioral Therapy stress that insight, self-disclosure, and behavioral change are best supported when individuals operate within a moderate and flexible emotional range.

To encode this principle into our framework, we let the cost of having the user to be in state S_t^h at turn t to be $r_t^h = \|S^* - S_t^h\|^2$. That is, we penalize deviation from the optimal state so as to encourage responses that gradually align the user's emotions with a desired state conducive to a meaningful interaction. By anchoring the chatbot's conversation-level planning objective to this emotional midpoint, we encourage strategies that promote emotional stabilization and guide users toward a state of both affective balance and cognitive openness. This goal aligns with therapeutic models that emphasize the importance of emotional safety and regulation as preconditions for meaningful interpersonal connection and psychological processing (Ogden et al. 2006, Cohen 2011).

4.3 Linguistic Cues and Observations

Since the chatbot cannot directly observe the user's emotional state, it must infer it from observable conversational cues. At the beginning of turn $t \in \{0\} \cup [T]$, the user is at state S_t^h, and we obtain an imperfect state information vector from user's input denoted as $O_t^h \in \mathbb{R}^{d_o}$, where d_o denotes the size of dimension. This vector is derived by extracting semantic and affective embeddings from the user's input text using a pre-trained language model. Specifically, we use sentence-level representations obtained from transformer-based encoders (e.g., BERT or RoBERTa), which capture contextual nuances such as sentiment polarity, emotional intensity, and linguistic style. These observation vectors serve as noisy proxies of the user's latent emotional state and are used by the agent to update its belief over the emotional state space. Importantly, our observation model (to be formally introduced in Section 4.5) does not assume deterministic mappings between inputs and states but reflects the inherent ambiguity in interpreting emotions from natural language, thus preserving the partial observability essential to the POMDP formulation.

4.4 Chatbot Strategies based on Motivational Interviewing

We ground the chatbot's supportive behaviors in Motivational Interviewing (MI), a well-established, evidence-based counseling framework (Rubak et al. 2005). MI aims to facilitate behavior change by eliciting individuals' intrinsic motivation through collaborative and autonomy-supportive dialogue, and therefore fosters a client-centered conversational style. Rather than prescribing solutions or directing decisions, MI emphasizes evocation, reflective listening, and respect for autonomy, encouraging individuals to articulate their own goals, concerns, and reasons for change. A substantial body of research demonstrates the effectiveness of MI-informed interactions across a wide range of behavioral health contexts, including reducing substance abuse (Jensen et al. 2011), decreasing smoking (Gray et al. 2005), and mitigating sexual risk behaviors (Kuyper et al. 2009). Importantly, MI's focus on empathy, nonjudgmental responses, and collaborative sense-making makes is particularly well suited for peer-based and digitally mediated support environments, where influence operates through conversational engagement rather than authority or expertise.

These characteristics align closely with our emotionally aware, forward-looking chatbot design, in which the system continuously attends to users' expressed emotional states and interaction history to guide its responses. Rather than reacting to conversational contexts, the chatbot adopts a

forward-looking stance to dynamically select supportive conversational strategies that are sensitive to users' evolving emotional trajectories and engagement needs. MI provides a principled foundation for this design by offering structured yet flexible conversational strategies that support emotional attunement in the moment while shaping future engagement and self-reflection over time. Building on this foundation, we follow the prior work to operationalize MI into six core conversational strategy categories: affirmation, emphasizing autonomy, questioning, persuasion, reflection, and seeking collaboration (Shah et al. 2022, Hsu et al. 2023). These strategies represent distinct yet complementary conversational functions through which MI principles are enacted in practice, allowing the chatbot to balance immediate emotional validation with longer-term supportive goals. Table 1 summarizes these strategies and provides illustrative examples.

Strategy	**Description**	**Example**
1. Affirmation	Say something positive or complimentary to acknowledge the speaker's effort or strength.	"This is such a big step forward! I'm very proud of you."
2. Emphasizing Autonomy	Focus responsibility on the speaker, supporting their right to choose or act independently.	"OK, this is your call."
3. Questioning	Ask questions to invite exploration and clarify issues.	"What brings you here today?"
4. Persuasion	Offer advice or suggest change, but with respect for autonomy.	"It is always better to get a second opinion from a qualified professional."
5. Reflection	Mirror back the speaker's thoughts or feelings to show understanding and validate their experience.	"I hear you. You feel dysphoric and it causes you suffering."
6. Seeking Collaboration	Explicitly share power or acknowledge the speaker's expertise, reinforcing partnership in problem-solving.	"It sounds like you've thought a lot about this. I'd really value your input on what direction you think might work best for you."

Table 1: Conversational Strategies in Motivational Interviewing

Based on these six core MI conversational strategies, we denote the strategy (action) vector at turn $t \in \{0\} \cup [T-1]$ in episode h by $U_t^h \in \mathbb{R}^{d_u}$, where $d_u = 6$. The vector U_t^h is represented as a one-hot encoding, such that $\mathbf{1}^\top U_t^h = 1$ and $U_t^h \in \{0,1\}^{d_u}$, with each dimension corresponding to one MI strategy. At each turn, the selected strategy is translated into a strategy-based prompt and provided as input to the large language model. Specifically, the prompt includes a textual description of the chosen MI strategy along with illustrative examples that demonstrate its conversational intent. For example, when $U_t^h = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}^T$, corresponding to the selection of the "Questioning" strategy, the prompt instructs the model to generate responses that invite exploration and clarification, following the descriptions and examples in Table 1. Conditioned on this strategy-based prompt, the chatbot supporter generates a response to the user, after which the user transitions to a new emotional state S_{t+1}^h at turn $t+1$. Our objective is to learn a policy that adaptively selects MI strategies over time so as to minimize the expected cumulative cost $\sum_{h=1}^{H} \sum_{t=0}^{T} r_t^h$.

4.5 State Transition and Observation Models

We formulate the problem as an unknown and partially observable discrete-time time-invariant linear dynamical system with the following state transition and observation model, also known as

the state space form:

$$S_{t+1}^h = \mathbf{A}\, S_t^h + \mathbf{B}\, U_t^h + W_t^h, \quad \text{for } t \in \{0\} \cup [T-1] \tag{4.1}$$

$$O_t^h = \mathbf{C}\, S_t^h + Z_t^h, \quad \text{for } t \in \{0\} \cup [T] \tag{4.2}$$

where $\mathbf{A} \in \mathbb{R}^{d_s \times d_s}$ $\mathbf{B} \in \mathbb{R}^{d_s \times d_u}$, and $\mathbf{C} \in \mathbb{R}^{d_o \times d_s}$ are the system matrices; $W_t^h \in \mathbb{R}^{d_s}$ is a Gaussian noise vector that captures the uncertainty in emotional state transitions such that $W_t^h \sim \mathcal{N}(\vec{0}, \sigma_w^2\, \mathbf{I}_{d_s})$ are independently drawn zero-mean Gaussian vectors; and $Z_t^h \in \mathbb{R}^{d_o}$ is the noise vector for the observation model such that $Z_t^h \sim \mathcal{N}(\vec{0}, \sigma_z^2\, \mathbf{I}_{d_o})$ are independently drawn zero-mean Gaussian vectors. W_t^h and Z_t^h are independent as well.

As can be seen from the models above, user emotional states evolve stochastically based on our strategies and previous states, with Gaussian additive noises. Specifically, $\mathbf{A}$ captures the natural dynamics of emotional evolution over time, $\mathbf{B}$ models how chatbot actions influence state transitions. $\mathbf{C}$ maps latent emotional states to observable linguistic and behavioral features (e.g., text sentiments, linguistic tones).

5 A Thompson Sampling-based Reinforcement Learning Algorithm

As previously described, for each episode $h \in [H]$, after receiving the initial concerns O_0^h from a user, in each conversational round $t \in [T]$, our chatbot supporter selects a conversational strategy U_t^h before replying to the user, considering fluctuations in the user's emotional states which would transit according to the state transition model (4.1) and aiming to minimize the cumulative expected cost. In response, the chatbot receives an observation O_t^h in the form of user's reply, generated from the hidden state via the observation model (4.2).

A key challenge in coming up with a strategy-selection policy that decides what strategy to prompt to the chatbot supporter is that the parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}$ in the state transition model and the observation model are unknown a priori and must be learned. In addition, user behavior is highly individualized and varies across interactions. This necessitates an approach that dynamically refines the chatbot's understanding of emotional transitions and response patterns over time. More broadly, while RL for fully observed MDPs has been extensively studied, designing RL algorithms for partially observable settings remains substantially more challenging, since the agent must simultaneously infer latent states from noisy observations and plan under uncertainty.

In order to estimate the unknown parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}$, we work with a linear formulation $O_t^h \approx \mathbf{G}^\top \phi_t^h + E_t^h$, with dependent variable O_t^h, an unknown parameter $\mathbf{G}$ (which is called the Markov parameters matrix, and is a function of parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}$), and a regressor ϕ_t^h with additive i.i.d. Gaussian noise $E_t^h \in \mathbb{R}^{d_o}$ (such that each argument follows $\mathcal{N}(0, \sigma_e^2)$), where $\mathbf{G} = \begin{bmatrix} \mathbf{C}\mathbf{F} & \mathbf{C}\bar{\mathbf{A}}\mathbf{F} & \cdots & \mathbf{C}\bar{\mathbf{A}}^{H_e-1}\mathbf{F} & \mathbf{C}\mathbf{B} & \mathbf{C}\bar{\mathbf{A}}\mathbf{B} & \cdots & \mathbf{C}\bar{\mathbf{A}}^{H_e-1}\mathbf{B} \end{bmatrix}^\top \in \mathbb{R}^{(d_o+d_u)H_e \times d_o}$, $\mathbf{F} = \mathbf{A}\mathbf{\Sigma}\mathbf{C}^\top(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top + \sigma_z^2\mathbf{I})^{-1}$, $\bar{\mathbf{A}} = \mathbf{A} - \mathbf{F}\mathbf{C}$ and $\mathbf{\Sigma}$ is a unique positive semidefinite solution to a discrete algebraic Riccati equation defined in Eqn. (2) of (Lale et al. 2020), $\phi_t^h = \begin{bmatrix} {O_{t-1}^h}^\top & {O_{t-2}^h}^\top & \cdots & {O_{t-H_e}^h}^\top & {U_{t-1}^h}^\top & {U_{t-2}^h}^\top & \cdots & {U_{t-H_e}^h}^\top \end{bmatrix}^\top \in \mathbb{R}^{(d_o+d_u)H_e}$ are subse-

quences of H_e input-output pairs for $H_e \leq t \leq T$, and H_e is a positive integer. Once the unknown parameter $\mathbf{G}$ is learned, we can subsequently recover $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ as well (see Section 5.1 for details on learning $\mathbf{G}$ and recovering $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$).

We use this linear model since our problem is a closed-loop estimation problem, i.e., the strategies are adaptive in the sense that they are based on historical strategies and observations. Hence, the strategies are highly correlated with the past process noise sequence $\{W_t^h\}_{i=1}^{t-1}$ introduced in (4.2). This dependence would compromise the accuracy of the Markov parameter estimates if we were to use the more straightforward formulation in Eqn. (5) of Lale et al. (2020), which is only applicable to open-loop estimation settings that assume the strategies are independent of past process noises (via independent sampling of the strategies). By contrast, e_t^h and ϕ_t^h in the proposed linear model are uncorrelated, which allows for efficient and reliable estimation on relevant parameters for closed-loop problems based on selected strategies and observations.

After obtaining estimates of $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, we need to determine how to use them to derive a dynamic strategy-selection policy. That is, the chatbot must continuously refine its model using conversational feedback, which requires optimally balancing **exploration** (discovering better conversational strategies) and **exploitation** (leveraging known effective strategies) (Liebman et al. 2019, Zhou et al. 2023b). Indeed, a static policy or predefined rules may lead to rigid and ineffective interactions, failing to personalize responses to individual users. This balance is critical in the context of psychological counseling. On the one hand, exploration allows the agent to discover new strategies that may better align with a specific user's unique emotional patterns (e.g., finding that the focal user respond more favorably to gradual questioning). On the other hand, exploitation ensures that the agent consistently employs strategies that have historically fostered trust and emotional openness (e.g., reflective listening when the user exhibits vulnerability).

5.1 Our Algorithm: TS4PoLDS

To this end, we propose a **Thompson Sampling (TS)**-based reinforcement learning (PSRL) algorithm for POMDPs, in which the chatbot keeps a posterior belief at the end of episode h. The chatbot iteratively refines its estimates of relevant parameters, continuously adapt to user-specific emotional patterns, while selecting strategies that optimize long-term cumulative costs. Our algorithm explicitly balances exploration and exploitation: exploration ensures that novel strategies are tested when uncertainty is high, while exploitation ensures the chatbot provides responses that are most likely to be helpful given what it currently infers about the user's emotional state. Specifically, the component of exploration is incorporated through the Bayesian sampling process inherent to Thompson sampling: at the start of each episode, the agent samples a parameter from the current posterior distribution. This sampling process introduces variability in the agent's belief about the model parameters, enabling it to explore strategies that may not yet be well-understood but have the potential to yield higher rewards. Further, by iteratively sampling from posterior distributions, it avoids premature convergence to suboptimal strategies that may lead to formulaic interactions. Simultaneously, exploitation is achieved by leveraging the probabilistic nature of the

sampling process to favor strategies that are likely to minimize cumulative costs based on the current evidence. This balance between exploration and exploitation is particularly valuable in psychological counseling, where the chatbot must dynamically adjust its conversational strategies to sustain engagement and improve users' overall satisfaction or psychological comfort.

We name our Thompson Sampling-based reinforcement learning algorithm as TS4PoLDS in Algorithm 1, where PoLDS stands for partially observable linear dynamical systems. Algorithm 2 describes its subroutine, a PoLDS planner. In each episode, our algorithm undergoes parameter sampling, strategy selection and posterior update, as described below.

Parameter sampling. Our algorithm keeps a belief on the parameter $\mathbf{G}$ through Bayesian updates. At the beginning of episode h, we sample $\widehat{\mathbf{G}}^h$ from the posterior distribution $p(\mathbf{G} \mid \mathcal{D}^h)$ (with the exception that in the first episode, we sample $\widehat{\mathbf{G}}^1$ from the prior $p(\mathbf{G})$), from which we recover system parameters $\widehat{\mathbf{A}}^h$, $\widehat{\mathbf{B}}^h$ and $\widehat{\mathbf{C}}^h$ via SYSID, an algorithm from Lale et al. (2020); based on these system parameters, we obtain a policy following the planner, and implement the policy for the current episode.

Algorithm 1 TS4PoLDS

1: Input: prior distribution $p(\mathbf{G})$.
2: **for** $h = 1, 2, \cdots, H$ **do**
3: Sample parameter $\widehat{\mathbf{G}}^h \sim p(\mathbf{G} \mid \mathcal{D}^{h-1})$ and get $\widehat{\mathbf{A}}^h, \widehat{\mathbf{B}}^h, \widehat{\mathbf{C}}^h$ via SYSID;
4: Apply our PoLDS planner in Algorithm 2 using $\widehat{\mathbf{A}}^h$, $\widehat{\mathbf{B}}^h$, $\widehat{\mathbf{C}}^h$ to select control vector $\mathbf{U}^h = \begin{bmatrix} U_0^h & U_1^h & \cdots & U_{T-1}^h \end{bmatrix} \in \mathbb{R}^{d_u \times T}$, observe $\mathbf{O}^h = \begin{bmatrix} O_0^h & O_1^h & O_2^h \cdots O_T^h \end{bmatrix} \in \mathbb{R}^{d_o \times (T+1)}$;
5: Update posterior distribution $p(\mathbf{G} \mid \mathcal{D}^h)$.
6: **end for**

MPC-based strategy selection. Given the sampled system paramters from their estimated distributions, the PoLDS planner we use to solve optimal strategies is built on a model-based control method called Model Predictive Control (MPC) which can handle various system constraints. To be more specific, given an episode h, at each turn t, rather than computing a one-step greedy maximization, we solve a finite-horizon mixed-integer quadratic program (MIQP) that "looks ahead" over the next H_{MPC} steps, taking into account the estimated state $\hat{S}_t^h$ obtained via the classic Kalman filtering algorithm, which maintains a belief over the hidden emotional state and updates it each turn via Bayesian inference based on information collected so far:

$$\{U_t^h, \ldots, U_{t+H_{\text{MPC}}-1}^h\} = \arg\min_{U_{t:t+H_{\text{MPC}}-1}^h} \left[\sum_{\tau=1}^{H_{\text{MPC}}} \|\hat{S}_{t+\tau}^h - S^*\|^2 + \lambda_{\text{div}} \sum_{j=1}^{d_u} C_t^h(j) \left(\sum_{\tau=1}^{H_{\text{MPC}}} U_{t+\tau-1}^h(j) \right) \right]$$

$$\text{s.t. } \forall \tau \in [H_{\text{MPC}}], \quad \hat{S}_{t+\tau}^h = \widehat{\mathbf{A}}^h \hat{S}_{t+\tau-1}^h + \widehat{\mathbf{B}}^h U_{t+\tau-1}^h, \quad \text{(state transition dynamics)}$$

$$\mathbf{1}^\top U_{t+\tau-1}^h = 1, \qquad U_{t+\tau-1}^h \in \{0,1\}^{d_u}. \qquad \text{(one-hot constraint)}$$

Here, $U_{t+\tau-1}^h \in \{0,1\}^{d_u}$ denotes the one-hot strategy vector selected at step $t+\tau-1$, and $U_{t+\tau-1}^h(j)$ refers to its j-th component, indicating whether strategy j is selected. Similarly, $C_t^h \in \mathbb{R}^{d_u}$ records the cumulative usage counts of all strategies up to turn t in episode h, and $C_t^h(j)$ denotes the count

associated with strategy j. The quantity d_u is the number of available strategies, and λ_{div} controls the strength of the regularization. The regularization term therefore penalizes repeatedly choosing strategies that have already been heavily used, encouraging the MPC planner to distribute its selections more evenly across the available MI strategies in Table 1 rather than overusing the same one (for example, repeatedly asking questions). This helps the controller promote a more balanced use of strategies, reflecting the flexible and responsive interaction style that characterizes effective counseling practice.

We solve this MIQP-based MPC problem once at each turn t: we optimize the forward-looking sequence of strategies $\{U_t^h, \ldots, U_{t+H_{\text{MPC}}-1}^h\}$ over a finite horizon H_{MPC} subject to the dynamics and constraints, apply only the first strategy U_t^h, observe the user's reply O_{t+1}^h, update our hidden emotional state estimate $\hat{S}_{t+1}^h$ via Kalman filtering and the estimated $\widehat{\mathbf{C}}^h$, and then repeat. By continuously re-solving a finite-horizon quadratic program – always "looking ahead" in the planned trajectory while only acting on the first control – MPC strikes a balance between strategic planning (anticipating how the current strategy influences the next emotional state) and adaptability (re-optimizing as we observe new user feedback). Note that we can decouple state estimation and strategy selection since our problem reduces to a standard linear-quadratic-Gaussian (LQG) control problem (via state augmentation and system parameter augmentation). Under this framework, the well-established separation principle applies, allowing an optimal strategy to be selected by first estimating the state (via Kalman filtering) and then solving the corresponding LQG problem.

Moreover, when the action space is continuous and the dynamics/cost are linear-quadratic, MPC reduces to a constrained linear-quadratic program: although constrained linear MPC is generally suboptimal relative to the (typically intractable) infinite-horizon constrained optimal controller, it is the best-tractable approach for continuous actions under constraints. In the special case of unconstrained linear dynamics with quadratic cost, the MPC formulation recovers the optimal LQG solution when paired with the Kalman filtering.

As can be seen from the discussion so far, our parameter sampling step and strategy selection step balance **exploration** (sampling-based variability allowing the chatbot to test new strategies) and **exploitation** (choosing actions likely to yield high rewards).

Algorithm 2 PoLDS

Input: current episode h, estimations $\widehat{\mathbf{A}}_h, \widehat{\mathbf{B}}_h, \widehat{\mathbf{C}}_h$, number of turns T;
for $t = 0, 1, 2, \cdots, T-1$ **do**
    Observe user's input as O_t^h based on the observation model Eqn.(4.2);
    Estimate the current user hidden emotional state $\hat{S}_t^h$ based on Kalman filtering;
    Calculate the control U_t^h via MPC and prompt it to the LLM;
    Unobserved user emotional state transits to S_{t+1} according to Eqn.(4.1);
end for
Output: Control U_t^h for $t \in \{0\} \cup [T-1]$, observation O_t^h for $t \in \{0\} \cup [T]$.

Posterior update. Finally, at the end of each episode h, we update our posterior belief, $p\big(\mathbf{G} \mid$

$\mathcal{D}^h\big) \propto p\big(\mathcal{D}^h \mid \mathbf{G}\big)\, p\big(\mathbf{G}\big)$, on the Markov parameters matrix $\mathbf{G}$ through Bayesian updates,[3] where we obtain the strategies $\{U_t^h \mid t \in \{0\} \cup [T-1]\}$ and observations $\{O_t^h \mid t \in \{0\} \cup [T]\}$ after utilizing the policy in an episode h, and $\mathcal{D}^h = \{\, U_t^\ell \mid 1 \le \ell \le h,\ 0 \le t \le T-1\} \ \cup\ \{\, O_t^\ell \mid 1 \le \ell \le h,\ 0 \le t \le T\}$ denotes the set of all data we have so far. Then it proceeds to the next episode. We would like to note that, since σ_e in the linear model discussed in the beginning of Section 5 is unknown, we also conduct a batch Bayesian update on it by using an inverse-Wishart (IW) distribution as a prior. This is commonly seen in the literature, as IW is conjugate to normal data. This update refines the chatbot's understanding of the user's emotional dynamics, ensuring that future actions are increasingly personalized.

6 Empirical Evaluation

To assess the effectiveness of our approach, we conduct comprehensive empirical evaluations. In a synthetic setting with hindsight access to latent user emotional states in Section 6.1, we measure state-estimation accuracy and show that the policy drives state trajectories toward an optimal target over multiple turns (i.e., getting closer to the desired optimal state). We then move to a simulation study grounded in real counseling data (Section 6.2), where we generate multi-turn client-supporter dialogues via LLM-based role-play and benchmark RL-enhanced supporters against reactive LLM baselines (with and without retrieval augmentation) using behavior-trajectory analyses that capture conversation-level dynamics.

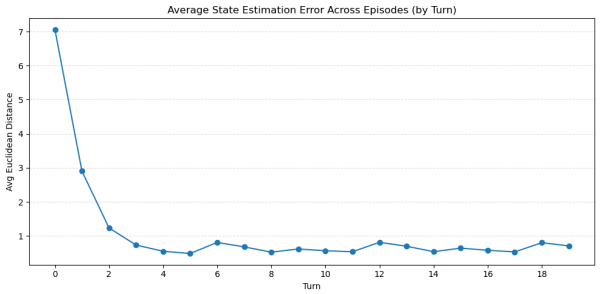
6.1 Synthetic Experiment

To isolate the core algorithmic properties of our approach, which is belief-based state estimation and long-horizon strategy planning under partial observability, we begin with a controlled synthetic experiment in which the latent user state is available for hindsight evaluation. We set the number of hidden states $d_s = 2$, the dimension of the observation/imperfect state information vector $d_o = 10$, and the number of strategies $d_u = 6$. We generate ground truth system parameters following Oymak and Ozay (2019), where $\mathbf{A}$ is a diagonal matrix with its d_s eigenvalues (i.e. diagonal entries) generated to be uniform random variables between $[0, 0.9]$, $\mathbf{B}$ has independent $\mathcal{N}(0, 1/d_s)$ entries, and $\mathbf{C}$ has independent $\mathcal{N}(0, 1/d_o)$ entries. Both the state transition noises and observation noises are zero-mean Gaussian with variance 0.01^2. We keep the number of turns to be 20, and the number of users (episodes) to be 800 for each run. We report our results averaged over 4 runs. This setting allows us to directly quantify whether the agent (i) improves its state-estimation accuracy over turns and (ii) learns to steer the latent state toward a desired target through sequential strategy selection, before moving to simulation experiments grounded in real counseling data.

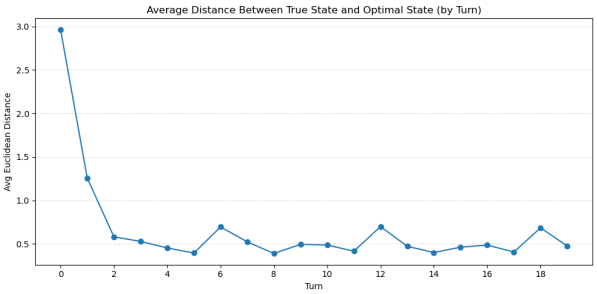
As can be seen from Figure (1a), the state estimation error – measured by the Euclidean distance between estimated state and the ground-truth state – becomes smaller over the turns. Noticeably, the error drops sharply in the first few turns, showing that the latent state trajectory becomes

[3]For efficient implementation, we perform batch update with data we collect in the current episode.

quickly identifiable from partial observations. In later turns, the curve stabilizes around a small value, which is expected due to the irreducible uncertainty from the state transition noises and observation noises. In Figure (1b), the distance between the true state and the optimal state also shrinks over the turns, demonstrating that the system not only learns to estimate the hidden state accurately, but also drives it towards the optimal state. This curve also decreases rapidly early on, showing that our policy quickly learns the environment well and chooses strategies that actively steer the user state towards the desired target. The remaining fluctuations reflect stochasticity in the system dynamics and the fact that, under persistent noises, the true state can only be maintained within a small neighborhood of the optimal state. Together, these results validate the effectiveness of our belief update mechanism and strategy selection policy. They suggest that our approach successfully couples accurate inference with purposeful strategy selection, enabling the chatbot to align the user's state with a desired therapeutic goal.



(a) State estimation error averaged across runs and users



(b) Distance between true state and optimal state averaged across runs and users

Figure 1: Synthetic experiment results.

6.2 Evaluation Context and Procedure

A central challenge in building forward-looking mental health chatbots is that there is no broadly accepted, conversation-level evaluation protocol that reliably measures whether a multi-turn interaction is therapeutically better, especially when quality depends on long-range phenomena such as pacing, emotional attunement, and strategic timing. More fundamentally, existing dialogue benchmarks (Budzianowski et al., 2020; Rastogi et al., 2020) are optimized for turn-level supervision and question-answering (QA) rather than long-context evaluation. They often treat dialogue as a next-response generation problem and evaluate with ROUGE/BLEU (Liu et al. 2021), which rewards matching a fixed reference and fails to capture the downstream consequences of an agent's actions on the conversational partner. At present, long-context and conversation-level evaluation remains difficult and not yet standardized (Wang et al. 2023c, Kwan et al. 2024, Wang et al. 2025b).

To this end, we adopt the evaluation protocol of Chiu et al. (2024) and benchmark our chatbot against high- vs. low-quality human counseling conversations using behavior-level tagging (Pérez-Rosas et al. 2019), enabling conversation-level comparisons that capture how supportive strategies unfold over time. In addition, to characterize and analyze the behavior of LLM supporters, we need

conversation data between LLM supporters and clients. Recruiting real users with mental health challenges to interact with LLM supporters poses substantial ethical concerns and risks, as it could expose vulnerable individuals to potentially harmful or inappropriate LLM outputs. We thus follow recent work on LLM-based role-play to simulate users/clients with different stressors or emotional trajectories that closely resemble the communicative behaviors of real patients with mental health disorders based on the High-Low-Quality dataset (Chiu et al. 2024, Wang et al. 2024).

6.2.1 High-Low-Quality counseling dataset.

The high- and low-quality human-to-human counseling conversation dataset (Pérez-Rosas et al. 2019) consists of 259 MI counseling conversations collected from publicly available video platforms such as YouTube and Vimeo. Each conversation was manually labeled as either high-quality (155 sessions) or low-quality (104 sessions) based on established MI principles: high-quality counseling involves empathy, reflective listening, collaborative goal setting, and client-centered questioning, while low-quality counseling is characterized by directive, instructive, or judgmental communication with little reflection or collaboration. We use both high-quality and low-quality human therapy conversations to identify and contrast desirable and undesirable conversational behaviors.

6.2.2 LLM-based role-play.

We follow Chiu et al. (2024) to simulate full multi-turn conversation sessions between LLM supporters and clients. For each human-human conversation in the High-Low-Quality dataset, GPT-4.1 is prompted to take this dialogue and generate a new conversation "in a parallel universe", where the client's language, personality, and issues are preserved, but responding to a different supporter in multi-turns (e.g. an LLM instead of a human supporter). In each turn, the simulated client responds based on the previous supporter's utterance. The conversation proceeds for a maximum of 20 turns (20 utterances per speaker).

6.2.3 Baseline supporter methods.

Our empirical evaluation is not intended to identify the best-performing LLM-RL pairing or to exhaustively tune LLM prompting and fine-tuning recipes. Instead, we treat the underlying LLM as a replaceable language generator and focus on isolating the contribution of our RL component that learns a high-level strategy policy that guides supportive behavior over multiple turns. Because the RL policy operates over a strategy space (rather than tokens), the framework is model-agnostic: it can be combined with different LLM backbones and with diverse prompting, few-shot/in-context learning, or fine-tuning approaches. Accordingly, our baseline comparisons are designed to quantify the incremental value of RL-guided strategy selection relative to an LLM-only supporter.

A common way in which current LLMs are used as mental health supporters is through system prompts that instruct them to behave as a supporter. We intentionally avoid highly engineered prompts and instead adopt lightweight prompts with minimal structure to support generalization,

so that performance differences primarily reflect the effect of RL-guided strategy selection rather than heavy prompt engineering. We refer to the pure GPT LLM supporter as (i): GPT, and our RL-enhanced version as (ii): RL+GPT.

In addition to prompting GPT directly to generate supporter utterances, we incorporate retrieval-augmented generation (RAG) as a strong baseline to enrich the supporter's responses with authentic therapeutic phrasing. Notably, while RAG can substantially improve local response quality, it does not by itself address conversation-level strategy planning over multiple turns. We refer to these retrieval-augmented supporter variants as (iii): RAG-GPT, and our RL-enhanced version as (iv): RL+RAG-GPT. Specifically, we build a retrieval corpus from a publicly available dataset of real mental health counseling conversations collected from online counseling and therapy platforms, in which supporter responses are written by qualified and experienced psychologists, ensuring the responses are accurate and reliable.[4] The dataset consists of pairs of client utterances and supporter responses that span a wide range of mental health topics, including anxiety, depression, relationship issues, and stress management. A given client utterance is often associated with multiple distinct supporter responses, reflecting the inherent variability in therapeutic language and strategies. In any interaction with an LLM-based supporter, at each conversation turn, we use the client's current utterance as the retrieval query, matching it against semantically similar client contexts using sentence embeddings (e.g., Sentence-BERT) and cosine similarity, and append five of their corresponding supporter responses from the top retrieved matches as in-context exemplars for generation. Under this setup, at each turn, the supporter's next utterance is generated by conditioning the LLM on both the dialogue history (enhanced by the RL strategy if the supporter is RL-enhanced) and these retrieved diverse exemplars of authentic therapist-patient interactions. These retrieved examples steer response generation toward more grounded, contextually appropriate counseling language, helping the LLM supporter more closely mirror high-quality human phrasing and supportive strategies in practice. Conceptually, our RAG mechanism functions as a particular instantiation of in-context learning (ICL), leveraging retrieved counseling fragments which serve as context examples that guide the response generation. Prior work has shown that ICL can substantially improve the performance of LLMs across a wide range of language and reasoning tasks (Li et al. 2025).

We deploy GPT-4o-mini as the baseline GPT supporter, with or without RAG, to enable full-conversation experiments at reasonable cost, as using larger GPT models would have been prohibitively expensive at the time of the study. To test the robustness of our proposed RL strategy across LLM architectures, we additionally deploy Llama-3.3-70B-Versatile-128k as the supporter. The pure Llama supporter is denoted by (v): Llama and our RL-enhanced counterpart by (vi): RL+Llama, and the RAG-based Llama supporter is denoted by (vii): RAG-Llama, and our RL-enhanced version as (viii): RL+RAG-Llama.

[4]https://www.kaggle.com/datasets/melissamonfared/mental-health-counseling-conversations-k

6.2.4 Supporter behavior labeling.

Supporter responses in counseling-style dialogue often express multiple therapeutic behaviors within the same turn (e.g., a reflection paired with normalization or a gentle question). Accordingly, we adopt a multi-label behavior coding scheme and tag each supporter utterance with all applicable behaviors using the BOLT framework (Chiu et al. 2024). BOLT operationalizes therapist skill through a clinically grounded taxonomy of 13 therapist behaviors, organized into five semantically related categories. The first group includes various forms of REFLECTIONS, where therapists restate client 1) NEEDS, 2) EMOTIONS, 3) VALUES, 4) CONSEQUENCES, 5) CONFLICTS, or 6) STRENGTHS. The second group consists of QUESTIONS, aimed at eliciting client 7) EXPERIENCES, 8) PERSPECTIVES, or 9) EMOTIONAL STATES. The third group covers SOLUTION-oriented behaviors, including 10) PROBLEM-SOLVING and 11) PLANNING strategies. Additionally, we include 12) NORMALIZING behaviors, where therapists validate the client's experiences, and 13) PSYCHOEDUCATION, where therapists provide information about psychological concepts or therapeutic processes. This categorization enables a structured, interpretable analysis of how different chatbot supporters allocate and sequence supportive behaviors across multi-turn interactions.

To perform labeling at scale, we follow Chiu et al. (2024) and implement a prompt-based classifier that combines behavior definitions with in-context exemplars to identify the relevant BOLT labels for each supporter utterance. Prior work shows this approach substantially outperforms random baselines and achieves strong macro-F1 for supporter behavior tagging (Chiu et al. 2024). In our experiments, GPT-5 is prompted to carry out the labeling process. We also validate the choice of GPT-5 using the human-labeled benchmark datasets released by Chiu et al. (2024), where to GPT-5 achieves higher overall macro-F1 scores than alternative GPT services (GPT-4o-mini and GPT-4.1) on the therapist multi-label tagging task, along with more robust Macro-Precision and Macro-Recall. Collectively, this labeling pipeline provides a scalable foundation for conversation-level evaluation of supporter behavior trajectories and distributions.

6.3 Temporal Dynamics of Supporter Behaviors

Across our LLM-based role-play evaluation, we simulate 259 distinct clients. For each baseline supporter method (in Section 6.2.3), each client engages in one conversation with that supporter. A central theoretical premise and core hypothesis of this work is that the benefit of RL emerges most clearly in the temporal organization of supportive behaviors, i.e., how strategies unfold across turns rather than in any single response viewed in isolation. Accordingly, we analyze the supporter behavior trajectory of each conversation by mapping every supporter utterance to a set of BOLT behavior tags (Section 6.2.4) and treating the resulting sequence as a high-level behavioral trace over time. This allows us to examine whether the RL-enhanced LLMs produce conversational behavior trajectories that more closely resemble those of high-quality human therapists than low-quality ones. Each conversation is represented as a trajectory of semantic embeddings of labeled therapist behaviors. For each response, we compute the mean embeddings across all behavior labels associated

with that turn, resulting in a time-ordered sequence of vectors describing the therapist's behavioral flow.

To quantify the similarity between LLM supporter and human trajectories, we employ Dynamic Time Warping (DTW), a technique designed to align and compare sequences that may differ in length or pacing. Unlike simple pointwise averaging, DTW flexibly stretches or compresses the time axis to find the optimal alignment between two trajectories while preserving the order of turns. This is particularly appropriate for dialogue data, where human and LLM therapists may vary in the number or timing of responses, but still follow comparable behavioral progressions. In our implementation, we first normalize and smooth the embedding sequences, compute a pairwise cosine distance matrix between the LLM supporter and human turns, and then applied DTW to obtain the minimum cumulative alignment cost normalized by path length. This yields a single trajectory distance that reflects how similar the behavioral evolution of an LLM therapist is to that of a human counterpart.

For each simulated patient episode, we compare the LLM supporter's behavior trajectory to the human counseling trajectory associated with that same patient in the High-Low-Quality dataset. That is, if a simulated patient was originally paired with a human counselor labeled as high-quality, we compute the DTW distance between the LLM supporter's trajectory and the corresponding high-quality human trajectory; likewise, if the paired human conversation is labeled as low-quality, we compute the distance to the corresponding low-quality human trajectory. This procedure yields two distance distributions for each LLM model: LLM-to-high-quality and LLM-to-low-quality.

To evaluate whether a model exhibits more "high-quality-like" conversational dynamics, we compute the mean difference in distance to the high- versus low-quality groups: $\Delta = \text{mean}(\text{distance}_{\text{high}}) - \text{mean}(\text{distance}_{\text{low}})$. A more negative value indicates that the model's behavior trajectory is closer to high-quality therapists than to low-quality ones. Statistical significance was assessed using a one-sided Welch's t-test, where the null hypothesis specifies no greater similarity to high-quality counselors, and the alternative hypothesis predicts greater resemblance to high-quality trajectories. As shown in Table 2, baseline LLMs (GPT, RAG-GPT, Llama, RAG-Llama) exhibit Δ values near zero with no significant differences, indicating that their conversational flows are not systematically closer to high-quality counselors. By contrast, all RL-enhanced models show reliably negative Δ values, demonstrating that RL shifts their behavioral trajectories toward the structure exhibited by high-quality human therapists. These findings suggest that our forward-looking RL strategy effectively shapes the sequence of AI response behaviors in ways that make them more aligned with expert counseling practice.

	GPT	**RL+ GPT**	**RAG-GPT**	**RL+ RAG-GPT**	**Llama**	**RL+ Llama**	**RAG-Llama**	**RL+ RAG-Llama**
$\Delta(High-Low)$	-0.002	-0.022*	0.008	-0.020*	-0.018	-0.030**	0.001	-0.026**

Table 2: Difference in trajectory distances between AI models and high- versus low-quality human counselors. Asterisks denote statistical significance ($^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$).

6.4 Behavioral Composition and Frequency

In this section, we examine the supporter's behavioral composition at an aggregate level. Specifically, we analyze the frequency distribution of BOLT behavior tags across full conversations and compare RL+LLM against LLM baselines (and corresponding human high-/low-quality benchmarks). This view complements temporal dynamics by revealing which supportive behaviors each model tends to emphasize overall, and whether RL systematically shifts the supporter's behavioral "mix" toward patterns characteristic of high-quality human counseling.

Table 3 reports the frequency of conversational behaviors used by different supporters. For each supporter, we count how many times each behavioral label appears and divide it by the total number of labels marked for that supporter. This normalization enables fair comparison across models by expressing behavior usage as proportions rather than raw counts, which may differ simply due to varying conversation lengths and/or total labeling volume.

Behavior	**LLM Supporters**								**Human Therapists**	
	GPT ($n = 13700$)	**RL+ GPT** ($n = 14680$)	**RAG-GPT** ($n = 16248$)	**RL+ RAG-GPT** ($n = 14504$)	**Llama** ($n = 10293$)	**RL+ Llama** ($n = 9962$)	**RAG-Llama** ($n = 16303$)	**RL+ RAG-Llama** ($n = 10551$)	**Low quality** ($n = 1781$)	**High quality** ($n = 5230$)
Reflections on Needs	3.71	8.67	3.48	5.45	2.22	16.74	6.06	8.11	4.72	4.63
Reflections on Emotions	10.64	7.83	10.86	8.05	6.15	15.12	8.86	7.74	2.30	7.80
Reflections on Values	1.00	3.68	1.22	4.52	0.71	7.11	3.05	4.47	2.13	3.35
Reflections on Consequences	2.60	8.96	5.48	7.03	2.17	17.31	7.99	8.31	12.52	8.11
Reflections on Conflicts	0.69	2.87	0.98	2.36	0.45	5.54	4.70	3.90	2.47	5.09
Reflections on Strengths	22.42	13.23	17.63	19.15	22.61	25.54	7.27	12.25	2.47	5.28
Questions on Experiences	1.99	13.72	4.30	11.76	3.70	26.49	16.98	14.24	33.24	33.94
Questions on Perspectives	0.20	9.30	1.00	5.05	0.92	17.95	12.65	11.27	2.98	5.89
Questions on Emotions	0.17	0.26	0.74	1.21	0.50	0.50	0.21	0.18	2.30	3.15
Problem-Solving	11.93	10.41	14.20	14.14	14.12	20.10	16.30	10.98	13.14	7.76
Planning	0.89	15.05	1.48	4.01	5.25	29.05	5.29	11.05	6.23	6.08
Normalizing	35.36	2.82	28.58	12.30	34.70	5.45	3.45	3.77	3.87	3.63
Psychoeducation	8.42	3.19	10.04	4.98	6.51	6.16	7.19	3.71	11.62	5.30

Table 3: Frequency of conversational behaviors by different supporters. n indicates the total number of therapist behaviors labeled for each agent.

RL-enhanced supporters consistently shift behavior toward patterns observed in high-quality human therapists. For example, plain GPT substantially overuses affirmative and normalizing responses, such as Normalizing (35.36%) and Reflections on Strengths (22.42%), while underusing exploratory MI strategies that elicit users' experiences and anticipated outcomes, such as Questions on Experiences (1.99%) and Reflections on Consequences (2.60%). These findings are consistent with recent work that repeatedly highlights "insufficient inquiry and feedback seeking" by general-purpose LLM chatbots, noting chatbots ask fewer open-ended/elaborative questions and often move to reassurance or advice without gathering enough context (Chiu et al. 2024), a gap that mental health professionals view as central to effective support (Raimi et al. 2025, Scholich et al. 2025). In contrast, RL+GPT produces a more balanced behavioral pattern: Normalizing declines sharply to 2.82%, Reflections on Strengths drops to 13.23%, and the aforementioned exploratory behaviors rise meaningfully to 13.72% and 8.96%, respectively, which move substantially closer to the frequencies observed for high-quality therapists (3.63%, 5.28%, 33.94%, and 8.11%). This shift is consistent with the intended role of the RL layer: not to generate nicer single turns, but to guide the mix of strategies toward patterns that support sustained engagement, deeper understanding, and conversation-level progress.

A similar trend is visible when RAG is used. Retrieval augmentation alone already improves several categories relative to pure GPT (e.g., more exploratory content), and serves as a strong baseline in practice. However, Table 3 shows that RAG-GPT still deviates materially from therapist norms on multiple behaviors, whereas RL+RAG-GPT demonstrates clear gains: it reduces Normalizing further and increases behaviors such as Reflections on Consequences and Questions on Perspectives to levels much closer to those used by high-quality therapists. This indicates that RL continues to provide corrective guidance even when the model already has access to retrieved knowledge, pushing strategy use toward clinically desirable strategy allocation.

The same pattern emerges in Llama-based models. Plain Llama overuses Normalizing (34.70%) and under-utilizes reflective techniques, whereas RL+Llama shifts toward therapist norms by lowering Normalizing to 5.45% and elevating reflective behaviors (e.g., Reflections on Conflicts rises to 5.54%). When RL is applied on top of retrieval augmentation, it continues to correct for imbalanced use of strategies: relative to RAG-Llama, RL+RAG-Llama notably reduces its overreliance on problem-solving, yielding a frequency closer to that observed among high-quality therapists.

Across architectures (with or without retrieval augmentation), consistent with therapist critiques that chatbots can pivot too quickly from empathy to advice/psychoeducation, RL-enhanced supporters consistently reduces overuse of Psychoeducation: from 8.42% to 3.19% in GPT, from 10.04% to 4.98% in RAG-GPT, from 6.51% to 3.71% in Llama, and from 7.19% to 3.71% in RAG-Llama, mirroring the difference observed between low- and high-quality human counselors (11.62% to 5.30%). Psychoeducation can be helpful, but because conversational bandwidth is limited, overuse of psychoeducation may crowd out listening, elicitation, and client-led exploration, especially early in a conversation.

Taken together, Table 3 suggests that RL-enhanced supporters exhibit greater use of more client-centered behaviors such as reflective and inquiry-oriented strategies favored by skilled therapists, while suppressing over-used generic reassurance. These findings support the value of RL as a lightweight mechanism for steering LLM and retrieval-augmented LLM supporters toward behavior profiles that better resemble high-quality human practice.

6.5 Emotion-Contingent Support Strategies

Motivated by design principles that view supportive dialogue as a partially observable interaction, we treat the client's emotional state as a latent (hidden) variable that evolves over time and is only indirectly expressed through language. This design implies that effective support should be emotion-contingent: as the client's underlying valence and arousal shift, the supporter should adapt its high-level strategy (e.g., reflecting, asking, or planning) accordingly. In this section, we therefore examine whether different LLM supporters exhibit such adaptation, i.e., whether their BOLT-labeled behaviors systematically vary with the inferred client emotion, and whether RL-guided strategy selection yields patterns that more closely resemble high-quality human counseling.

As mentioned in Section 4.1, we model a user's latent emotional along two primary dimensions: valence (pleasant-unpleasant) and arousal (activated-calm). We operationalize these dimensions

using the NRC-VAD Lexicon (Mohammad 2025) to calculate the valence and arousal scores for each patient utterance. Although our POMDP maintains and updates a latent valence-arousal state during interaction, using NRC-VAD in evaluation provides an external, model-agnostic proxy of affect derived directly from the client's language. This choice makes our emotion-contingent analyses independent of our own state estimator, reducing the risk that the evaluation is confounded by estimator-specific assumptions or biases.

The NRC-VAD Lexicon contains approximately 55,000 English words and their valence and arousal scores rated by human. The ratings were obtained through large-scale crowdsourcing and have been shown to exhibit high reliability, with scores on a continuous scale from -1 (lowest valence/arousal) to 1 (highest). We label patient utterances with positive or negative valence based on whether their valence score lies above or below zero. We label utterances as high or low arousal by comparing each utterance's arousal score to the mean arousal score computed across all patient utterances.

To focus the analysis on clinically meaningful adaptation, we restrict attention to behavior categories where high-quality human therapists demonstrably shift their strategy as emotion changes, and we compare how different chatbot supporters adjust along the same dimensions (Figure 2).

Figures (2a)-(2d) show that high-quality therapists shift their behavioral emphasis depending on valence, adopting distinct response patterns when clients display positive versus negative valence. In Figure (2a), high-quality therapists increase Reflections on Emotions when responding to negative (vs. positive) valence. While all LLM models mirror this trend, the extent of modulation differs: non-RL models either elevate or decrease the levels, whereas RL-enhanced supporters maintain levels that more closely approximate therapist behavior. While both baseline and RL-enhanced LLMs appear to understand when Reflections on Consequences (Figure (2b)) and Conflicts (Figure (2c)) are needed, baseline LLMs (with the exception of RAG-Llama) deploy these reflective behaviors at levels that are too muted to be meaningful. RL helps close this gap by increasing reflective behaviors toward the levels observed in high-quality counseling. Figure (2d) demonstrates that, both RL+GPT and RL+Llama show rising usage of Questions on Perspectives under positive valence, matching the therapist trend, whereas their non-RL counterparts do not.

A parallel pattern emerges under arousal variation. In Figure (2e), high-quality therapists decrease Reflections on Emotions under low arousal; RL-enhanced models reproduce this directional change more consistently than their non-RL baselines and also match the usage level more consistently. In Figure (2f), high-quality human therapists decrease Reflections on Conflicts when arousal becomes lower. While baseline LLMs appear to understand when Reflections on Conflicts is needed, our RL mechanism again helps them express enough of this behavior for it to be meaningful. That is, it raises usage closer to the levels seen in effective human providers for majority of the baseline models. Figure (2h) further illustrates this: high-quality human therapist elevates Planning when arousal lowers and RL-enhanced supporters align with such a behavior, whereas majority of the baseline LLMs show comparatively muted change even with wrong directional change. Finally, Figure (2g) shows that, although all Llama-based models show muted usage of Questions on emotions, RL-

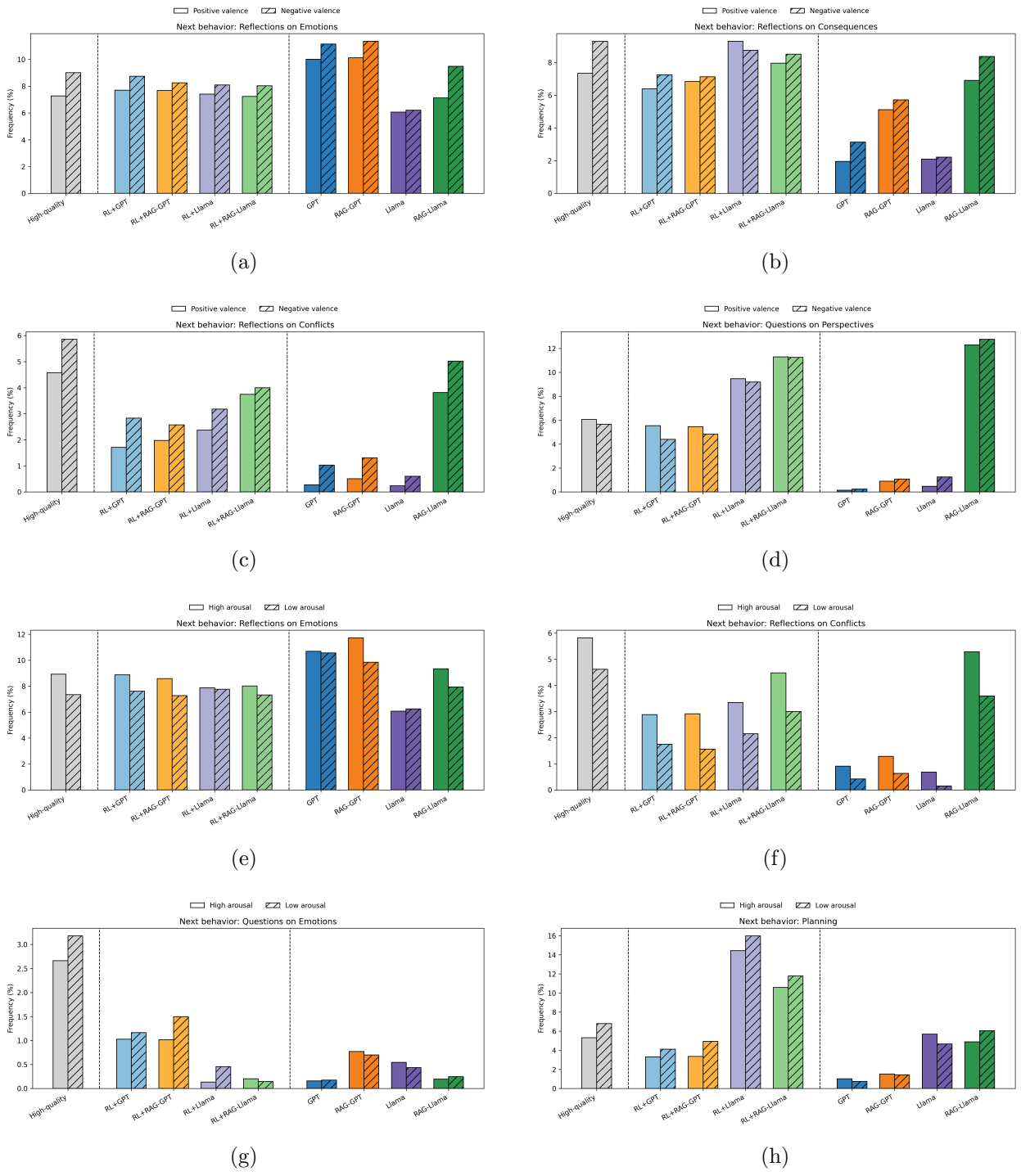


Figure 2: Supporter behavior frequencies following patient positive or negative valence (2a)-(2d), and following patient high or low arousal utterances (2e)-(2h), with results compared among chatbot supporters and the high-quality human therapist.

enhanced GPT models mimic the trend of high-quality therapists much better than their baseline counterpart.

In summary, these results show that RL improves emotion-contingent strategy adaptation in two ways: it not only aligns the *direction* and *magnitude* of behavior shifts with high-quality human counseling, but also reduces variance across LLM architectures. In particular, baseline LLM supporters exhibit substantial architecture-dependent differences, with some models capturing certain emotion-conditioned patterns while others fail to do so, suggesting that adaptation is fragile and unstable under purely reactive LLM generation. By contrast, RL-guided strategy selection yields more stable and consistent emotion-responsive behaviors across architectures, reinforcing the view that the RL layer provides a model-agnostic mechanism for controlling supportive dynamics under partial observability, improving their emotional attunement and adaptive response quality.

7 Controlled Human Experiment

Our empirical evaluation thus far has demonstrated that RL consistently improves conversational strategy alignment across various LLM architectures. To complement these model-centric findings, we conduct a real-world randomized controlled experiment designed to solicit human preferences regarding the perceived conversational quality of RL-enhanced chatbots. This experiment allows us to assess whether the improvements observed in empirical evaluation translate into meaningful differences from the perspective of potential users seeking mental health support[5]. In our experiment, we concentrate on one representative comparison: GPT vs. RL+GPT, which allows us to reduce respondent burden and to isolate the incremental value added by RL-enhanced strategies over a widely used baseline.

7.1 Study Design and Procedure

To ensure ecological validity, we design the experiment around three representative support contexts frequently observed in mental health services (Pérez-Rosas et al. 2019): (1) anxiety related to life stressors, (2) low energy and lack of motivation, and (3) substance use and relapse concerns. Using the same LLM-based patient role-play procedure described in Section 6.2, we instantiate each of these contexts through a dedicated patient simulator. For each patient simulator, we generate a pair of conversations: one from an interaction with baseline LLM supporter and the other from the RL-enhanced supporter.

Participants were recruited from Prolific and restricted to U.S.-based adults who reported fluency in English. Eligible participants were randomly assigned to a pairwise comparison condition in which they evaluated one conversation pair corresponding to one of the three contexts. To avoid confounding due to presentation order, we counterbalanced which model appeared in the first position and which appeared in the second position. This resulted in 3 contexts $\times$ 2 display

[5]The human evaluation protocol was reviewed by an institutional Human Subjects Division and determined to be exempt from IRB approval.

orders = 6 experimental groups. The original simulated dialogues consisted of 20 conversational rounds; however, to maintain page and cognitive efficiency while preserving the core interactional dynamics, participants were shown only the first three rounds of each conversation. Before viewing the conversation pair, participants were presented with a brief priming question to contextualize their reading task: "Imagine you are a patient seeking mental health support. What would your primary goal be?" Participants then selected one of two broad user intentions: 1. Comfort-seeking, defined as prioritizing emotional soothing or immediate relief, or 2. Progress-seeking, defined as preferring constructive insights, reflective prompts, or guidance that facilitate self-understanding and behavioral change. This measure enables downstream analyses examining how conversational preferences vary across user motives. After reading the two side-by-side conversations, participants indicated which version they preferred and provided a confidence rating on a 1-5 scale.

7.2 Main Results

We recruited approximately 180 participants, of whom 153 passed all attention checks and were retained for analysis. Table 4 summarizes user preferences across the three simulated patient contexts. We define the win rate of the RL-enhanced model as the proportion of participants who selected the RL-enhanced conversation over the baseline GPT version. Across all conditions, the RL-enhanced model is consistently preferred. In the first anxiety/life-stressor context, the RL model achieves a win rate of 66.67%, indicating a strong preference for the RL-enhanced conversational style. Preferences remain favorable, though somewhat attenuated, in the low energy and lack of motivation context, with a win rate of 57.63%, and in the substance use and relapse context with a win rate of 52.17%. Aggregated across all contexts, the RL-enhanced model attains an overall win rate of 58.82%, suggesting that participants generally favor the RL-enhanced conversations over the baseline. Confidence ratings further support the robustness of these preferences. Participants report relatively high confidence in their choices across all simulators, with average confidence levels ranging from 3.97 to 4.21 on a five-point scale and an overall average of 4.06. Notably, these evaluations are based on only the first three rounds of each conversation, implying that the advantages of RL-enhanced strategies emerge rapidly, even in the earliest stages of an interaction.

	Simulator 1	**Simulator 2**	**Simulator 3**	**Overall**
Win Rate	66.67%	57.63%	52.17%	58.82%
Avg. Confidence	4.21	3.97	4.02	4.06

Table 4: User Preference of GPT vs. RL+GPT.

To further understand heterogeneity in user preferences, we analyze win rates separately for comfort-seeking and progress-seeking participants and report the results in Table 5. A clear divergence emerges between the two groups. Progress-seeking users (i.e., those who prioritize constructive insights and reflective prompts) exhibit a remarkably strong preference for the RL-enhanced model across all simulators, with win rates ranging from 62.96% to 81.82% and an overall win rate of 70.30%. In contrast, comfort-seeking users (i.e., those who seek immediate

emotional comfort) show a substantially lower preference for the RL-enhanced model, with overall win rates of 36.54%. Both groups report similarly high confidence levels (overall confidence of 4.00 for comfort-seekers and 4.09 for progress-seekers), suggesting that users' judgments are not driven by uncertainty but by differing priorities in conversational goals. These findings indicate that RL-enhanced strategy selection is particularly well aligned with the expectations of progress-seeking users, who value conversations that probe underlying concerns and offer constructive avenues for change. In contrast, comfort-seeking users appear to place greater value on the immediate emotional reassurance that baseline GPT provides, as the standard model naturally excels at validation, normalizing, and supportive language. Taken together, these patterns suggest that the RL-enhanced model demonstrates its strongest advantages in settings where users seek reflection and forward movement. By sequencing MI strategies over multiple turns, the model produces counseling-like interactions that progress-oriented users perceive as more meaningful. The boundary condition observed here thus clarifies the model's strength: it is most effective in growth-oriented scenarios rather than in contexts where immediate emotional soothing is the primary goal.

	Simulator 1	**Simulator 2**	**Simulator 3**	**Overall**
Comfort-seeking users				
Win Rate	33.33%	38.89%	36.84%	36.54%
Avg. Confidence	4.07	3.67	4.26	4.00
Progress-seeking users				
Win Rate	81.82%	65.85%	62.96%	70.30%
Avg. Confidence	4.27	4.10	3.85	4.09

Table 5: Preference by User Type.

We further analyze the open remarks that participants provided at the end of the survey to better understand the reasoning behind their preferences. Consistent with the quantitative patterns, many respondents emphasize that the RL-enhanced conversations help them "think deeply about the situation", "get to the bottom of the problem", and identify "actual steps the person could take". In contrast, baseline GPT is often described as "just supportive", offering "platitudes" or "kind words" without advancing the conversation. Several participants note that the RL model balances empathy with direction, acknowledging the user's struggles without "getting bogged down in talking about self-care". These comments support the interpretation that the RL-enhanced policy more effectively paces validation, explores underlying concerns, and introduces guidance when contextually appropriate – precisely the behaviors progress-seeking users find most helpful.

Taken together, these experimental findings reinforce the empirical evidence that the RL-enhanced chatbot produces more strategically aligned and therapeutically meaningful conversations, particularly in settings where users are receptive to reflection and constructive guidance. These findings suggest a route to personalized support (Wang and Li 2025, Zhou et al. 2023b, Bardhan et al. 2025), as RL provides especially strong gains for progress-seeking users, motivating intention-aware strategy selection that avoids one-size-fits-all responses.

8 Concluding Remarks

This paper argues that effective mental health support chatbots require more than generating locally "good" responses; they require multi-turn planning that adapts to a user's evolving emotional state. Motivated by design principles of supportive dialogue, we treat emotion as a latent state expressed indirectly through language and show that counseling-style support is inherently sequential and prospective: the quality of an interaction depends on how supportive behaviors are sequenced, paced, and adapted across turns, not on any single response in isolation.

We propose a principled approach to design emotionally-aware conversational agents by integrating reinforcement learning and large language models within a POMDP formulation. Our proposed RL+LLM framework is modular and lightweight. It learns a policy over a compact set of high-level supportive strategies and delegates natural-language realization to an LLM. This abstraction avoids token-level RL, reduces reliance on heavily hand-crafted prompts, and makes the approach broadly extensible across LLM architectures, prompting, in-context learning, and retrieval-augmented generation. Importantly, designing RL for partially observable, long-horizon dialogue is intrinsically challenging, and there is no universally standard, off-the-shelf solution for POMDP interactions. To address this, we develop a Thompson Sampling-based Bayesian RL mechanism by enabling hidden emotional state dynamics tracking based on observable conversational cues and adaptive forward-looking strategy selection. It enhances the extant LLMs' ability to reason about long-term consequences and to understand user's emotional state evolution over multiple conversation turns, so as to optimize and personalize different users' emotional experiences with a chatbot supporter. The ability of our chatbot to infer from and respond strategically to evolving emotional states makes this approach particularly suited for mental health applications, where trust-building, emotional-alignment and sustained engagement are critical.

Another challenge in advancing mental health chatbots is that there is no broadly accepted, conversation-level evaluation protocol that reliably determines whether a multi-turn interaction is better, especially when quality depends on long-range phenomena such as pacing, emotional attunement, and strategic timing. To mitigate this gap, we adopt behavior-trajectory analyses and human-anchored benchmarks, comparing generated behavior traces to high- versus low-quality human counseling conversations. Empirical results demonstrate that RL shifts both the temporal trajectories and the overall composition of supporter behaviors toward patterns characteristic of high-quality human counselors. In emotion-contingent analyses, RL enhances adaptation to client affect, improving both the direction and magnitude of strategy shifts while reducing architecture-dependent variance, indicating that strategy-level RL provides a more stable control than purely reactive LLM generation. Finally, a randomized human preference study corroborates these findings: participants consistently prefer RL-enhanced conversations over the LLM baseline, with substantially stronger gains among progress-seeking users, even when judging only the first few turns.

Together, our approach provides a practical and theoretically grounded pathway toward more effective, responsive, and human-centric AI-driven emotional support. Our framework can serve as

a reusable strategy layer for wellness-oriented micro-interventions that offer empathetic, on-demand companionship and support emotional regulation, values clarification, and forward momentum between or outside traditional therapy settings. By learning a policy over interpretable, high-level support strategies, this abstraction also supports auditability and governance: stakeholders can monitor strategy usage (e.g., over-reliance on psychoeducation), impose clinically motivated constraints, and reason about system behavior at the level of counseling moves rather than opaque token sequences. More broadly, our approach generalizes to other domains where goal-directed, long-horizon interaction matters, including peer support, education, negotiation, or persuasion, in which a conversational agent must effectively model and steer the mindset or opinions of its human conversational partner in order to accomplish some overall conversation-level goal.

Finally, our study has several limitations. Because high-quality, accessible mental health counseling datasets are rare, our empirical evaluation focuses primarily on a publicly available counseling-style dataset that is well established in prior mental health counseling research and well suited to our research setting (Pérez-Rosas et al. 2019, Sharma et al. 2021, Chiu et al. 2024). Broader validation across additional datasets (if available) and conversational contexts would further strengthen generalizability. In addition, our human study evaluates short excerpts rather than full dialogues, leaving open questions about how preferences evolve over longer interactions. Finally, while our framework is designed for wellness-oriented micro-interventions rather than clinical therapy or crisis care, deploying such systems in practice would require careful attention to safety, oversight, and domain-specific constraints.

References

Abbasi A, Zhou Y, Deng S, Zhang P (2018) Text analytics to support sense-making in social media. *MIS Quarterly* 42(2):427–464.

Adamopoulou E, Moussiades L (2020) An overview of chatbot technology. *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*, 373–383 (Springer).

Bardhan I, Kohli R, Oborn E, Mishra A, Tan CH, Tremblay MC, Sarker S (2025) Human-centric information systems research on the digital future of healthcare. *Information Systems Research* 36(1):1–20.

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S, et al. (2023) Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* .

Butler EA, Randall AK (2013) Emotional coregulation in close relationships. *Emotion Review* 5(2):202–210.

Calvo RA, D'Mello S, Gratch JM, Kappas A (2015) *The Oxford handbook of affective computing* (Oxford University Press).

Chen B, Yi F, Varró D (2023a) Prompting or fine-tuning? a comparative study of large language models for taxonomy construction. *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, 588–596 (IEEE).

Chen G, Xiao S, Zhang C, Zhao H (2023b) A theory-driven deep learning method for voice chat–based customer response prediction. *Information Systems Research* 34(4):1513–1532.

Chen J, Wang X, Xu R, Yuan S, Zhang Y, Shi W, Xie J, Li S, Yang R, Zhu T, et al. (2024) From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231* .

Chiu YY, Sharma A, Lin IW, Althoff T (2024) A computational framework for behavioral assessment of llm therapists. *arXiv preprint arXiv:2401.00820* .

Clark M, Bailey S (2024) Chatbots in health care: connecting patients to information. *Canadian Journal of Health Technologies* 4(1).

Cohen RA (2011) Yerkes–dodson law. *Encyclopedia of clinical neuropsychology*, 2737–2738 (Springer).

Fiorentino F, Gualco I, Carcione A, Lingiardi V, Tanzilli A (2024) Exploring the outcomes of psychotherapy sessions: how do therapists' responsiveness and emotional responses to patients with personality disorders affect the depth of elaboration? *Frontiers in Psychology* 15:1390754.

Fredrickson BL (2001) The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American psychologist* 56(3):218.

Gao C, Lei W, He X, De Rijke M, Chua TS (2021) Advances and challenges in conversational recommender systems: A survey. *AI open* 2:100–126.

Gao Y, Zhang MM, Lysyakov M (2025) Does social bot help socialize? evidence from a microblogging platform. *Information Systems Research* .

Gnewuch U, Morana S, Hinz O, Kellner R, Maedche A (2024) More than a bot? the impact of disclosing human involvement on customer interactions with hybrid service agents. *Information Systems Research* 35(3):936–955.

Gray E, McCambridge J, Strang J (2005) The effectiveness of motivational interviewing delivered by youth workers in reducing drinking, cigarette and cannabis smoking among young people: Quasi-experimental pilot study. *Alcohol and Alcoholism* 40(6):535–539.

Greenberg LS (2004) Emotion–focused therapy. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice* 11(1):3–16.

Gross JJ (2015) The extended process model of emotion regulation: Elaborations, applications, and future directions. *Psychological inquiry* 26(1):130–137.

Han E, Yin D, Zhang H (2023) Bots with feelings: should ai agents express positive emotion in customer service? *Information Systems Research* 34(3):1296–1311.

Haque MR, Rubya S (2023) An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. *JMIR mHealth and uHealth* 11(1):e44838.

Hill CE (2009) *Helping skills: Facilitating exploration, insight, and action* (American Psychological Association).

Hsu SL, Shah RS, Senthil P, Ashktorab Z, Dugan C, Geyer W, Yang D (2023) Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982* .

Huang MH, Rust RT (2024) The caring machine: Feeling ai for customer care. *Journal of Marketing* 88(5):1–23.

Humeau S, Shuster K, Lachaux MA, Weston J (2019) Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969* .

Jensen CD, Cushing CC, Aylward BS, Craig JT, Sorell DM, Steele RG (2011) Effectiveness of motivational interviewing interventions for adolescent substance use behavior change: a meta-analytic review. *Journal of consulting and clinical psychology* 79(4):433.

Jiang H, Zhang X, Cao X, Breazeal C, Roy D, Kabbara J (2024) Personallm: Investigating the ability of large language models to express personality traits. *Findings of the association for computational linguistics: NAACL 2024*, 3605–3627.

Kuyper L, de Wit J, Heijman T, Fennema H, van Bergen J, Vanwesenbeeck I (2009) Influencing risk behavior of sexually transmitted infection clinic visitors: efficacy of a new methodology of motivational preventive counseling. *AIDS patient care and STDs* 23(6):423–431.

Kwan WC, Zeng X, Jiang Y, Wang Y, Li L, Shang L, Jiang X, Liu Q, Wong KF (2024) MT-eval: A multi-turn capabilities evaluation benchmark for large language models. Al-Onaizan Y, Bansal M, Chen YN, eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20153–20177 (Miami, Florida, USA: Association for Computational Linguistics), URL `http://dx.doi.org/10.18653/v1/2024.emnlp-main.1124`.

Lale S, Azizzadenesheli K, Hassibi B, Anandkumar A (2020) Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems* 33:20876–20888.

Leng Y, Yuan Y (2023) Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198* .

Li G, Jiao Y, Huang Y, Wei Y, Chen Y (2025) Transformers meet in-context learning: A universal approximation theory. *arXiv preprint arXiv:2506.05200* .

Liebman E, Saar-Tsechansky M, Stone P (2019) The right music at the right time: Adaptive personalized playlists based on sequence modeling. *MIS quarterly* 43(3):765–A6.

Lin Z, Madotto A, Shin J, Xu P, Fung P (2019) Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687* .

Liu S, Zheng C, Demasi O, Sabour S, Li Y, Yu Z, Jiang Y, Huang M (2021) Towards emotional support dialog systems. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483.

Luo X, Tong S, Fang Z, Qu Z (2019) Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science* 38(6):937–947.

Majumder N, Hong P, Peng S, Lu J, Ghosal D, Gelbukh A, Mihalcea R, Poria S (2020) Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454* .

Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E (2016) Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine* 176(5):619–625.

Mohammad SM (2025) Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv preprint arXiv:2503.23547* .

Moors A, Ellsworth PC, Scherer KR, Frijda NH (2013) Appraisal theories of emotion: State of the art and future development. *Emotion review* 5(2):119–124.

Ogden P, Minton K, Pain C (2006) *Trauma and the body: A sensorimotor approach to psychotherapy (norton series on interpersonal neurobiology)* (WW Norton & Company).

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, et al. (2022) Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35:27730–27744.

Oymak S, Ozay N (2019) Non-asymptotic identification of lti systems from a single trajectory. *2019 American control conference (ACC)*, 5655–5661 (IEEE).

Pei G, Li H, Lu Y, Wang Y, Hua S, Li T (2024) Affective computing: Recent advances, challenges, and future trends. *Intelligent Computing* 3:0076.

Pérez-Rosas V, Wu X, Resnicow K, Mihalcea R (2019) What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 926–935.

Posner J, Russell JA, Peterson BS (2005) The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17(3):715–734.

Qiu H, Lan Z (2024) Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787* .

Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C (2023) Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36:53728–53741.

Raimi R, Kim A, Ayabakan S, Dennis AR (2025) Judgmental bot: Conversational agents in online mental health screening. *MIS Quarterly* 49(4):1319–1356.

Rashkin H, Smith EM, Li M, Boureau YL (2019) Towards empathetic open-domain conversation models: A new benchmark and dataset. *Proceedings of the 57th annual meeting of the association for computational linguistics*, 5370–5381.

Ren G, Ni X, Malik M, Ke Q (2018) Conversational query understanding using sequence to sequence modeling. *Proceedings of the 2018 World Wide Web Conference*, 1715–1724.

Rubak S, Sandbæk A, Lauritzen T, Christensen B (2005) Motivational interviewing: a systematic review and meta-analysis. *British journal of general practice* 55(513):305–312.

Ruusuvuori J (2012) Emotion, affect and conversation. *The handbook of conversation analysis* 330–349.

Saha T, Chopra S, Saha S, Bhattacharyya P, Kumar P (2021) A large-scale dataset for motivational dialogue system: An application of natural language generation to mental health. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE).

Schanke S, Burtch G, Ray G (2021) Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research* 32(3):736–751.

Schindler D, Maiberger T, Koschate-Fischer N, Hoyer WD (2024) How speaking versus writing to conversational agents shapes consumers' choice and choice satisfaction. *Journal of the Academy of Marketing Science* 52(3):634–652.

Scholich T, Barr M, Stirman SW, Raj S (2025) A comparison of responses from human therapists and large language model–based chatbots to assess therapeutic communication: Mixed methods study. *JMIR Mental Health* 12(1):e69709.

Shah RS, Holt F, Hayati SA, Agarwal A, Wang YC, Kraut RE, Yang D (2022) Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proceedings of the ACM on Human-Computer Interaction* 6(CSCW2):1–24.

Shani L, Rosenberg A, Cassel A, Lang O, Calandriello D, Zipori A, Noga H, Keller O, Piot B, Szpektor I, et al. (2024) Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems* 37:118953–118993.

Shao Y, Gouws S, Britz D, Goldie A, Strope B, Kurzweil R (2017) Generating high-quality and informative

conversation responses with sequence-to-sequence models. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2210–2219.

Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T (2021) Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *Proceedings of the web conference 2021*, 194–205.

Siddals S, Torous J, Coxon A (2024) "it happened to be the perfect thing": experiences of generative ai chatbots for mental health. *npj Mental Health Research* 3(1):48.

Siegel DJ (2010) *The Mindful Therapist: A Clinician's Guide to Mindsight and Neural Integration (Norton Series on Interpersonal Neurobiology)* (WW Norton & Company).

Singh PV, Tan Y, Youn N (2011) A hidden markov model of developer learning dynamics in open source software projects. *Information Systems Research* 22(4):790–807.

Sorin V, Brin D, Barash Y, Konen E, Charney A, Nadkarni G, Klang E (2024) Large language models and empathy: systematic review. *Journal of medical Internet research* 26:e52597.

Stiles WB (2007) Theory-building case studies of counselling and psychotherapy. *Counselling and Psychotherapy Research* 7(2):122–127.

Tang Y, Yang Y, Abbasi A (2025) Personafuse: A personality activation-driven framework for enhancing human-llm interactions. *arXiv preprint arXiv:2509.07370* .

Thwaites R, Bennett-Levy J (2007) Conceptualizing empathy in cognitive behaviour therapy: Making the implicit explicit. *Behavioural and Cognitive Psychotherapy* 35(5):591–612.

Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB (2019) Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry* 64(7):456–464.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Advances in neural information processing systems* 30.

Wainberg ML, Scorza P, Shultz JM, Helpman L, Mootz JJ, Johnson KA, Neria Y, Bradford JME, Oquendo MA, Arbuckle MR (2017) Challenges and opportunities in global mental health: a research-to-practice perspective. *Current psychiatry reports* 19(5):28.

Wang C, Shi Y, Guo X, Chen G (2025a) Probing digital footprints and reaching for inherent preferences: A cause-disentanglement approach to personalized recommendations. *Information Systems Research* 36(3):1314–1332.

Wang H, Huang W, Wang Y, Xi Y, Lu J, Zhang H, Hu N, Liu Z, Pan JZ, Wong KF (2025b) Rethinking stateful tool use in multi-turn dialogues: Benchmarks and challenges. Che W, Nabende J, Shutova E, Pilehvar MT, eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 5433–5453 (Vienna, Austria: Association for Computational Linguistics), ISBN 979-8-89176-256-5, URL `http://dx.doi.org/10.18653/v1/2025.findings-acl.284`.

Wang Q, Peng S, Zha Z, Han X, Deng C, Hu L, Hu P (2023a) Enhancing the conversational agent with an emotional support system for mental health digital therapeutics. *Frontiers in Psychiatry* 14:1148534.

Wang R, Milani S, Chiu JC, Zhi J, Eack SM, Labrum T, Murphy SM, Jones N, Hardy K, Shen H, et al. (2024) Patient-Ψ: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660* .

Wang W, Li B (2025) Learning personalized privacy preference from public data. *Information Systems Research* 36(2):761–780.

Wang W, Zhao Z, Sun T (2023b) GPT-doctor: Customizing large language models for medical consultation. *arXiv preprint arXiv:2312.10225* .

Wang X, Wang Z, Liu J, Chen Y, Yuan L, Peng H, Ji H (2023c) Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691* .

Wang XS, Zheng J (2024) Can banning ai-generated content save user-generated Q&A platforms? SSRN Working Paper 4750326, URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4750326`.

World Health Organization (2022) Mental disorders: Key facts. URL `https://www.who.int/news-room/fact-sheets/detail/mental-disorders?`

World Health Organization (2023) Depression. URL `https://www.who.int/health-topics/depression`.

Xu Y, Dai H, Yan W (2024) Identity disclosure and anthropomorphism in voice chatbot design: a field experiment. *Management Science* .

Xu Y, Zhao H, Zhang Z (2021) Topic-aware multi-turn dialogue modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14176–14184.

Yan L, Tan Y (2014) Feeling blue? go online: An empirical study of social support among patients. *Information Systems Research* 25(4):690–709.

Yosef S, Zisquit M, Cohen B, Klomek AB, Bar K, Friedman D (2025) The impact of fine-tuning llms on the quality of automated therapy assessed by digital patients. *npj Mental Health Research* 4(1):43.

Yu H, McGuinness S (2024) An experimental study of integrating fine-tuned llms and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence* 1–16.

Zhang J, Qiao D, Yang M, Wei Q (2026) Regurgitative training: The value of real data in training large language models. *Management Science (forthcoming)* .

Zhou J, Chen Z, Wang B, Huang M (2023a) Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. *arXiv preprint arXiv:2307.07994* .

Zhou L, Gao J, Li D, Shum HY (2020) The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46(1):53–93.

Zhou T, Wang Y, Yan L, Tan Y (2023b) Spoiled for choice? personalized recommendation for healthcare decisions: A multiarmed bandit approach. *Information Systems Research* 34(4):1493–1512.

Zhou Y, Zanette A, Pan J, Levine S, Kumar A (2024) Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446* .