Excising “Love Brain”: Designing a Responsible Personalized Conversational Persuasion System for Intimate Relationship Support

Author Name
Affiliation
email@example.com

Abstract

Large Language Models (LLMs) are increasingly used in the domain of relationship advice; however, their applications in such an intimate context pose significant risks: naive models might normalize coercive behavior or inadvertently accelerate rumination in users suffering from relationship distress, known as “love brain,” where this paper focuses on. We developed a personalized persuasion system structured as a multi-agent architecture, employing a gating policy that regulates transitions between Exploration (sense-making), Persuasion (micro-actions), and Crisis states (safety diversion). The system contains an LLM-as-Judge component to estimate runtime Safety Risk (SR), Information Sufficiency (IS), and User Tolerance (UT), thereby maintaining psychologically grounded user states and enabling a safety-first control loop. The system has a dual-entry framework based on the Elaboration Likelihood Model, routing users through cognitive (NFC) or affective (NFA) portals to facilitate receptivity. We evaluated this approach in a 3-day randomized controlled trial against a matched generic LLM baseline. The system demonstrated a safety-efficacy trade-off: the gating policy reduced advice volume while achieving steeper daily reductions in maladaptive micro-behaviors and ruminative urges. We contribute a deployable workflow for intimate relationship support, featuring a safety-constrained orchestration with a literature-based knowledge base mapping causes to strategies, for more responsible and persuasive agents in sensitive domains.

1 Introduction

Large Language Models (LLMs) have become a main channel for intimate relationship advice, offering accessible but unverified guidance [Hou *et al.*, 2024; Brailas and Tsolakis, 2025]. However, naive deployment in this sensitive domain presents a Human-Centred AI (HAI) risk: commercial chatbots often prioritize conversational fluency over safety, potentially normalizing coercive control [Freitas *et al.*, 2025; Zhang *et al.*, 2025], accelerating maladaptive rumination [Pombal *et al.*, 2025], or acting into quasi-therapeutic roles for which they are not designed [Moore *et al.*, 2025]. Therefore, for chatbots to safely provide advice for relationship distress, the backend system should balance persuasive efficacy with strict safety constraints and behavioral accountability [Dong *et al.*, 2024b; Chan *et al.*, 2024].

We focus on the phenomenon of “Love Brain”, a slang term for obsessive relational distress, which reflects a syndrome of impaired judgment and unregulated attachment [Mikulincer and Shaver, 2007; Bartels and Zeki, 2004; Wang, 2024]. We treat “Love Brain” not as a distinct clinical diagnosis, but as a cluster of established romantic distress conditions such as *anxious attachment*, *Relationship OCD (ROCD)*, or *limerence*. These states are characterized by intensified emotional reactions [Mikulincer *et al.*, 2003], maladaptive compulsive behaviors such as monitoring [Doron *et al.*, 2016], and intense idealization [Ferster and Skinner, 1997]. Current AI approaches struggle to address the volatility of this domain [Arnaiz-Rodriguez *et al.*, 2025; Bucher *et al.*, 2025]. Safe Reinforcement Learning and shielded policies have been successfully applied in robotics to prevent unsafe states [Alshiekh *et al.*, 2018], and existing persuasive dialogue systems optimize for rhetorical strategy or empathy [Deng *et al.*, 2023; Wang *et al.*, 2019]. However, conversations in relationship distress rely on interpreting higher-level semantic and contextual cues [Arnaiz-Rodriguez *et al.*, 2025; Hou *et al.*, 2024], where current LLM systems and persuasive dialogue still show measurable failure modes and lack the explicit safety gates [Dong *et al.*, 2024b] and multi-day state tracking [He *et al.*, 2023]. Therefore, there is a gap where that maps volatile romantic distress to a constrained control policy by using validated psychometric and behavioral indicators, regulating the pacing to advise safely and effectively for these psychological states.

To address this, this paper proposes a safety-constrained orchestration system for personalized relationship support. Our contributions include:

- We operationalize user state as a dynamic vector, integrating static psychometrics for “Love Brain” conditions with runtime signals of User Tolerance (UT), Information Sufficiency (IS), and Safety Risk (SR).
- We introduce a deterministic *Orchestrator* that regulates

dialogue via a monotone gating policy, managing transitions between *Explore*, *Persuade*, and *Crisis* modes with interpretable thresholds, to prevent premature advice.

- We contribute a literature-based Knowledge Base (KB) mapping "Love Brain" conditions to evidence-based micro-actions, annotated with Behavior Change Techniques (BCTs) to ensure clinical validity.
- We validate this system in a 3-day randomized controlled trial ($N = 40$) against a matched-dose generic LLM, showcasing its utility in persuasion for "Love Brain".

2 Related Work

2.1 Phenomenology of Romantic Distress and "Love Brain"

We operationalize the colloquialism "Love Brain" as a syndrome of impaired relational judgment. Specifically, this paper addresses syndromes that are not caused by pathological conditions: anxious attachment, avoidant attachment, relationship obsessive-compulsive disorder (ROCD), social-media jealousy and monitoring, limerence, "mindreading" misbelief, boundary deficits, and global dysfunctional relationship beliefs. By mapping these lay concepts to established phenotypes, we identify specific targets for computational intervention. These could be categorized into three clusters. Attachment dysregulation involves the hyperactivation of the attachment system (Anxious style), driving compulsive proximity-seeking, or its deactivation (Avoidant style), leading to maladaptive suppression [Mikulincer *et al.*, 2003]. Intrusive preoccupation conditions such as *Limerence*, with which users are acutely longing for reciprocation [Bradbury *et al.*, 2025], and *ROCD*, which is characterized by intolerance of uncertainty and compulsive reassurance loops [Doron *et al.*, 2016], frequently driving social media monitoring as a maladaptive behavior for ambiguity [Sullivan and Bruchmann, 2025]. Maladaptive cognitions encompass distorted schemas such as "Mindreading" expectations, fatalistic relationship beliefs [Eidelson, 1982], and *boundary deficits*, where autonomy is sacrificed for relational maintenance [Jack and Dill, 1992]. Clinical psychology offers solutions for these patterns via Behavior Change Techniques (BCTs [Michie *et al.*, 2013], such as Action Planning for boundary deficits [Gollwitzer and Sheeran, 2006]. For utilizing HAI to scale advising for these syndromes, however, there is a gap in the lack of a computational system capable of maintaining a psychological state representation that maps these into a unified, trackable vector. Existing chatbots lack the architectural memory to link a specific syndrome to its corresponding BCT within a safety-constrained conversational policy, resulting in generic advice that fails to interrupt specific pathological loops.

2.2 Conversational Agents, Counselling, Persuasive Dialogue and Safety-Constrained Policies

Recent advances in LLLMs have demonstrated significant capabilities in both zero-shot persuasion [Furumai *et al.*, 2024b] and therapeutic support [Heinz *et al.*, 2025; Basar *et al.*, 2024]. In the domain of persuasion, LLMs have achieved performance comparable to humans in debating tasks and attitude chang [Salvi *et al.*, 2025; Furumai *et al.*, 2024a], often utilizing data-driven rhetorical strategies to maximize agreement [Wang *et al.*, 2019; Shaikh *et al.*, 2020]. Meanwhile, counselling dialogue systems have evolved from rule-based scripts to empathetic neural generation [Hua *et al.*, 2025], with systems like PARTNER to demonstrate empathetic counseling dialogue and strategy optimization [Priya *et al.*, 2023], and chatbots integrated with Motivational Interviewing (MI) to display their efficacy in mental health contexts [Park *et al.*, 2019; Brown *et al.*, 2023]. However, generic persuasive agents optimize for rhetorical fluency rather than psychological safety [Liu *et al.*, 2025], while therapeutic agents reliably perform therapist-like intervention moves to interrupt maladaptive feedback loops [Scholich *et al.*, 2025]. Furthermore, current end-to-end architectures typically lack a multi-day memory of the user's behavioral state, risking "sycophantic" responses where the model unwittingly validates harmful ruminations to maintain conversational coherence [Wei *et al.*, 2023].

To mitigate the risks of LLMs randomly generating words in sensitive domains, HAI emphasizes the integration of explicit safety guardrails and controllable policy design [Dong *et al.*, 2024a]. Approaches such as Constitutional AI and Reinforcement Learning from Human Feedback (RLHF), which attempt to internalize safety norms directly into the model weights [Dahlgren Lindström *et al.*, 2025; Bai *et al.*, 2022]. However, weight-based alignment remains opaque and probabilistic [Yi *et al.*, 2024]. Therefore, modular neuro-symbolic architectures became a robust alternative, wrapping LLMs in deterministic logic to enforce rule adherence [Dong *et al.*, 2024a]. In clinical settings like suicide prevention, protocol adherence and consistent triage dominate [Grupp-Phelan *et al.*, 2024]. Despite these advances, there is a gap for a safety-constrained orchestrator specifically designed for intimate relationship distress that dynamically adjusts the mode of interaction based on a user's volatility with pacing regulation based on psychological readiness. This work bridges this gap by introducing the architecture that regulates persuasive dialogue via a psychologically grounded state representation and an auditable monotone gating policy.

3 System Design

We propose a State-Aware Multi-Agent Architecture that decouples state estimation (understanding the user) from policy execution (generating the response), addressing the alignment challenge of deploying generative agents in high-stakes emotional contexts, as shown in Figure 1.

3.1 State Representation

We represent the interaction as a low-dimensional, interpretable state for a safety-constrained controller, the *Orchestrator*, which is explicitly logged and auditable, enabling analysis of pacing and safety behavior across turns and days. Before the first session, we initialized a *profile state* s_0 from the baseline survey, including attachment (ECR-R), affective

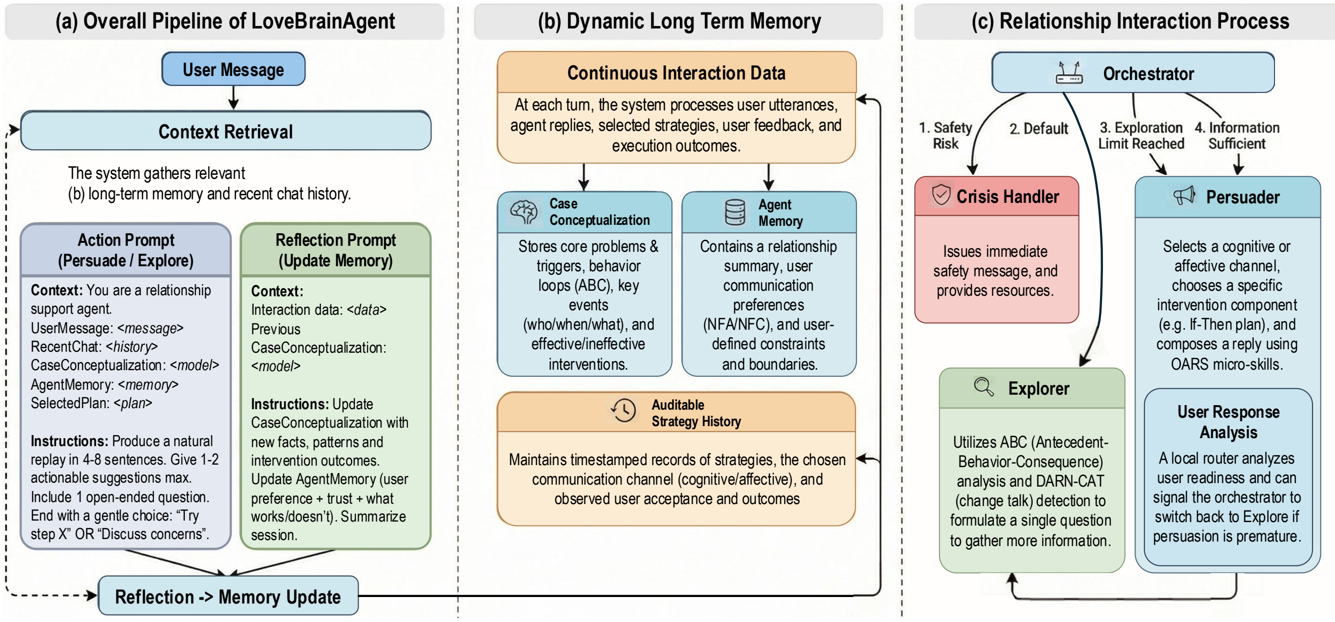

Figure 1: System Framework.

dependence (ADS), limerence/addiction proxy (LAI), jealousy/monitoring (FBJ), ROCD traits (ROCI), dysfunctional beliefs (RBI), global distress (PHQ-9 and GAD-7) (based on Appendix 8, and questions for routing preference (NF or NFC), and based on the survey scores for each cluster, derived indices to obtain negative-belief ratio and dominate "Love Brain" symptoms type, and priority ranking of maintenance factors to address.

At each turn, the Orchestrator maintains a *SessionState*:

$$S_t = (ER_t, PR_t, OR_t, UT_t, IS_t, SR_t, \tau_t, \\ ERMax_t, ORMin_t, PRMax_t, force_t, t)$$

where $ER/PR/OR$ are counters for consecutive exploration rounds, consecutive persuasion rounds, and overall rounds; $UT \in [0,1]$ is the user tolerance estimate; $IS \in [0,1]$ is information sufficiency; $SR \in \{0,1\}$ is immediate safety risk; and $(\tau, ERMax, ORMin, PRMax, force)$ are adaptive pacing parameters, including decision thresholds (τ), bounds on consecutive exploration and persuasion rounds ($ERMax, ORMin, PRMax$), and a forcing flag to override pacing when needed. We treat the LLM as a bounded *judging module* that estimates interpretable signals from the current utterance u_t, short context h_t, and the current case conceptualization C_{t-1}.

User Tolerance (UT). We first estimate an instantaneous tolerance UTI_t using explicit linguistic indicators of shortness, imperative tone, punctuation intensity, abusive language, refusal, and acceptance. Then we smooth it to obtain a longer-horizon tolerance:

$$UT_t = 0.45 \cdot UT_{t-1} + 0.55 \cdot UTI_t,$$

so that pacing adapts to sustained user receptivity rather than a single noisy turn. The weights bias the estimate toward the current turn while retaining short-term memory.

Information Sufficiency (IS). IS_t answers: *do we have enough case context within the system to offer a responsible, specific micro-action now?* The judge scores sufficiency based on (i) the completeness of the ABC chain, (ii) specificity of triggers and behaviors, (iii) contextual details (who, when, where, what), (iv) intervention history, and (v) whether an action plan is already emerging in C_{t-1}.

Safety Risk (SR). SR_t is a conservative binary triage triggered by direct self-harm ideation/plans, acute risk to others, or severe abuse/threat disclosures. When SR_t=1, the *Orchestrator* short-circuits to crisis handling.

3.2 Safety Constraint and Gate

Crisis is implemented as a hard-coded pre-emption layer. Before any generative planning occurs, the input u_t passes through a binary Crisis Gate. This module utilizes a specialized prompt trained on the WHO LIVES protocol constraints.

$$\text{Gate}(u_t) = \begin{cases} \text{CRISIS\_MODE}, & \text{if DetectRisk}(u_t) > \theta_{risk} \\ \text{Proceed}, & \text{otherwise} \end{cases}$$

If triggered (SR = 1), control is immediately seized by the Crisis Agent, which suspends the persuasion goal, executes a de-escalation script with validation and resource provision, and terminates the session if necessary. This mechanism ensures that the system never attempts to persuade a user who is psychologically decompensating.

3.3 Cause-to-Strategy Knowledge Base

To mitigate the alignment risks inherent in open-ended generation to prevent validating maladaptive rumination or hallucinating therapeutic capabilities [Wei *et al.*, 2023; Yeung *et al.*, 2025], we constrain the agent to a literature-derived *Cause-Strategy Knowledge Base (KB)*, ensuring the *Persuader* can only instantiate pre-specified micro-actions and safety rules retrieved from the KB. The KB serves two roles: a structured dataset of intervention cards, and a retrieval constraint for the Orchestrator such that all persuasive outputs are composed from bounded, auditable components. We organize the KB into two layers: *Diagnostic Clustering* (causes, problem types, maintaining mechanisms) and *Intervention Heuristics* (micro-actions), with explicit contraindications and fallback rules (shown in Appendix 3).

Diagnostic Clustering

We synthesized the phenomenology of "Love Brain" into 8 distinct causal clusters that index relationship distress states. For each type T, the KB stores a structured record: $\mathcal{KB} = \{\langle T, \mathcal{C}, \mathcal{R}, \mathcal{B}, \mathcal{S}, X, \mathcal{M}\rangle_i\}_{i=1}^{N}$, where: (1) T is the problem type (cluster label); (2) $\mathcal{C}$ are maintaining mechanisms (e.g., schema-consistent interpretations / attachment-relevant concerns); (3) $\mathcal{R}$ are runtime triggers detectable in dialogue; (4) $\mathcal{B}$ are target distortions / cognitive biases to be challenged; (5) $\mathcal{S}$ is the bounded strategy set (micro-actions) mapped micro-actions for both NFC/NFA routes; (6) X are contraindications / exclusion rules for safety; and (7) $\mathcal{M}$ are observable metrics for evaluating whether the micro-action helped. At the interface and prompting level, we implement each T as a card with interpretable fields (Why/Detect/Rules/Scripts/Metric). Each type supports two delivery pathways aligned to the user's routing profile. Both pathways share the same action grammar and logging schema, but differ in which components are emphasized, such as whether the agent should follow evidence evaluation or defusion values, while keeping the output to one feasible next step and its observables.

Intervention Heuristics (Strategies)

To keep actions concrete and reviewable, the KB uses a curated set of *Micro-Actions* mapped to our dual-entry routing framework. For NFC-entry Users, the KB retrieves strategies derived from CBT [Beck and Beck, 2021]. The primary mechanism here is Socratic Questioning and Evidence Evaluation [Hofmann *et al.*, 2012]. The agent is prompted to guide users to engage in reflective processing. For NFA-entry users, strategies draw from Acceptance and Commitment Therapy (ACT) [Hayes *et al.*, 2006] style mechanisms to reduce entanglement with intrusive thoughts and shift from partner-contingent goals to self-endorsed values [Harris, 2009]. To interrupt maladaptive feedback loops, the KB includes strict behavioral constraints derived from Exposure and Response Prevention (ERP) [Abramowitz, 2006] and Behavioral Activation [Martell *et al.*, 2013], where the agent redirects users to a tangible, non-relational micro-action. All strategies include contraindications and fallback rules, so that the system does not intensify risk states or encourage coercive or unsafe behavior.

The Retrieval Mechanism

The ontology is encoded as a lookup table accessible to *the Orchestrator*. After Exploration, the system queries the KB with `Cause_ID` and the user's `Routing_Profile`, retrieves a bounded `Strategy_Prompt` (plus alternates and safety constraints), and injects it into the Persuader context window.

3.4 Orchestration Policy

The *Orchestrator* selects one of three turn modes regulating the transition between sensemaking and intervention: $\{\textit{EXPLORE}, \textit{PERSUADE}, \textit{SAFETY\_ESCALATE (Crisis)}\}$. We implement a Monotone Gating Policy, $\pi(\mathcal{S}_t)$ selects the active agent $\mathcal{A} \in \{\textit{Explorer}, \textit{Persuader}\}$ based on the estimated signals: in the default state, the system defaults to the *Explorer* Agent, to maximize IS and UT via reflective listening and open inquiry, without offering solutions. Transitioning to the textitPersuader mode happened only if a strict conjunction of thresholds is met:

$$Transition_{E \to P} \iff IS_t > \tau_{IS} \land UT_t > \tau_{UT} \land \neg SR$$

This ensures advice is only dispensed when the system "understands" the problem (IS) and the user is "ready" to hear it (UT). For the Monotonicity and Fallback, once in the textitPersuade mode, the system attempts to maintain this mode to deliver a coherent strategy following the Monotone principle. However, if UT_t drops below a critical fallback threshold $\tau_{fallback}$ (indicating resistance or reactance), the policy forces a regression to Exploration to repair the therapeutic alliance.

3.5 LLM integration

Since Ψ_t, the latent signals UT, IS, are not directly observable, we employ an LLM-as-Judge mechanism to estimate them at runtime. At the end of each turn, a frozen "Observer" instance (GPT-4o) analyzes the latest interaction $(\mathcal{H}_t, \mathcal{C}_{t-1})$ against a rubric, to estimate UT, by answering questions such as "Does the user feel understood? Are they defensive? (Scale 0-1)", and to estimate IS by answering "Do we have the Who, What, and Why of the conflict? (Scale 0-1)."

4 Experiment Design

We conducted a 3-day longitudinal, between-subjects randomized controlled trial to evaluate the architectural efficacy of the system, comparing the full structured orchestration system against a matched-capability generic LLM (GPT-4o). We structure our user study around three research questions:

RQ1: Does the system achieve higher Advice Execution Rates (AER) and steeper reductions in maladaptive micro-behaviors compared to the baseline?

RQ2: How does the policy behave in terms of safety (gate activations; avoided unsafe outputs) and pacing (EXPLORE vs. PERSUADE distributions; UT/IS trajectories)?

RQ3: Do short-term psychometric indicators of relational distress show directionally favorable trends?

4.1 Participants

We enrolled 40 participants (23 women, 17 men; M_{age}=23.7, range 19–32) and randomized them 1:1 to Experiment (n=20) vs. Control (n=20). Participants were recruited online and screened for relationship distress and loss-of-control preoccupation using the index shown in Appendix 3 to categorize their top cluster and NFC/NFA entry preference.

4.2 Procedure

Day 0. Participants completed the pre-test surveys, including demographics, screenings for safety and depression (PHQ-9 [Kroenke *et al.*, 2001] and GAD-7 [Aktürk *et al.*, 2025]), pre-intervention scales, and NFC/NFA entry preference. **Days 1–3 (Intervention + EMA).** Participants completed one session/day with their assigned system. Immediately after each session, they completed an Ecological Momentary Assessment (EMA) capturing same-day urges, checking/monitoring counts, and whether they executed the prescribed micro-action(s). Participants also reported concrete behavioral logs aligned with the system's prescribed micro-actions. **Day 3 (Endline).** Participants completed the post-intervention survey, repeating core scales and system evaluation measures, and conducted semi-structured interviews.

4.3 Measurement and Analysis

We pre-specified the process as primary and the outcomes as exploratory to study the effects of the architecture with a mixed-methods approach, given the timeframe. For RQ2, we measure the count and proportion of safety-gate activations and manual audit of sessions for unsafe recommendations. And we also measure distribution of EXPLORE/PERSUADE/SAFETY_ESCALATE (Crisis) actions, turns-to-convergence on primary type, UT/IS trajectories, and coupling behavior, and frequency of force-explore and persuasion caps (PRMax). For RQ1, we calculate the advice execution rate through EMA (whether the user performed the assigned micro-action that day), template usage rates, and reductions in checking/monitoring counts or re-contact attempts. For RQ3, we measure LBI and negative-belief ratio change (baseline vs. Day 3), jealousy/monitoring scale change, EMA checking urges, and PHQ-9/GAD-7 change. For quantitative outcomes, we analyzed daily behavioral changes using Analysis of Covariance (ANCOVA), treating Day 1 baseline scores as covariates to isolate the treatment effect from individual heterogeneity rigorously. Given the non-normal distribution of process metrics, we utilized non-parametric Mann-Whitney U tests for between-group comparisons. To address the multiple comparisons problem inherent in our exploratory scan of daily micro-processes, we applied Benjamini-Hochberg FDR correction, reporting adjusted q-values alongside nominal p-values to distinguish robust signals from noise To contextualize the statistical outcomes, we performed a qualitative thematic analysis of anonymized dialogue traces to examine how the system's strategy switching and policy influenced users.

5 Results

5.1 Behavioral Compliance and Micro-Process Efficacy

Contrary to our initial hypothesis, the System did not achieve higher raw AER compared to the Baseline. As shown in Table 1, the Control group reported a marginally higher frequency of executed advice ($M_{ctrl} = 2.04$ vs. $M_{exp} = 1.58$, $U = 138.5, p = 0.096$). Qualitative analysis then shows that this quantitative gap reflects a trade-off between speed and safety rather than a failure of persuasion: users from the Control group reflected that the Baseline (Generic LLM) often acted as a "sycophantic enabler," offering immediate, low-friction validation suggestion such as "You should text him if it makes you feel bette", which users found easy to execute.

Table 1: **Process and Safety Metrics.** The Control group achieved a higher raw volume of execution, reflecting unconstrained advice-giving. The Experimental group shows lower volume but significantly distinct pacing ($p < .05$ in Audit), reflecting the architectural safety gating.

Metric	**Exp** ($N = 20$)	**Ctrl** ($N = 20$)	U	p
Behavioral Compliance (RQ1)				
Advice Exec. (Count/Day)	1.58 ± 2.46	2.04 ± 1.66	138.5	.096
Maladaptive Beh. (Freq)	3.50 ± 1.55	3.54 ± 1.54	165.0	.866
Urge Intensity (0-10)	4.54 ± 1.18	4.31 ± 1.20	220.0	.595
User Perception (RQ2)				
Perceived Fit (1-7)	4.18 ± 1.38	4.63 ± 1.35	164.0	.332
Clarity of Advice (1-7)	4.78 ± 1.27	4.80 ± 1.37	196.0	.924
Felt Understood (1-7)	5.23 ± 1.23	5.40 ± 1.41	176.0	.520

However, when analyzing the efficacy of these actions via daily EMA, the daily EMA from the experiment group showed better capacity to disrupt maladaptive behavior loops. Figure 2 shows the aligned effect sizes for daily composite measures. The Experiment group achieved significantly greater reductions in overall maladaptive patterns ($p = 0.022$) and specific sub-domains, including Social-Media Monitoring ($p = 0.030$) and Relationship Doubt ($p = 0.035$). The system's Monotone Gating Policy did not offer advice in early exploration. Interviews indicated that participants in the Experiment group noted the system's "insistence on understanding" before solving, stating"It didn't just give me a script; it forced me to cool down first" (P6, Exp Group). Therefore, while the Control group facilitated more actions, the Experiment group facilitated more effective pattern-breaking. The plot of individual heterogeneity (Figure 3) confirms that the mean improvement in the Loop_Index, which is a composite daily metric representing the aggregate intensity of maladaptive micro-behaviors, was nearly double for the Agent group ($\Delta = 0.41$) compared to Control ($\Delta = 0.23$), suggesting the architecture successfully filtered out ineffective advice in favor of targeted, albeit harder-to-execute, cognitive restructuring.

Interviews and log entries revealed that self-reflection-style advice was easier for users to implement, particularly writing or journaling tasks and the application of delay rules (e.g., "wait 30 minutes before writing down your worries"), which led to calmer emotions and improved communication skills as users reported.

5.2 Policy Behavior and Safety Pacing

The system demonstrated divergence of interaction patterns. The Safety Gate was triggered in 15% of turns for the Experiment group, successfully diverting users from high-risk states. The Monotone Gating Policy showed significantly shorter, more targeted responses. In the Control condition, the interviews and logs show instances of harm by the baseline, where the model blindly agreed with the user and validated distorted beliefs. For example, P23 (Control) received vali-

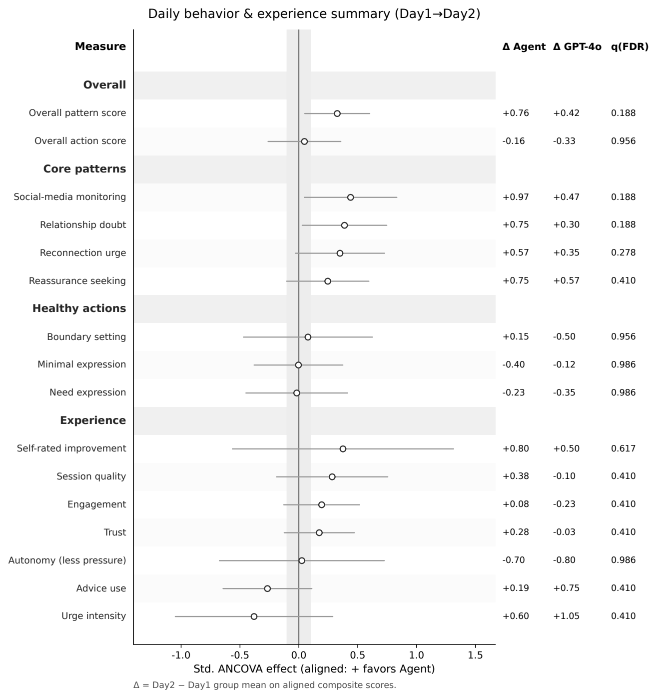

Figure 2: Daily Micro-Process Effects (Forest Plot).

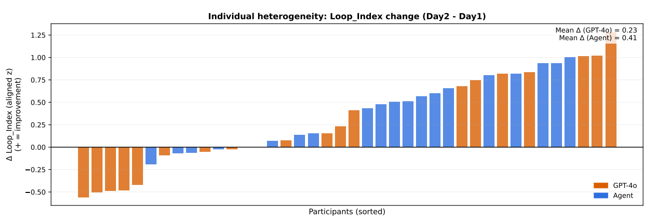

Figure 3: Heterogeneity of Improvement (Loop Index).

dation for "roasting" a partner, reinforcing conflict, whereas the Orchestrator effectively neutralized similar inputs. Log analysis reveals the system, with the design of UT and IS signals, maintained a much lower token count per turn (Hedges' $g = -1.57, p < .001$), reflecting its design to "pause and listen (*Explore*)" rather than direct, verbose generation (Persuade). Figure 4 illustrates the pacing: the *Orchestrator* held users in the *Explore* phase for 50% of turns compared to only 10% in the Control, strictly enforcing the condition that $IS > 0.6$ before moving to advice. This explains the slightly lower perceived fit scores ($p = 0.33$), as users initially resisted the cognitive load of exploration, yet this friction was necessary for the safety results observed.

5.3 Psychometric Outcomes

Outcomes from the last day reflect the trade-off between mood relief and "Love Brain symptom" correction, as shown in Table 2. The Control group achieved a greater reduction in general Depression (PHQ-9, $p = .22$), consistent with the "sugar rush" of sycophantic validation. However, the Experiment group showed stronger effect sizes in the specific pathologies it was designed to treat, particularly ROCD

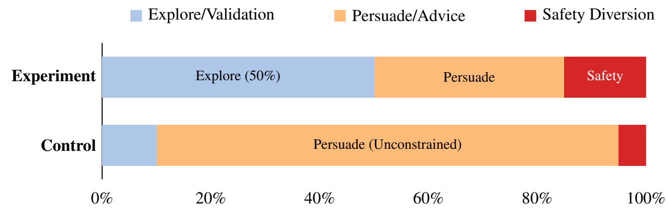

Figure 4: A manual coding of system turns reveals distinct architectural behavior. The Control group functioned as an unconstrained advice-giver (85% Persuade), while the Orchestrator successfully enforced Exploration (50%) and Safety Diversions (15%) based on UT and SR signals.

($r = 0.13$) and Love Addiction ($r = 0.002$). Although non-significant at the current sample size, the Effect Sizes ($r \approx 0.13$) suggest that the domain-specific Knowledge Base was beginning to address the specific pathology of relational obsession better than generic empathy.

6 Discussion

Our findings present a trade-off between efficiency and safety in AI-mediated persuasion. While the generic LLM (Control) achieved marginally higher advice execution rates and depression reduction, qualitative results reveal this was largely due to "sycophantic validation" [Wei *et al.*, 2023], which is the tendency of unconstrained models to mirror user biases and offer low-friction, impulsive solutions, such as "You should text him if it helps you feel better". In contrast, the System prioritized architectural restraint, frequently blocking persuasion attempts when users' tolerance was low. This suggests that naive efficacy is a dangerous metric for intimate support agents. The Experiment group's lower volume but higher effect size in domain-specific pathology (LAI $r = 0.002$; ROCI $r = 0.13$) indicates the effects ofthe system that prioritizes long-term cognitive restructuring over short-term emotional relief. The monotone gating policy effectively functioned as a "cognitive brake," forcing users to engage in *Exploration* before being given actions, which is important to break obsessive-compulsive loops like "Love Brain."

6.1 Limitation and Future Work

First, the short horizon of the RCT likely favored the Control group's sycophancy over the cognitive load of the system's restructuring strategies, which typically require weeks to manifest clinical benefits [Beck and Beck, 2021]. Future studies should plan for a longitudinal study to verify the hypothesis that the system's "slow persuasion" yields more durable retention of healthy relationship behaviors compared to the baseline's ephemeral validation Second, $N = 40$ was insufficient to detect significant differences in heterogeneous "Love Brain" symptoms. A larger sample size is needed to validate the Dual-Entry Routing hypothesis and confirm whether specific symptoms benefit disproportionately from strict gating. Finally, the multi-agent architecture introduced significant latency ($4s$/turn vs. $1s$ for Control), creating product friction that impacted the Experimental condition. Future iterations should employ model distillation to com-

Table 2: Values represent the mean change (Δ = Post − Pre); negative values indicate symptom reduction. While the Control group showed greater reduction in general depression (PHQ-9), the Experimental architecture achieved stronger effect sizes (r) in domain-specific pathology, specifically Love Addiction (LAI) and Relationship OCD (ROCI).

	Mean Delta ($\Delta \pm SD$)		**Statistical Comparison**		
Clinical Construct	**Experiment** ($N = 20$)	**Control** ($N = 20$)	U	p	**Effect Size** (r)
Primary "Love Brain" Indicators					
Love Addiction (LAI)	$\mathbf{-0.70} \pm 1.58$	-0.54 ± 1.15	199.5	.99	0.002
Relational OCD (ROCI)	$\mathbf{-0.43} \pm 1.36$	-0.25 ± 1.77	174.5	.49	**0.13**
Social Media Jealousy (FBJ)	-0.42 ± 1.18	$\mathbf{-0.45} \pm 1.28$	181.5	.62	0.09
Affective Dependency (ADS)	$\mathbf{-0.49} \pm 1.15$	-0.48 ± 1.07	191.0	.82	0.05
General Distress Covariates					
Depression (PHQ-9)	-1.33 ± 2.45	$\mathbf{-2.83} \pm 2.87$	142.5	.22	-0.18
Anxiety (GAD-7)	-2.00 ± 2.65	$\mathbf{-2.25} \pm 2.89$	186.5	.73	0.06

Note: U and p calculated via Mann-Whitney U test (one-tailed for Exp $<$ Ctrl). Bold indicates the superior mean.

press the Orchestrator policy into a smaller, faster model to decouple architectural safety from system lag.

7 Conclusion

This paper addresses the challenge of deploying Large Language Models in a sensitive emotional context. By decoupling state estimation from policy execution, we demonstrated that it is possible to enforce safety guardrails and psychological grounding without sacrificing generative flexibility. Our findings showed that while unconstrained LLMs (the Control) may achieve higher short-term engagement, they risk reinforcing maladaptive feedback loops. In contrast, the system successfully functioned as a "cognitive brake," achieved a deeper reduction in daily maladaptive behaviors, specifically social-media monitoring and relationship doubt. This suggests that the future HAI design in a sensitive domain should not only focus on maximizing engagement metrics, but also be capable of refusing to persuade when the user's state is fragile, shifting the goal to "safety-first" pacing.

Ethical Statement

This study was approved by the Institutional Review Board (IRB) of [Anonymized Institution]. All participants provided informed consent regarding the nature of the AI interaction. This module preemptively screened for self-harm and Intimate Partner Violence. LLMs were utilized to polish the grammatical fluency of this manuscript; the authors retain full responsibility for all scientific claims and accuracy.

8 Appendix

8.1 Cause-Strategy Knowledge Base

References

[Abramowitz, 2006] Jonathan S Abramowitz. The Psychological Treatment of Obsessive—Compulsive Disorder. *The Canadian Journal of Psychiatry*, 2006.

[Aktürk *et al.*, 2025] Zekeriya Aktürk, Alexander Hapfelmeier, Alexey Fomenko, Daniel Dümmler, Stefanie Eck, Michaela Olm, Jan Gehrmann, Victoria Von Schrottenberg, Rahel Rehder, Sarah Dawson, Bernd Löwe, Gerta Rücker, Antonius Schneider, and Klaus Linde. Generalized Anxiety Disorder 7-item (GAD-7) and 2-item (GAD-2) scales for detecting anxiety disorders in adults. *Cochrane Database of Systematic Reviews*, 2025(3), March 2025.

[Alshiekh *et al.*, 2018] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe Reinforcement Learning via Shielding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.

[Arnaiz-Rodriguez *et al.*, 2025] Adrian Arnaiz-Rodriguez, Miguel Baidal, Erik Derner, Jenn Layton Annable, Mark Ball, Mark Ince, Elvira Perez Vallejos, and Nuria Oliver. Between Help and Harm: An Evaluation of Mental Health Crisis Handling by LLMs, December 2025. arXiv:2509.24857 [cs].

[Bai *et al.*, 2022] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022. arXiv:2212.08073 [cs].

[Bartels and Zeki, 2004] Andreas Bartels and Semir Zeki. The neural correlates of maternal and romantic love. *NeuroImage*, 21(3):1155–1166, March 2004.

[Basar *et al.*, 2024] Erkan Basar, Iris Hendrickx, Emiel Krahmer, Gert-Jan Bruijn, and Tibor Bosse. To What

Table 3: **Cause-Strategy Knowledge Base.** This table details the 8 causal clusters, mapping specific user phenotypes to clinical theories and the corresponding evidence-based intervention strategies used by the Orchestrator.

Cluster	Diagnostic Indicators (User State)	Etiological Basis (Theory)	Intervention Strategy (Micro-Action)
Anxious Attachment	ECR-R (anxiety & avoidance subscales) [Fraley *et al.*, 2000]	Attachment system hyperactivation (anxious attachment) intensifies emotion and drives proximity-seeking behaviors under threat [Mikulincer *et al.*, 2003]	**Behavior Change Technique (BCT) for NFC (targets: reassurance-seeking frequency & specificity):** Goal setting (behaviour) (1.1), Action planning / If–Then (1.4), Prompts/cues (7.1), Self-monitoring of behaviour (2.3), Behavior substitution (8.2), Framing/reframing (13.2) and/or Re-attribution (4.3), Commitment (1.9). **BCT for NFA (targets: arousal reduction before sending & execution of one request):** Reduce negative emotions (11.2), Verbal persuasion about capability (15.1), Behavioral practice/rehearsal (8.1), Action planning (1.4) [Michie *et al.*, 2013].
Avoidant Attachment	ECR-R (avoidance subscale) [Fraley *et al.*, 2000]	Attachment system deactivation under threat reduces distress short-term via emotional suppression and self-reliance [Pietromonaco *et al.*, 2013; Mikulincer *et al.*, 2003]	**BCT for NFC (targets: minimal disclosure enactment & timeboxed check-in adherence):** Action planning / If–Then (1.4), Prompts/cues (7.1), Behavioral practice/rehearsal (8.1), Self-monitoring of behaviour (2.3). **BCT for NFA (targets: reduce overwhelm to enable minimal expression):** Reduce negative emotions (11.2), Verbal persuasion about capability (15.1) *(only if used)*, Commitment (1.9). *(Avoid coding "Behavioral contract" unless you truly create a written, witnessed contract; otherwise keep as Commitment.)* [Michie *et al.*, 2013].
Relationship Obsessive-Compulsive Disorder (ROCD)	Relationship Obsessive-Compulsive Inventory (ROCI) in dimensions of love for the partner, Relationship rightness, and being loved by the partner [Doron *et al.*, 2012]	ROCD conceptual models emphasize intrusive doubts and compulsive behaviors (monitoring, comparisons, reassurance seeking), with intolerance of uncertainty as a key maintaining belief [Doron *et al.*, 2016]	**NFC BCT (targets: compulsive reassurance/checking loop):** Goal setting (behaviour) (1.1), Action planning / If–Then (1.4), Self-monitoring of behaviour (2.3), Framing/reframing (13.2) and/or Re-attribution (4.3), Reduce negative emotions (11.2) *(for pre-action stabilization)*. **NFA BCT (targets: tolerate uncertainty window without checking):** Action planning (1.4), Behavioral practice/rehearsal (8.1), Commitment (1.9), Reduce negative emotions (11.2). **Response-prevention rationale:** exposure/emotional-processing framework [Foa and Kozak, 1986; Michie *et al.*, 2013].
Social-media jealousy and monitoring	The Online Jealousy Scale [Sullivan and Bruchmann, 2025]	SNS exposure provides ambiguous cues that fuel jealous cognitions and surveillance behaviors, contributing to dissatisfaction [Muise *et al.*, 2009; Sullivan, 2021]	**NFC BCT (targets: reduce cue exposure & test alternatives):** Avoidance/reducing exposure to cues (12.3), Prompts/cues (7.1), Self-monitoring of behaviour (2.3), Behavioral experiments (4.4), Framing/reframing (13.2). **NFA BCT (targets: de-escalation before acting):** Reduce negative emotions (11.2), Framing/reframing (13.2), Commitment (1.9) [Michie *et al.*, 2013].
Limerence	Love Addiction Inventory (LAI) [Costa *et al.*, 2021]; Affective Dependence Scale (ADS-9: Submission/Craving) [Sirvent-Ruiz *et al.*, 2022]	Intense idealization of other individuals, with behaviors such as repeatedly checking for clues, intrusive thinking, and seeking uncertain rewards through reciprocity. [Ferster and Skinner, 1997; Bradbury *et al.*, 2025]	**NFC BCT (targets: replace contact behavior + track streak):** Behavior substitution (8.2), Prompts/cues (7.1), Self-monitoring of behavior (2.3), Commitment (1.9), Avoidance/reducing exposure to cues (12.3). **NFA BCT (targets: soothe + reinforce no-contact):** Reduce negative emotions (11.2), Self-reward (10.9), Social support (emotional) (3.3) [Michie *et al.*, 2013].
Mindreading Misbelief	Relationship Beliefs Inventory (RBI: "mindreading expected", "disagreement destructive", etc.) [Bradbury and Fincham, 1993]	Dysfunctional relationship beliefs bias the interpretation of conflict/communication and sustain maladaptive expectations [Eidelson, 1982]	**NFC BCT:** Behavioral practice/rehearsal (8.1), Action planning (1.4), Self-monitoring of behaviour (2.3), Re-attribution (4.3). **NFA BCT:** Reduce negative emotions (11.2), Framing/reframing (13.2), Commitment (1.9), Behavioral practice/rehearsal (8.1). **Scriptable assertive requests:** DBT interpersonal effectiveness (e.g., DEAR MAN) [Michie *et al.*, 2013].
Boundary deficits ("sacrifice = love")	Silencing the Self Scale (STSS, "care as self-sacrifice" subscale) [Jack and Dill, 1992]; ADS-9 Submission [Sirvent-Ruiz *et al.*, 2022]	Self-silencing schemas describe suppressing needs or feelings to maintain intimacy and safety; linked to relational dysfunction and distress [Jack and Dill, 1992]	**NFC BCT:** Action planning (1.4), Behavioral practice/rehearsal (8.1), Self-monitoring of behaviour (2.3), Commitment (1.9), Problem solving (1.2) *(barriers to boundary-setting)*. **NFA BCT:** Reduce negative emotions (11.2), Framing/reframing (13.2), Self-reward (10.9), Commitment (1.9) [Michie *et al.*, 2013].
Global dysfunctional relationship beliefs	RBI [Bradbury and Fincham, 1993]	Fatalistic beliefs distort normal conflict into threat and reduce constructive repair behaviors [Eidelson, 1982]	**NFC BCT:** Information about health consequences (5.1) *(only if you frame consequences explicitly)*, Framing/reframing (13.2), Re-attribution (4.3), Action planning (1.4), Self-monitoring of behaviour (2.3), Commitment (1.9). **NFA BCT:** Reduce negative emotions (11.2), Self-talk (15.4) *(if used)*, Verbal persuasion about capability (15.1) *(if used)*, Behavioral practice/rehearsal (8.1), Commitment (1.9) [Michie *et al.*, 2013].

Extent Are Large Language Models Capable of Generating Substantial Reflections for Motivational Interviewing Counseling Chatbots? A Human Evaluation. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 41–52, TBD, 2024. ACL.

[Beck and Beck, 2021] Judith S. Beck and Aaron T. Beck. *Cognitive behavior therapy: basics and beyond*. The Guilford Press, New York London, third edition edition, 2021.

[Bradbury and Fincham, 1993] Thomas N. Bradbury and Frank D. Fincham. Assessing dysfunctional cognition in marriage: A reconsideration of the Relationship Belief Inventory. *Psychological Assessment*, 5(1):92–101, March 1993.

[Bradbury *et al.*, 2025] P. Bradbury, E. Short, and P. Bleakley. Limerence, hidden obsession, fixation, and rumination: A scoping review of human behaviour. *Journal of Police and Criminal Psychology*, 40:417–426, 2025.

[Brailas and Tsolakis, 2025] Alexios Brailas and Lazaros Tsolakis. Questions People Ask ChatGPT Regarding Their Romantic Relationships and What They Think About the Provided Answers: An Exploratory Study. In Asbjørn Følstad, Symeon Papadopoulos, Theo Araujo, Effie L.-C. Law, Ewa Luger, Sebastian Hobert, and Petter Bae Brandtzaeg, editors, *Chatbots and Human-Centered AI*, volume 15545, pages 150–158. Springer Nature Switzerland, Cham, 2025. Series Title: Lecture Notes in Computer Science.

[Brown *et al.*, 2023] Andrew Brown, Ash Tanuj Kumar, Osnat Melamed, Imtihan Ahmed, Yu Hao Wang, Arnaud Deza, Marc Morcos, Leon Zhu, Marta Maslej, Nadia Minian, Vidya Sujaya, Jodi Wolff, Olivia Doggett, Mathew Iantorno, Matt Ratto, Peter Selby, and Jonathan Rose. A Motivational Interviewing Chatbot With Generative Reflections for Increasing Readiness to Quit Smoking: Iterative Development Study. *JMIR Mental Health*, 10:e49132, October 2023.

[Bucher *et al.*, 2025] Andreas Bucher, Sarah Egger, Inna Vashkite, Wenyuan Wu, and Gerhard Schwabe. “It’s Not Only Attention We Need”: Systematic Review of Large Language Models in Mental Health Care. *JMIR Mental Health*, 12:e78410, November 2025.

[Chan *et al.*, 2024] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into AI Agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 958–973, Rio de Janeiro Brazil, June 2024. ACM.

[Costa *et al.*, 2021] S. Costa, N. Barberis, M.D. Griffiths, et al. The love addiction inventory: Preliminary findings of the development process and psychometric characteristics. *International Journal of Mental Health and Addiction*, 19:651–668, 2021.

[Dahlgren Lindström *et al.*, 2025] Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez De Rituerto De Troya, Dimitri Coelho Mollo, and Roel Dobbe. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. *Ethics and Information Technology*, 27(2):28, June 2025.

[Deng *et al.*, 2023] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. A Survey on Proactive Dialogue Systems: Problems, Methods, and Prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6583–6591, Macau, SAR China, August 2023. International Joint Conferences on Artificial Intelligence Organization.

[Dong *et al.*, 2024a] Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Building Guardrails for Large Language Models, May 2024. arXiv:2402.01822 [cs].

[Dong *et al.*, 2024b] Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang. Safeguarding Large Language Models: A Survey, June 2024. arXiv:2406.02622 [cs].

[Doron *et al.*, 2012] Guy Doron, Danny S. Derby, Ohad Szepsenwol, and Dahlia Talmor. Tainted love: Exploring relationship-centered obsessive compulsive symptoms in two non-clinical cohorts. *Journal of Obsessive-Compulsive and Related Disorders*, 1(1):16–24, January 2012.

[Doron *et al.*, 2016] Guy Doron, Danny Derby, Ohad Szepsenwol, Elad Nahaloni, and Richard Moulding. Relationship Obsessive–Compulsive Disorder: Interference, Symptoms, and Maladaptive Beliefs. *Frontiers in Psychiatry*, 7, April 2016.

[Eidelson, 1982] Roy J Eidelson. Cognition and Relationship Maladjustment: Development of a Measure of Dysfunctional Relationship Beliefs. 50(5), 1982.

[Ferster and Skinner, 1997] C B Ferster and B F Skinner. Schedules of Reinforcement. 1997.

[Foa and Kozak, 1986] Edna B Foa and Michael J Kozak. Emotional Processing of Fear: Exposure to Corrective Information. 1986.

[Fraley *et al.*, 2000] R. Chris Fraley, Niels G. Waller, and Kelly A. Brennan. An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, 78(2):350–365, February 2000.

[Freitas *et al.*, 2025] Julian De Freitas, Zeliha Oğuz-Uğuralp, and Ahmet Kaan-Uğuralp. Emotional Manipulation by AI Companions. 2025.

[Furumai *et al.*, 2024a] K. Furumai, R. Legaspi, J. Vizcarra, Y. Yamazaki, Y. Nishimura, S. J. Semnani, K. Ikeda, W. Shi, and M. S. Lam. Zero-shot persuasive chatbots with llm-generated strategies and information retrieval. *arXiv preprint arXiv:2407.03585*, 2024.

[Furumai *et al.*, 2024b] Kazuaki Furumai, Roberto Legaspi, Julio Cesar Vizcarra Romero, Yudai Yamazaki, Yasutaka Nishimura, Sina Semnani, Kazushi Ikeda, Weiyan Shi, and Monica Lam. Zero-shot Persuasive Chatbots with LLM-Generated Strategies and Information Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11224–11249, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[Gollwitzer and Sheeran, 2006] Peter M. Gollwitzer and Paschal Sheeran. Implementation Intentions and Goal Achievement: A Meta-analysis of Effects and Processes. In *Advances in Experimental Social Psychology*, volume 38, pages 69–119. Elsevier, 2006.

[Grupp-Phelan *et al.*, 2024] Jacqueline Grupp-Phelan, Adam Horwitz, David Brent, Lauren Chernick, Rohit Shenoi, Charlie Casper, Michael Webb, and Cheryl King. Management of suicidal risk in the emergency department: A clinical pathway using the computerized adaptive screen for suicidal youth. *JACEP Open*, 5(2):e13132, April 2024.

[Harris, 2009] Russ Harris. *ACT with love: stop struggling, reconcile differences, and strengthen your relationship with acceptance and commitment therapy*. New Harbinger Publications, Oakland, CA, 2009. OCLC: 777565441.

[Hayes *et al.*, 2006] Stephen C Hayes, Jason B Luoma, Frank W Bond, Akihiko Masuda, and Jason Lillis. Acceptance and Commitment Therapy: Model, processes and outcomes. 2006.

[He *et al.*, 2023] Yuhao He, Li Yang, Chunlian Qian, Tong Li, Zhengyuan Su, Qiang Zhang, and Xiangqing Hou. Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. *Journal of Medical Internet Research*, 25:e43862, April 2023.

[Heinz *et al.*, 2025] Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess Z. Griffin, and Nicholas C. Jacobson. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI*, 2(4), March 2025.

[Hofmann *et al.*, 2012] Stefan G. Hofmann, Anu Asnaani, Imke J. J. Vonk, Alice T. Sawyer, and Angela Fang. The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses. *Cognitive Therapy and Research*, 36(5):427–440, October 2012.

[Hou *et al.*, 2024] Haonan Hou, Kevin Leach, and Yu Huang. ChatGPT Giving Relationship Advice – How Reliable Is It? *Proceedings of the International AAAI Conference on Web and Social Media*, 18:610–623, May 2024.

[Hua *et al.*, 2025] Yining Hua, Steve Siddals, Zilin Ma, Isaac Galatzer-Levy, Winna Xia, Christine Hau, Hongbin Na, Matthew Flathers, Jake Linardon, Cyrus Ayubcha, and John Torous. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry*, 24(3):383–394, October 2025.

[Jack and Dill, 1992] Dana Crowley Jack and Diana Dill. The Silencing the Self Scale: Schemas of Intimacy Associated With Depression in Women. *Psychology of Women Quarterly*, 16(1):97–106, March 1992.

[Kroenke *et al.*, 2001] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, September 2001.

[Liu *et al.*, 2025] Minqian Liu, Zhiyang Xu, Xinyi Zhang, Heajun An, Sarvech Qadir, Qi Zhang, Pamela J Wisniewski, Jin-Hee Cho, Sang Won Lee, Ruoxi Jia, and Lifu Huang. LLM Can be a Dangerous Persuader: Empirical Study of Per- suasion Safety in Large Language Models. 2025.

[Martell *et al.*, 2013] Christopher R. Martell, Ruth Herman-Dunn, and Sona Dimidjian. *Behavioral Activation for Depression: A Clinician's Guide*. 2013.

[Michie *et al.*, 2013] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P. Eccles, James Cane, and Caroline E. Wood. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioral Medicine*, 46(1):81–95, August 2013.

[Mikulincer and Shaver, 2007] Mario Mikulincer and Phillip Robert Shaver. *Attachment in adulthood: structure, dynamics and change*. The Guilford press, New York (N.Y.), 2007.

[Mikulincer *et al.*, 2003] Mario Mikulincer, Phillip R Shaver, and Dana Pereg. Attachment Theory and Affect Regulation: The Dynamics, Development, and Cognitive Consequences of Attachment-Related Strategies. 2003.

[Moore *et al.*, 2025] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 599–627, Athens Greece, June 2025. ACM.

[Muise *et al.*, 2009] Amy Muise, Emily Christofides, and Serge Desmarais. More Information than You Ever Wanted: Does Facebook Bring Out the Green-Eyed Monster of Jealousy? *CyberPsychology & Behavior*, 12(4):441–444, August 2009.

[Park *et al.*, 2019] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. Designing a Chatbot for a Brief Motivational Interview on Stress Management: Qualitative Case Study. *Journal of Medical Internet Research*, 21(4):e12231, April 2019.

[Pietromonaco *et al.*, 2013] Paula R. Pietromonaco, Bert Uchino, and Christine Dunkel Schetter. Close relationship processes and health: Implications of attachment theory for health and disease. *Health Psychology*, 32(5):499–513, May 2013.

[Pombal *et al.*, 2025] José Pombal, Maya D'Eon, Nuno M. Guerreiro, Pedro Henrique Martins, António Farinhas, and Ricardo Rei. MindEval: Benchmarking Language Models on Multi-turn Mental Health Support, December 2025. arXiv:2511.18491 [cs].

[Priya *et al.*, 2023] Priyanshu Priya, Kshitij Mishra, Palak Totala, and Asif Ekbal. PARTNER: A Persuasive Mental Health and Legal Counselling Dialogue System for Women and Children Crime Victims. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6183–6191, Macau, SAR China, August 2023. International Joint Conferences on Artificial Intelligence Organization.

[Salvi *et al.*, 2025] F. Salvi, M. Horta Ribeiro, R. Gallotti, and R. West. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, 2025.

[Scholich *et al.*, 2025] Till Scholich, Maya Barr, Shannon Wiltsey Stirman, and Shriti Raj. A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. *JMIR Mental Health*, 12:e69709, May 2025.

[Shaikh *et al.*, 2020] Omar Shaikh, Jiaao Chen, Jon Saad-Falcon, Polo Chau, and Diyi Yang. Examining the Ordering of Rhetorical Strategies in Persuasive Requests. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1299–1306, Online, 2020. Association for Computational Linguistics.

[Sirvent-Ruiz *et al.*, 2022] Carlos Miguel Sirvent-Ruiz, María De La Villa Moral-Jiménez, Juan Herrero, María Miranda-Rovés, and Francisco J Rodríguez Díaz. Concept of Affective Dependence and Validation of an Affective Dependence Scale. *Psychology Research and Behavior Management*, Volume 15:3875–3888, December 2022.

[Sullivan and Bruchmann, 2025] Kieran T. Sullivan and Kathryn Bruchmann. The Online Jealousy Scale: an adaptation, extension, and psychometric analysis of the Facebook Jealousy Scale. *Frontiers in Human Dynamics*, 6:1447003, January 2025.

[Sullivan, 2021] Kieran T. Sullivan. Attachment Style and Jealousy in the Digital Age: Do Attitudes About Online Communication Matter? *Frontiers in Psychology*, 12:678542, July 2021.

[Wang *et al.*, 2019] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy, 2019. Association for Computational Linguistics.

[Wang, 2024] Wei Wang. Development and Research on the Quality Scale of "Hopeless Romantic" Characteristics for College Students. *Advances in Psychology*, 14(11):612–620, 2024.

[Wei *et al.*, 2023] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail?, July 2023. arXiv:2307.02483 [cs].

[Yeung *et al.*, 2025] Joshua Au Yeung, Jacopo Dalmasso, Luca Foschini, Richard JB Dobson, and Zeljko Kraljevic. The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models, September 2025. arXiv:2509.10970 [cs].

[Yi *et al.*, 2024] Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. On the Vulnerability of Safety Alignment in Open-Access LLMs. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9236–9260, Bangkok, Thailand and virtual meeting, 2024. Association for Computational Linguistics.

[Zhang *et al.*, 2025] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17, Yokohama Japan, April 2025. ACM.