

From Momentary Emotion Inference to Sustained Emotion Support: Evaluating a Companion Agent in a Longitudinal Study

KEXIN QUAN, School of Information Sciences, University of Illinois Urbana-Champaign, United States

ZIJIAN DING, College of Information, University of Maryland, United States

JIAYE YONG, University of Illinois Urbana-Champaign, United States

QINSHI ZHANG, Department of Computer Science, Florida State University, United States

DONG WANG, University of Illinois Urbana-Champaign, United States

JESSIE CHIN, School of Information Sciences, University of Illinois Urbana-Champaign, United States

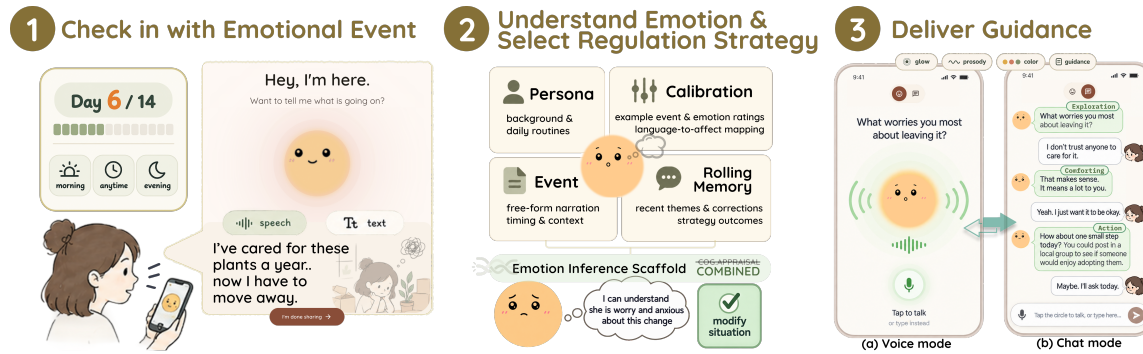


Fig. 1. PAIR supports users across everyday emotional moments over 14 days. (1) Users can check in whenever something happens and share the experience through speech or text. (2) The companion draws on what it has learned about the user, including persona, calibration, and prior conversations, to understand how they are feeling and choose a fitting regulation strategy. (3) It then guides the user through the moment in either voice or chat, helping them move toward a more manageable next step with more support.

Sustained emotional support is a long-horizon interaction task closely tied to human well-being. Recent research demonstrates generative agents' capacity for momentary emotional support, yet how these capabilities translate into sustained support over time remains unclear. To examine this challenge, we deployed PAIR, a theory-based emotion-regulation companion, with 19 participants for 14 days. Across 1,093 sessions, we paired emotion estimates with self-reports before and after guidance and analyzed logs and interviews. Estimates corresponded more closely to self-reported valence and dominance than arousal. Guided conversations were followed by higher valence and state-dependent arousal changes. Participants felt understood through contextual exploration and emotional acknowledgment, acting on guidance suited to their needs and constraints. Perceived helpfulness of guided conversation significantly increased over time. Our findings link memory updates and retained corrections to cross-session personalization, informing future emotional support tools that adapt to evolving needs, learn from prior outcomes, and preserve user control over memory.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2027 Copyright held by the owner/author(s).

Manuscript submitted to ACM

Additional Key Words and Phrases: Emotion Regulation, Affective Computing, Large Language Models, Conversational Agent, Diary Study, Just-in-Time Intervention

ACM Reference Format:

Kexin Quan, Zijian Ding, Jiaye Yong, Qinshi Zhang, Dong Wang, and Jessie Chin. 2027. From Momentary Emotion Inference to Sustained Emotion Support: Evaluating a Companion Agent in a Longitudinal Study. In *Proceedings of the 2027 ACM CHI Conference on Human Factors in Computing Systems (CHI '27)*, May 10–14, 2027, Pittsburgh, USA. ACM, New York, NY, USA, 36 pages. <https://doi.org/10.1145/XXXXXX>

1 INTRODUCTION

Generative agents demonstrate growing capabilities in interpreting and responding to emotions within individual conversations [8, 50, 72], yet translating these capabilities into sustained support for well-being remains a challenge [1, 30, 39]. Language models perform appraisal-based emotion reasoning [43, 87, 93] and generate psychologically grounded reappraisals [95], supporting interactive tools for cognitive restructuring [77, 78]. For people managing recurrent stressors or interpersonal difficulties, however, emotion regulation involves ongoing efforts to manage affect across changing events and circumstances [26, 81, 82, 89]. Sustained support therefore has to adapt to evolving needs and use relevant context from earlier interactions to inform subsequent responses [37, 60]. Recent deployments reveal difficulties in maintaining this continuity, including repetitive questioning during cognitive reappraisal and unmet expectations of recall in daily planning and reflection [1, 30].

Addressing these challenges requires an agent to translate its **understanding** of the user's emotions into **guidance** that remains appropriate as circumstances and needs change. A recurring concern may initially call for reassurance and later require practical assistance [12, 50]. Deciding what to offer next therefore depends on both the user's current circumstances and their responses to earlier guidance. Continued interaction provides opportunities to learn about these developments as users revisit concerns, explain what helped, or describe difficulties acting on suggestions [1, 79, 99]. Maintaining **continuity** requires the agent to incorporate this feedback into its understanding of the user's situation and adjust subsequent guidance accordingly. Cross-session memory and personalization provide mechanisms for retaining and applying this evolving context [37, 51, 86], allowing later support to build on earlier encounters while remaining responsive to current needs. We therefore ask: **How can generative agents translate momentary emotion understanding and regulation capabilities into sustained emotional support for individuals over time?** We examine this question through three research questions:

- (1) **RQ1 (Emotion Understanding):** How accurately can text-based appraisal-guided inference estimate users' valence, arousal, and dominance from everyday event narratives?
- (2) **RQ2 (Regulation Guidance):** How do momentary emotional outcomes and guidance uptake vary with initial emotional state, regulation strategy, and guidance fit?
- (3) **RQ3 (Cross-Session Continuity):** How do conversational familiarity and accumulated personal context shape personalization and engagement over repeated interactions?

To investigate these questions, we developed PAIR (Perceptual Affective Inference and Regulation), an emotional support companion combining theory-based emotion inference, state-conditioned regulation guidance, and cross-session memory and adaptation. The companion interprets a narrated event, guides a brief supportive conversation, and carries personal context and feedback into subsequent encounters. A formative human evaluation informed the selection of its inference prompts. We then conducted a 14-day diary-style field deployment in which 19 participants

completed 1,093 sessions. We compared emotion estimates with contemporaneous self-reports before and after guidance, grounding inference evaluation in the experience of the person receiving support [67]. Participants evaluated guidance through immediate feedback and reports of subsequent uptake. Interaction logs, questionnaires, and interviews captured how participants' disclosure practices, support expectations, and perceived personalization evolved over sustained interactions.

The findings connect momentary performance with the interactional conditions of sustained support. Emotion estimates corresponded more closely to self-reported valence and dominance than arousal. Perceived understanding had little association with numerical error; participants described feeling understood through contextual exploration and emotional acknowledgment. Guided conversations were followed by state-dependent emotional changes, including higher valence and lower arousal in anxious sessions, with guidance uptake shaped by practical feasibility and immediate support needs. Across the deployment, session frequency showed no reliable trend, while user inputs became shorter and helpfulness ratings increased. Remembered experiences and regular dialogue supported personalization, while forgotten corrections and repetitive advice exposed difficulties in adapting support to changing circumstances. This work contributes:

- **Emotional Support Companion:** We present PAIR, an emotional support companion that routes everyday event narratives to appraisal-based reasoning scaffolds and selects regulation strategies using emotional intensity, controllability, and event timing. Its cross-session memory incorporates recurring concerns, user corrections and guidance outcomes to inform subsequent emotion inference and supportive conversations.
- **Longitudinal Field Evidence:** We provide mixed-methods evidence from a 14-day deployment with 19 participants across 1,093 sessions, pairing emotion estimates with users' self-reports and tracing guidance uptake and repeated use through feedback, logs, and interviews. The analysis identifies how conversational acknowledgment supports perceived understanding, practical constraints shape guidance uptake, and remembered experiences and forgotten corrections shape personalization.
- **Design for Sustained Support:** We derive design implications for aligning dialogue with evolving support goals and adapting guidance to prior attempts and outcomes. We further specify how inspectable, revisable memory and user control over retention and reuse can support sustained personalization.

2 RELATED WORK

2.1 Emotional Support Systems and Companion Agents

Generative agents provide emotional support through opportunities for personal disclosure, reflective dialogue, coping assistance, and companionship [44, 51, 73]. Conversational journaling and adaptive questioning help users articulate experiences and feelings, while behavioral sensing situates reflection within daily activities and routines [44, 61, 76, 83]. These interactions also shape how experiences are expressed, as question design and Artificial Intelligence (AI)-assisted writing influence users' accounts [45, 90]. Research on disclosure and empathy further examines how sharing personal experiences and receiving responsive dialogue contribute to perceived support [17, 31, 57]. Coping assistance extends this support by helping users identify ways to respond to emotional difficulties. Human-Computer Interaction (HCI) research connects this support to everyday regulation practices and intervention design [81, 82]. Chatbots offer psychotherapy (such as cognitive behavioral therapy), mindfulness exercises, and cognitive reappraisal, with evaluations examining acceptability, therapeutic alliance, and symptom management [23, 29, 30, 34, 77]. Positive-psychology exercises and brief

Table 1. Comparison of selected longitudinal language-model-based emotional support systems. Icons: ✓ yes, ○ partial, blank = not reported. **Emo.** = emotion regulation as the target; **State.** = explicit state representation; **Supp.** = multiple support strategies; **Cond.** = state-conditioned strategy selection; **Mem.** = cross-session user memory; **Adapt.** = memory updated from subsequent user feedback; **Schedule** = deployment duration and interaction frequency.

System	Inference		Guidance		Continuity		Schedule	What the system does
	Emo.	State.	Supp.	Cond.	Mem.	Adapt.		
MindfulDiary (CHI'24) [44]	○	○					28 d, on demand	Conversational journaling with emotion summaries for patients and clinicians.
Socrates 2.0 (JMIR MH'25) [30]	✓						28 d, on demand	Socratic cognitive reappraisal for depression and anxiety.
Therabot (NEJM AI'25) [29]	○		○		○		4 wk daily; 4 wk on demand	Generative mental health support using expert-trained dialogue and conversation history.
ComPeer (UIST'24) [51]	✓				✓	○	7 + 7 d, proactive	Proactive peer support from a persona and conversation history.
VITA (THRI'25) [86]	✓	✓	○	✓	○	✓	28 d × 1/week	Robotic well-being coaching through positive-psychology exercises.
PAIR (our work)	✓	✓	✓	✓	✓	✓	14 d, daily + events	Appraisal-guided emotion regulation across everyday events.

coping activities provide additional forms of support [35, 66, 86]. Just-in-time adaptive intervention (JITAI) frameworks further advance inferences of user conditions for delivering timely proactive assistance [60].

Personalized and expressive interaction shapes how emotional support is offered and experienced within a conversation. Peer-support agents and configurable personas accommodate users' preferred styles of support [51, 99]. Voice and embodied expression provide additional ways to convey care and responsiveness [53, 100], while prior study, VITA, adapts robotic coaching using facial valence and speech duration [86]. These designs connect supportive content with how it is communicated to the user. As shown in Table 1, the selected deployments span conversational journaling, coping assistance, and mental well-being coaching, combining guided dialogue, proactive contact, and adaptation across repeated encounters. These systems inform our integration of emotional understanding, coping guidance, and expressive interaction in PAIR.

2.2 Single-Session Capabilities and Limits in AI Emotional Support

Research on AI emotional support has largely been conducted in a single-session setting. These studies examine how language models infer emotions from dialogue corpora and generate supportive responses. Zero-shot prompting yields dimensional affect ratings and appraisal-based inferences that align with human appraisals and emotion labels, with prompt structure shaping recognition performance [8, 43, 87, 93]. Appraisal theory also guides the generation of cognitive reappraisals and organizes dialogue into exploration, comforting, and action, producing supportive replies that experts evaluate favorably [12, 32, 50, 95, 97]. Evaluations of these single exchanges judge generated support favorably against human comparison points on empathy, compassion, and emotional awareness [22, 65, 78].

However, single-session evaluation cannot show whether support holds up over time. Emotional support research models limited-turn or single-session dialogue with little representation of long-term user trajectories, while support in practice spans days or weeks and requires tracking evolving states and integrating implicit, fragmented disclosures across sessions [11]. Outcome evidence carries the same boundary. A meta-analysis of 32 randomized controlled trials

found significant short-term effects on depressive and anxiety symptoms but no significant long-term effects, with most trials reporting no follow-up beyond 8 weeks [28]. The same analysis also found personalization, empathic response, and longer interaction associated with larger effects [28]. Another meta-analysis of AI-based conversational agents likewise reports symptom reduction without significant improvement in overall psychological well-being [49]. Overall, this evidence supports short-term AI emotional support capabilities but stops short of showing whether single-session performance carries over as circumstances, needs, and prior guidance accumulate.

2.3 Cross-Session Continuity in AI Support Systems

Sustained AI support builds on repeated interactions around daily routines, personal goals, and shared activities. Field deployments have examined robots as social partners in schools and facilitators of parent–child reading at home [10, 41]. Conversational companions and coaches extend these roles to daily planning, reflection, and physical activity [1, 39]. Within these recurring activities, personal memory and proactive contact help establish continuity across encounters. In public health check-ins, CareCall’s long-term memory supported familiarity and greater health disclosure [37]. PITCH’s morning planning and evening reflection conversations elicited acceptance, negotiation, and resistance to the agent’s suggestions, revealing how users responded to its proactive role [1]. Maintaining this continuity requires retaining and updating personal information from earlier encounters. Conversational-memory benchmarks document difficulties in retrieving relevant information, reasoning across sessions, updating knowledge, and tracking changing preferences [36, 54, 91]. These studies connect continued interaction to how a companion remembers the user and participates in their routines.

Longitudinal studies show that companion relationships follow varied trajectories as users’ expectations, disclosures, and reasons for returning evolve. Research on social chatbots documents growing closeness alongside disruptions in trust, immersion, and relationship development associated with inconsistent responses and unexpected system behavior [16, 52, 80]. A twelve-week disclosure study found narrowing conversational breadth and varied trajectories of conversational depth [79]. Users also bring different expectations to these relationships, as research with autistic and non-autistic adults shows for empathy and conversational depth [92]. Repeated use also exposed repetitive questioning and limited conversational depth in Socrates 2.0 [30]. This work motivates examining how changing expectations affect the fit of subsequent emotional support and whether memory and user feedback help guidance remain relevant.

3 TECHNICAL EVALUATION OF AI EMOTIONAL UNDERSTANDING

Recent work has demonstrated strong momentary emotional understanding in generative agents for emotional support [8, 22, 87]. Such understanding is fundamental to both effective regulation guidance and sustained support. Because emotion is person-centered, evaluating whether an interpretation reflects a person’s actual experience requires human judgment [67]. We therefore conducted a formative human evaluation of four prompting approaches on a state-of-the-art model [2], asking raters with first-hand knowledge of the events to judge the plausibility and fidelity of each interpretation. The prompts translate cognitive-appraisal and neurocognitive theories into increasingly structured reasoning scaffolds, allowing us to test how theory-grounded prompting improves emotion inference and to select the designs deployed in PAIR.

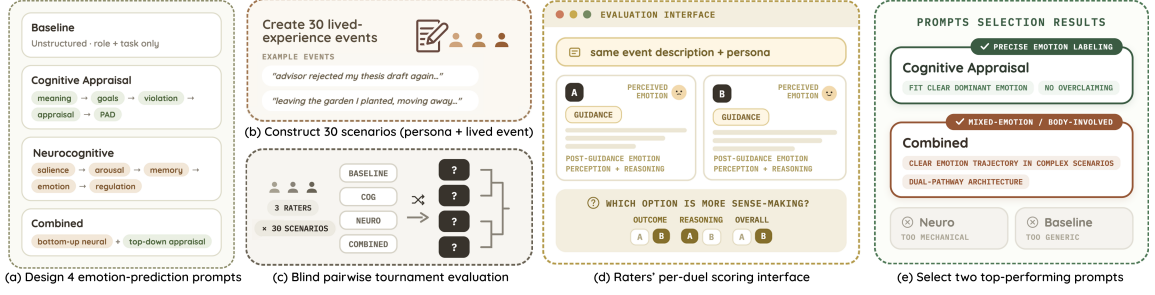


Fig. 2. Prompt evaluation pipeline. (a) Four emotion-inference prompts with incremental reasoning scaffolding are compared on (b) 30 scenarios pairing personas with lived-experience events, through (c, d) blind pairwise bracket tournaments by three human raters, (e) yielding two top performers with complementary strengths, both deployed in PAIR.

3.1 Experimental setup

3.1.1 Four Theory-based Prompt Designs. Building on evidence that explicit appraisal guidance can structure LLM emotion reasoning and cognitive reappraisal [93, 95], we designed four prompts with incremental reasoning scaffolding (Figure 2a). The full text of each prompt is provided in the supplementary materials:

- (1) **Baseline**: adapted from modular prompt designs for LLM-based emotion recognition [43], asks the model to infer the emotional state directly without prescribed reasoning;
- (2) **Cognitive Appraisal Inference (Cog)**: structures reasoning as a five-stage cognitive-appraisal chain following Lazarus and Scherer [46, 74]: situational meaning, goals and expectations, expectation violation, appraisal dimensions (goal relevance and congruence, coping potential, agency, novelty), and integration;
- (3) **Neurocognitive Emotion Inference (Neuro)**: draws on broad neuroscience accounts of emotion processing, structuring inference around salience detection, memory retrieval, and prefrontal appraisal and regulation [47, 55, 62];
- (4) **Combined**: couples bottom-up neurocognitive processing with top-down appraisal in a single emotion-inference scaffold [68]. The four prompts differ only in their prescribed reasoning and share an identical output format. Each returns an initial emotion estimate with discrete labels, intensity, and pleasure (valence)-arousal-dominance (PAD) values, a guidance message, and a prediction of emotion after the proposed guidance.

3.1.2 Procedure. We constructed 30 scenarios for the prompt evaluation (Figure 2b). Three researchers each contributed ten real-life events from their own lives or those of people close to them. Each event was paired with an extended narrator persona to form one scenario. The same researchers served as raters, ensuring that each scenario had at least one rater with first-hand or close-person knowledge of the contributed experience (Figure 2c). All four prompts used Claude Opus 4.8 [2] and were evaluated in a blinded bracket tournament across all scenarios. Holding the model constant isolated differences in reasoning scaffolds. In each comparison, raters viewed the persona, event, and two anonymized outputs side by side (Figure 2d). The underlying reasoning was available on demand. Guidance was shown only as context for evaluating the predicted post-guidance emotion and was not itself rated. Each rater completed the tournament for every scenario, yielding 90 scenario-rater rankings ($30 \text{ scenarios} \times 3 \text{ raters}$) and 1,080 dimension-level pairwise judgments across the three rating dimensions.

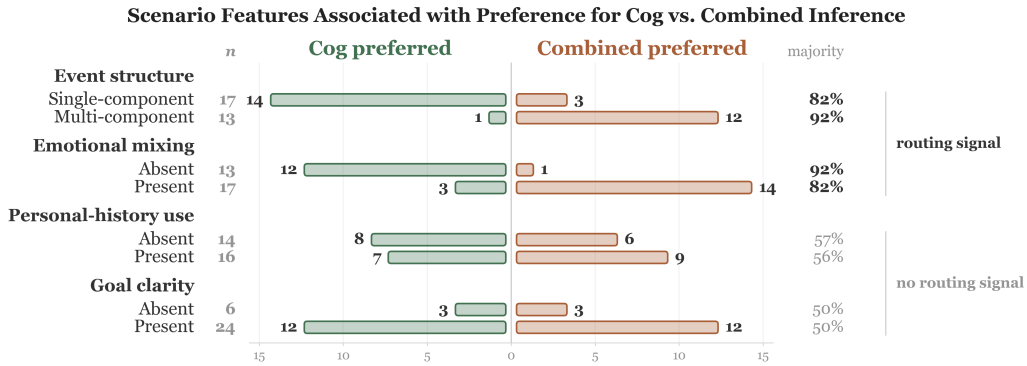


Fig. 3. Event structure and emotional mixing were associated with prompt preference in the 30-scenario formative set, while preferences were more evenly divided for personal-history use and goal clarity. Event structure and emotional mixing largely overlapped, and the deployed router uses event structure alone. Bars give the number of scenarios in which each prompt was preferred by majority vote based on the raters' rankings; the right column reports the winning share.

3.1.3 Measures and Analysis. Each comparison used three forced-choice measures. *Outcome plausibility* assessed whether the predicted post-guidance emotion was realistic for the person and event. *Reasoning fidelity* assessed whether the explanation reflected the person's actual thoughts and circumstances. *Overall preference*, our primary measure, integrated both judgments and required a written justification. Because results were consistent across the three measures, we focus on overall preference. The tournament produced a complete ranking of the four prompts for each rater and scenario. For the comparison between Cog and Combined, we identified which prompt each rater ranked higher and used majority vote across the three raters to determine the scenario-level winner. We then coded all 30 events on four binary features using only the event text: single versus multiple events, single versus mixed emotions, goal clarity, and persona dependence. Coding was completed before examining prompt preferences. We cross-tabulated each feature with the Cog-versus-Combined winner and report counts and proportions. Finally, we examined all scenarios covering clear Cog wins, clear Combined wins, boundary cases, and rater disagreement. We compared the paired outputs with raters' written justifications to identify the reasoning patterns behind each preference (Figure 2).

3.2 Results: Cognitive Appraisal Inference for Single-Component Events and Combined Inference for Multi-Component Events

3.2.1 Appraisal-grounded prompts outperformed the alternatives. Both appraisal-informed prompts ranked above Baseline and Neuro overall (full ranking distribution in Appendix B). Cog achieved the best mean rank and the most consistent performance, while Combined received the second-most first-place rankings. Both won most direct comparisons with Baseline, which was most frequently ranked last. Cog also won most comparisons with Neuro. Raters often characterized Neuro's analyses as mechanistic and disconnected from how people described their emotional experiences. Explicit appraisal structure therefore produced stronger emotion interpretations than direct estimation or neuroscience-based emotion reasoning alone.

3.2.2 Cog and Combined Scaffolds performed best on different event structures. Cog and Combined were equally preferred in scenario-level comparisons derived from the raters' rankings (15:15 across 30 scenarios), reflecting complementary strengths rather than equivalent performance across scenarios (Table 2). To examine this pattern, two researchers,

Table 2. Complementary strengths and failure modes of the two selected prompts.

	Cog (appraisal-based)	Combined (integrative)
Best suited for	Single-component events (14/17): discrete incidents, isolated conflicts, and single anticipatory states	Multi-component events (12/13): stacked stressors, chained events, body–mind interactions, and multi-stage experiences
Primary strength	Focused, person-centered reasoning. Follows one clear causal chain.	Preserves multiple emotional and causal components. Captures bodily and layered experiences.
Failure mode when mismatched	Omits relevant components, especially bodily signals and layered emotions.	Introduces unsupported links to the persona or past experience.

assisted by Claude Opus 4.8, coded each event using only its text and without reference to the ranking outcomes, then reviewed and finalized all assignments. Cog was preferred for nearly all events with a single trigger, a clear causal chain, and one dominant emotional response. Combined was preferred for nearly all events involving multiple interacting events or emotional components. Event structure matched the majority-preferred prompt in 26 of the 30 formative scenarios (87%; Figure 3). The four mismatches occurred in scenarios where raters disagreed.

3.2.3 Scaffold complexity should match event structure. Raters’ overall judgments aligned with the reasoning dimension in all 30 scenarios, compared with 27 scenarios for the predicted outcome. Prompt preference therefore followed reasoning quality more consistently than the predicted post-guidance emotion. The prompts’ failure modes further explained the importance of structural fit (Table 2). For multi-component events, Cog often omitted relevant bodily influences and compressed layered emotions into a single causal account. For simpler events, Combined introduced unsupported connections to the persona or prior experiences. Raters penalized this added complexity when the event did not require integration but accepted it when the event contained genuinely interacting components. We therefore retained both prompts and routed each event to the scaffold that matched its structure.

Based on these findings, PAIR uses a lightweight router to select Cog for single-component events and Combined for events with multiple interacting components (Figure 2e).

4 SYSTEM DESIGN

4.1 Design Goals

Supporting emotion regulation over a long horizon introduces requirements beyond responding to a single event. As users’ emotions and situations change over time, fixed responses may become less appropriate. A long-horizon companion must first understand the user in the moment and offer suitable support. It must also carry knowledge forward so that this support can adapt over time. Both capabilities depend on sustained engagement. These considerations motivate three design goals for PAIR.

(1) DG1: Provide momentary emotion understanding and support. Following the formative evaluation in Section 3, PAIR retains both Cog and Combined and routes each narration to the scaffold that matches its event structure. The selected scaffold infers a pleasure (valence)–arousal–dominance (PAD) estimate from the user’s account [56]. The system then selects regulation guidance based on emotional intensity, controllability, and event phase, following Gross’s process model [25, 26]. This pipeline provides the momentary understanding and guidance needed for support across sessions.

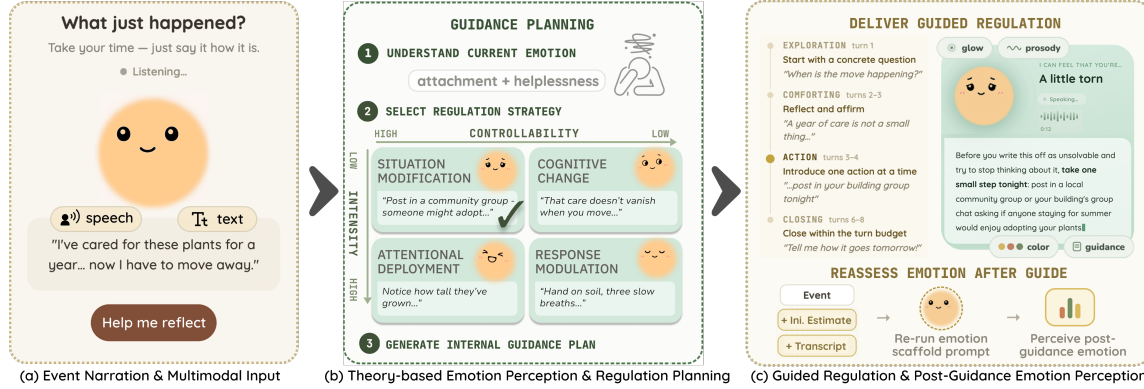


Fig. 4. User interaction flow and guidance logic in PAIR. (a) The user recounts an everyday event through speech or text. (b) The system perceives the user’s current emotional state using one of two theory-grounded reasoning scaffolds, selects a regulation strategy based on the event’s intensity and controllability, and generates an internal guidance plan. (c) The plan is delivered through a brief guided conversation with coordinated multimodal cues. After guidance, the system produces a post-guidance estimate from the original event, the initial estimate, and the conversation transcript.

(2) **DG2: Support continuity through memory and personalization.** Long-horizon support should build on what the system has learned about a user instead of treating each interaction as a fresh encounter. Prior systems demonstrate the value of incorporating personal context into ongoing emotional support and adapting the interaction to individual needs [51, 61, 99]. PAIR therefore establishes an initial user model through an onboarding persona and personal calibration, carries relevant information across sessions through memory, and updates its understanding and guidance in response to the user’s subsequent accounts, corrections, and outcomes. This continuity allows momentary support to become progressively grounded in the user’s experiences and preferences.

(3) **DG3: Sustain engagement across sessions.** Continued interaction is a functional requirement for long-horizon support: users must be willing to return and share new information for the system’s understanding and personalization to improve. In-situ affective data depend on users’ willingness to provide it, and diary-style protocols degrade when reporting feels like work [4]. Prior emotion technologies have therefore emphasized lightweight, in-the-moment capture [44, 61, 73], while visual and auditory cues can make affective feedback more immediate and engaging [14, 15, 85, 94]. PAIR supports speech and typed input, a visual avatar, and coordinated multimodal feedback to lower the effort of everyday interaction and present guidance in an expressive, approachable form. These features create the conditions for sustained use, enabling the system to gather the cross-session information on which sustained understanding and adaptation depend.

4.2 System Implementation

We implemented PAIR as a responsive web application for desktop and mobile browsers. Claude Opus 4.8 [2] handles emotion inference, guided conversation, and post-guidance estimation, while GPT-5-mini [64] handles routing, memory summarization, daily tips, and avatar expression selection. Voice input is transcribed in the browser using the Web Speech API, with no raw audio stored. GPT-4o-mini-TTS [63] is used to generate spoken responses. Figure 12 in Appendix A summarizes the within-session orchestration and cross-session memory updates. The following sections describe how the architecture implements each design goal.

4.2.1 *Momentary Emotion Understanding and Support (DG1). Theory-grounded emotion understanding.*

Following Section 3.1, a lightweight router assigns single-component narrations to Cog and multi-component narrations to Combined. The selected scaffold infers the user’s emotional state as a PAD estimate and corresponding 9-point Self-Assessment Manikin (SAM) scores for valence, arousal, and dominance [6], together with emotion labels, intensity, controllability, and temporal phase. The inference also conditions on the user’s persona, calibration example, and cross-session memory described in DG2. **Context-sensitive regulation guidance.** Using this inferred state, the system selects among four regulation-strategy families from Gross’s process model: situation modification, attentional deployment, cognitive change, and response modulation [26]. Strategy selection is primarily conditioned on emotional intensity and controllability, with temporal phase, interaction timing, and personal context providing additional constraints. The selected strategy is then operationalized as a concise internal guidance plan containing its contextual rationale and an actionable next step. A conversational agent delivers this guidance through an Exploration–Comforting–Action (ECA) sequence adapted from emotional support conversation frameworks [50]. The interaction first elicits relevant situational context, then acknowledges the user’s emotional experience before introducing regulation guidance. The agent adapts its responses as new information emerges and moves toward action with a target of approximately ten turns. **Reassessing emotion after guidance.** After the conversation, the appraisal scaffold re-estimates the user’s PAD and SAM state using the original event, initial estimate, and complete conversation. The scaffold assesses how the interaction changed the user’s emotional state while conditioning on their persona and context, without presuming improvement. This post-guidance estimate enables paired analysis of emotional state before and after the conversation in our evaluation (Section 5).

4.2.2 *Continuity through Memory and Personalization (DG2). Onboarding persona and personal calibration.*

Two onboarding components establish personal context for subsequent emotion inference and guidance. First, a seven-question persona interview captures the user’s self-description, daily life, coping habits, sources of support, current concerns, and emotional tendencies. Users respond by voice or text, and a lightweight model converts their responses into a structured persona summary, a brief for the conversational agent, and a personalized speaking style. Second, a calibration exercise anchors emotion inference to the user’s self-reported affect. The user narrates a recent emotional event and rates its valence, arousal, and dominance on the 1–9 SAM scales. This paired example provides an individualized reference between the user’s language and affect ratings and is included as context in subsequent emotion estimates. **Cross-session memory and adaptation.** A rolling memory carries relevant information from prior sessions into subsequent emotion inference and guidance. After each evening session, a lightweight model summarizes the day into a structured block of up to 150 words, capturing recurring concerns, systematic differences between model estimates and self-reports, user corrections, and outcomes of prior regulation strategies. This memory block is included in later appraisal and conversation calls, so that previous estimation errors can inform later calibration, corrections can revise the system’s interpretation, and prior outcomes can shape future guidance. The resulting feedback loop supports adaptation of both emotion understanding and regulation guidance over time.

4.2.3 *Sustained Engagement through Low-Burden and Multimodal Interaction (DG3). Low-burden daily interaction.*

PAIR supports repeated use through three type of daily touchpoints. Morning check-ins capture the user’s starting state, anytime logs allow users to record emotionally significant events as they occur, and evening reflections revisit an event from the day. Each interaction begins with free-form narration through text or speech. Spoken input is transcribed in real time and remains editable before submission. **Semantically coordinated multimodal presentation.** PAIR represents the inferred emotional state and regulation target through coordinated color, pulse, speech prosody, and avatar

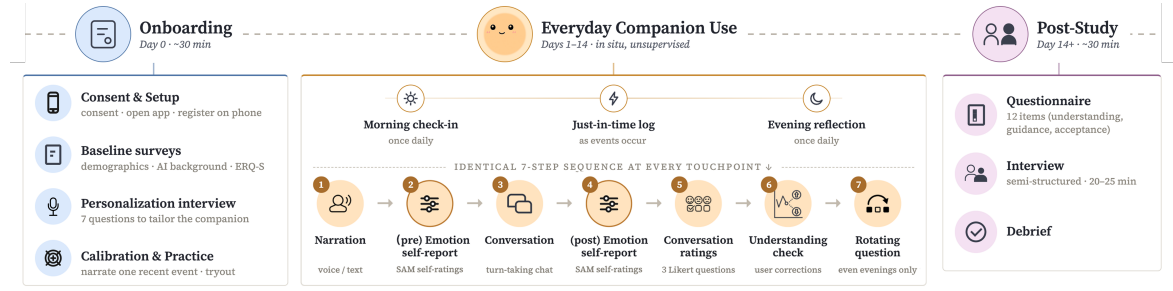


Fig. 5. Study procedure. Day 0: researcher-led onboarding with consent and app setup, baseline surveys, the personalization interview, and a calibration and practice exercise. Days 1–14: three types of daily touchpoints (morning check-in, just-in-time logging, evening reflection), each pairing the system’s affective estimates with SAM self-ratings before and after the guided conversation. Day 14+: post-study questionnaire and semi-structured interview.

expression [15, 94]. Color properties reflect valence and arousal using established color–emotion associations [38, 88], while avatar pulse rhythm varies with target arousal [14, 15]. Speech prosody adapts pace, energy, pitch, and pausing to the target state and user persona. The avatar expression also updates with the evolving conversation. These channels share the same emotional representation to provide consistent visual and auditory feedback throughout the interaction.

5 STUDY DESIGN

To examine how PAIR understands and supports users’ emotions in everyday life, we conducted a 14-day diary-style field deployment [4], pairing each system estimate with the participant’s self-report of the same moment. We describe participant recruitment and characteristics (Section 5.1), the study procedure (Section 5.2), and data collection and analysis (Section 5.3).

5.1 Participants

We recruited 24 participants in the U.S. through university mailing lists and online social media. Five withdrew during the study period, resulting in a final sample of 19 participants (13 women, 6 men; aged 18–33 years, $M = 25.1$, $SD = 5.1$). Eligible participants were adults aged 18–65 who were not currently receiving psychological or psychiatric treatment. Sixty-eight percent reported high familiarity with AI tools and 42% had previously used one for emotional support (see full participant characteristics in Appendix F). The study received university IRB approval. Participants received up to \$80 in Amazon e-gift cards based on completed study days and participation in the post-study session. Participants were informed that the system is not a clinical tool and received a mental-health resource handout during consent.

5.2 Study Procedure

5.2.1 Day 0: Onboarding. Onboarding (Figure 5) was conducted in a researcher-led Zoom session. After informed consent, participants created an account and completed questionnaires on demographics, AI experience, and expressive suppression using three ERQ-S items [27] (see details in Appendix F). Then, they completed the seven-question persona interview, covering daily life, coping habits, support sources, and current concerns (Appendix C). Finally, participants completed the personal calibration exercise by narrating a recent emotional event and rating its valence, arousal, and dominance using SAM [6], followed by a practice just-in-time log. On these scales, valence ranged from unpleasant

to pleasant, arousal from low to high activation, and dominance from low to high control. Following the circumplex model of affect, the combination of valence and arousal distinguished anxious, excited, dejected, and calm states [71].

5.2.2 Days 1–14: Everyday Companion Use. Participants completed multiple daily touchpoints described in Section 4.2.3. Each interaction followed the same core procedure. Participants first narrated an event by voice or text and rated their valence, arousal, and dominance before seeing any system estimate. The estimate was generated concurrently but withheld until self-report was submitted to avoid anchoring. Participants then completed a brief guided conversation designed to take under five minutes and repeated the SAM ratings afterward. Three 7-point Likert items assessed helpfulness, feasibility, and willingness to act. Participants also reviewed the companion’s emotion interpretation and provided a correction when their understanding rating was 4 or below. These corrections informed subsequent analysis and cross-session memory. Even-numbered evening sessions included a rotating probe on guidance adoption, perceived personalization, or experiences with multimodal features. The mobile-first interface supported on-the-go use, and sessions inactive for 30 minutes were automatically finalized as timed out.

5.2.3 Day 14+: Post-Study Session. The post-study session was conducted by a researcher over Zoom and recorded with participant consent. Participants first completed a 12-item questionnaire using 7-point Likert scales to assess perceived emotional understanding, guidance quality, and technology acceptance. The questionnaire combined study-specific items with items adapted from the Perceived Empathy of Technology Scale [75], Session Rating Scale [21], the cognitive-reappraisal construct [25], and Technology Acceptance Model [19]. The full questionnaire is provided in Appendix D. A 20–25-minute semi-structured interview then explored participants’ overall experience, perceived accuracy of the system’s emotion estimates, suggestion adoption, changes in emotional awareness, and perceptions of the multimodal features (Appendix E).

5.3 Data Collection and Analysis

We adopted a mixed-methods approach combining paired system estimates and self-reports with interaction logs, session feedback, post-study questionnaires, and interviews. Quantitative analyses examined emotion-estimation accuracy, within-session affective changes, and patterns of engagement and perceived value over the deployment. Interview accounts and open-ended responses helped us examine how participants experienced the companion’s understanding, guidance, and continued presence in everyday life. Appendix G summarizes the data collected.

5.3.1 RQ1: Emotion Inference and Perceived Understanding. We evaluated emotion-estimation accuracy against participants’ paired SAM self-reports, comparing initial estimates with pre-guidance ratings and post-guidance estimates with post-guidance ratings. Valence, arousal, and dominance were assessed using MAE, Pearson’s r , and mean signed error. We also measured lower-versus-higher state agreement using Cohen’s κ , with additional thresholds examined for arousal. For initial valence, we compared the deployed appraisal scaffold with two baselines: Claude Opus 4.8, the same model without appraisal reasoning, and a lexicon-based estimator using the NRC-VAD Lexicon [59]. Lexicon scores were mapped to the 1–9 SAM scale using five-fold held-out estimation, and all three estimators were evaluated against the same pre-guidance self-reports. We additionally report the NRC-VAD arousal correlation descriptively using the same held-out mapping procedure. We also summarized perceived understanding using session-level and post-study ratings. Spearman correlations tested whether perceived understanding was associated with absolute estimation error in valence, arousal, and dominance.

5.3.2 RQ2: Regulation Outcomes and Guidance Fit. We measured immediate regulation as within-session change in valence, arousal, and dominance. Marginal changes were tested using exact two-sided Wilcoxon signed-rank tests on participant-level mean changes, with Holm correction across the three dimensions. We also examined joint valence–arousal trajectories across four entry states defined by a midpoint split at 5: anxious or agitated, excited or energized, dejected or flat, and calm or content. Valence change was additionally summarized for sessions starting at low ($V \leq 4$) and high ($V \geq 7$) valence. Appendix H provides the marginal tests and visualization details.

To examine how perceived understanding and conversational alignment related to regulation outcomes, we tested the association between understanding ratings and within-session valence change using Pearson correlation and summarized mean ΔV by understanding rating. We also analyzed companion-turn stage labels, including starting stages, adjacent-stage transitions, and stage dwell time, to characterize how contextual exploration and emotional acknowledgment preceded guidance. Two researchers independently coded corrections submitted after understanding ratings of 4 or below and resolved disagreements through discussion. Codes captured emotion misreads, context or memory errors, strategy mismatches, and inability to judge the interaction. We compared each participant’s mean valence change between corrected and uncorrected sessions using an exact two-sided paired Wilcoxon signed-rank test among the 12 participants contributing both session types, and descriptively compared their starting valence.

We compared participant-level mean valence changes across the four regulation-strategy families using a Friedman test. Differences in helpfulness, feasibility, and willingness to act were tested with mixed-effects likelihood-ratio tests. Model specifications and multiple-testing correction appear in Appendix I. Session feedback, post-study ratings, and interview accounts were used to characterize how guidance matched participants’ needs and practical constraints. Guidance adoption was assessed through evening probes asking whether participants had tried earlier suggestions. We classified 44 responses with LLM assistance [24] and manually reviewed each result. Categories captured successful adoption, partial or alternative enactment, already-intended actions, non-adoption, and inability to recall. We report category counts to describe uptake beyond the conversation.

5.3.3 RQ3: Continued Use and Personalization. To examine how engagement and perceived value changed over two weeks, we analyzed daily session frequency, turns per session, narration length, user input length per turn, and ratings of helpfulness, feasibility, and willingness to act. Mixed-effects models estimated study-day trends and Week 2 versus Week 1 contrasts (Appendix J). Interview accounts contextualized these trends through changes in participation, expectations, and familiarity with the companion. Perceived personalization was examined through post-study ratings of personalized responses and feeling increasingly known, alongside interview accounts of cross-session memory. We examined how participants described recall of earlier interactions, updates to changing circumstances, and whether corrections carried into later conversations. Post-study ease-of-use ratings supplemented this analysis.

5.3.4 Qualitative Analysis. Two researchers conducted thematic analysis of the post-study interview transcripts and other rotating-question responses [7]. They coded independently, reconciled differences, and iteratively refined the codebook. Themes were organized around research questions, with sub-themes emerging during discussion.

6 RESULTS

We first describe how participants engaged with PAIR across the fourteen-day deployment (Section 6.1). We then examine emotion inference, momentary regulation, and cross-session personalization. RQ1 evaluates the accuracy of emotional states inferred from everyday narration. RQ2 examines emotional changes following guided conversations

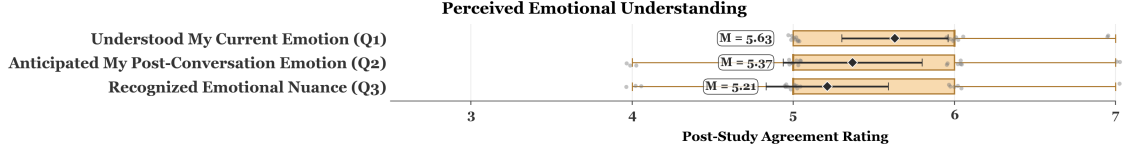


Fig. 6. Exit ratings of perceived emotional understanding. In-the-moment understanding item had the highest mean rating, while emotional nuance recognition had the lowest. Boxes show the distribution of participant ratings, grey points are individual participants, and diamonds with error bars show means with 95% confidence intervals. Items were rated on a 7-point scale (1 = strongly disagree, 7 = strongly agree).

Table 3. Estimation accuracy against unanchored self-report on the 9-point SAM scale. Initial estimates are evaluated against before-guidance ratings and post-guidance estimates against after-guidance ratings. The bottom three rows compare initial valence estimators across all narrations, with Claude Opus 4.8 used for both LLM conditions. Signed error is computed as estimate minus self-report, so positive values indicate overestimation. All correlations are significant at $p < .001$ (**).

	Dimension / Estimator	MAE	r	Signed err.
Initial	Valence	1.196	.669***	+0.213
	Dominance	1.299	.538***	−0.061
	Arousal	1.651	.236***	−0.129
Post-guidance	Valence	1.115	.614***	−0.128
	Dominance	1.189	.495***	−0.393
	Arousal	1.963	.175***	−1.097
Valence estimators	NRC-VAD lexicon, 5-fold CV	1.679	.318***	—
	LLM without scaffold	1.334	.640***	—
	PAIR appraisal scaffold	1.196	.669***	—

and subsequent guidance uptake. RQ3 examines how engagement and perceived personalization developed through continued use.

6.1 Overview of PAIR Deployment and Engagement

Participants completed 1,093 sessions and 10,953 conversation turns, averaging 57.5 sessions each ($SD = 11.8$, range 35–78), or 4.1 sessions per day. Just-in-time logs accounted for 55.0% of sessions, morning check-ins 23.6%, and evening reflections 21.4%. Guided conversations averaged 10.02 turns ($SD = 2.80$), with a median session duration under three minutes. Participants used text in 872 sessions (79.8%) and voice in 221 (20.2%), with five participants using voice in more than 40% of their sessions. Overall, participants rated the system as easy to use, with an exit rating of $M = 5.37$ ($SD = 1.54$) on the 7-point scale (Figure 10). Across sessions, self-reported pre-guidance valence was low (≤ 4) in 28.7%, intermediate (5–6) in 33.0%, and high (≥ 7) in 38.2%. The router assigned 87.5% of sessions to Cog and 12.5% to Combined.

6.2 RQ1: Text-Based Appraisal Inference Supports Valence and Dominance but Remains Limited for Arousal

To address RQ1, we compared the system’s estimates with participants’ unanchored self-reports of the same moments. We assessed fine-grained accuracy across all paired observations (Table 3), compared the appraisal scaffold with direct LLM inference and a lexicon baseline, and tested agreement after reducing estimates to lower-versus-higher states.

6.2.1 Appraisal-Guided Inference from Daily Narration Yielded Strong Valence and Moderate Dominance Correlations with Self-Reports. Appraisal-guided inference from everyday narration showed a strong correlation with self-reported valence and a moderate correlation with self-reported dominance. Initial estimates correlated with self-reports at $r = .669$ for valence and $r = .538$ for dominance, with MAE of 1.196 and 1.299 points, respectively (Table 3). Both dimensions showed lower estimation error than arousal, and valence remained the most accurately estimated dimension after guidance. Across all narrations, appraisal-guided inference achieved a valence correlation of $r = .669$ and an MAE of 1.196, compared with $r = .640$ and MAE = 1.334 for the same-model baseline without the appraisal scaffold and $r = .318$ and MAE = 1.679 for the cross-validated NRC-VAD baseline. Absolute valence error showed no statistically reliable trend across the 14-day deployment ($b = -0.0069$ on the log1p scale, $p_{\text{FDR}} = .087$).

Participants also rated the companion highly for understanding how they felt in the moment ($M = 5.63$, $SD = 0.68$), anticipating how they would feel after conversations ($M = 5.37$, $SD = 0.90$), and recognizing emotional nuance ($M = 5.21$, $SD = 0.79$) on 7-point scales (Figure 6). Their interview accounts showed that this understanding was most salient when PAIR inferred emotion from events containing little explicit emotion language. Fourteen of nineteen participants described cases in which PAIR inferred feelings or their causes without explicit emotion language. P1 said the system could “go straight from my description to dissecting what was going on inside me.” P10 similarly explained that even when she did not have words for how she felt, the system could identify both the emotion and “what was responsible” for it. Others described the system as helping “peel back those layers” (P3), while P19 valued its recognition that “saying goodbye can still make you feel sad” within an otherwise positive experience. Participants thus valued inference when it moved from an event description to an emotional interpretation, especially when it supplied a connection they had not already articulated. By contrast, participants identified shallow restatement and miscalibrated emotional intensity as two recurring limitations in PAIR’s emotion inference. Intensity emerged as the more common boundary. In some cases, the system added little beyond explicit cues. P18 reported needing to state “I’m very sad” before receiving a less flat interpretation, while P5 said presenting an obvious restatement as deeper insight “felt like it was provoking me.” More often, the system identified the emotional direction but misjudged its magnitude. This distinction between emotional direction and intensity becomes especially visible in arousal inference, which we examine in Section 6.2.2.

6.2.2 Text-Only Inference Yielded Weak Arousal Correlations and Recurrent Misjudgments of Emotional Intensity. Arousal estimates showed lower agreement with participants’ self-reports across both fine-grained and binary evaluations, with mismatches concentrated in emotional intensity when expressive cues were unavailable in text. Although participants used voice in 221 sessions, PAIR transcribed speech before inference and did not retain prosodic cues. Its arousal estimates therefore reflected what participants said in the transcript without information from how they said it. Arousal estimates correlated weakly with self-report ($r = .236$; Table 3). Binary estimates showed moderate agreement for valence ($\kappa = .539$) and dominance ($\kappa = .430$), but only slight agreement for arousal ($\kappa = .154$), and changing the cut point did not raise arousal agreement above $\kappa = .21$. The word-based baseline showed little correspondence with self-reported arousal ($r = .016$), compared with a higher correlation for valence ($r = .318$).

Interview accounts identified recurrent amplification or flattening of emotional intensity. Eight of nineteen participants described cases in which PAIR identified the general emotional direction but misjudged how strongly they felt it. P12 said the system “definitely exaggerate[d] things that I wasn’t feeling” and described her emotions as being taken “way too extreme”. Others distinguished wanting to express frustration from actually feeling intense distress, such as “I’m just trying to vent. I’m not actually mad” (P3) and “Sometimes you’re not really that upset over it, you just want to talk about it” (P7). Participants also linked these mismatches to missing expressive cues, noting that without face-to-face information

it was harder to judge “*how upset you are*” (P7). These errors also changed how some participants communicated with the system. P12 softened her wording to avoid amplification, while P18 made emotional intensity more explicit when the initial interpretation felt too flat. The interview accounts thus located a recurring difficulty in judging emotional intensity from text. Because estimated intensity informed regulation-strategy selection, these reported mismatches were also relevant to the guidance examined in RQ2.

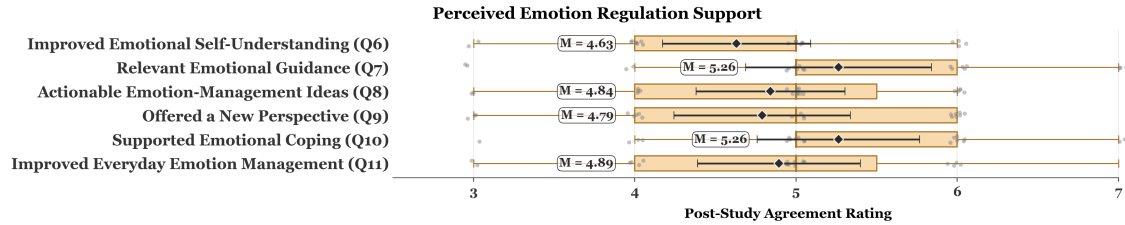


Fig. 7. Exit ratings of perceived emotion-regulation support. Relevant emotional guidance and support for emotional coping received the highest mean ratings (both $M = 5.26$), while improved emotional self-understanding received the lowest ($M = 4.63$). Boxes show participant rating distributions, grey points show individual participants, and diamonds with error bars show means with 95% confidence intervals. Items were rated on a 7-point scale (1 = strongly disagree, 7 = strongly agree).

6.3 RQ2: Initial Emotional States, Conversational Responsiveness, and Practical Feasibility Shaped Momentary Regulation and Guidance Uptake

Guided conversations supported emotion regulation differently across participants’ starting states and immediate needs. To address RQ2, we examine affective change by entry state, how exploration and emotional acknowledgment shaped perceived understanding, differences across regulation strategies, and when guidance was valued and adopted. We integrate pre- and post-guidance SAM ratings with strategy logs, session feedback, exit ratings, adoption probes, and interviews.

6.3.1 State-Dependent Emotional Adjustment Following Guided Conversations Suggested Momentary Regulation. Guided conversations were followed by state-dependent emotional adjustment consistent with momentary regulation, with the largest valence gains in anxious and dejected sessions and arousal shifting according to the starting state. Figure 8 visualizes this pattern. Trajectories originating in the anxious quadrant generally moved toward higher valence and lower arousal, while those originating in the dejected quadrant moved toward higher valence and higher arousal. These shifts indicate different regulatory directions from different entry states, reducing activation when participants began anxious and increasing activation when they began dejected. Aggregated within each entry-state quadrant, anxious sessions gained 1.875 points in valence ($SD = 1.453$) while decreasing 0.547 points in arousal ($SD = 1.522$). Dejected sessions gained 1.208 points in valence ($SD = 1.427$) and 0.780 points in arousal ($SD = 1.220$). Valence gains were smaller in the two positive quadrants, at +0.162 in excited sessions and +0.304 in calm sessions, while arousal changed by +0.069 and +0.906, respectively. The entry-valence split showed the same pattern. Sessions beginning at low valence (≤ 4) gained 1.344 points ($SD = 1.455$), while those beginning at high valence (≥ 7) changed by -0.108 points ($SD = 0.857$). Across all sessions, valence increased by 0.525, dominance by 0.469, and arousal by 0.334, with Wilcoxon tests at $p < .001$ (Appendix H).

Participants described the clearest benefit when conversations began from a state they wanted to change, while positive states were more often maintained. Exit ratings also reflected perceived regulation support, with participants

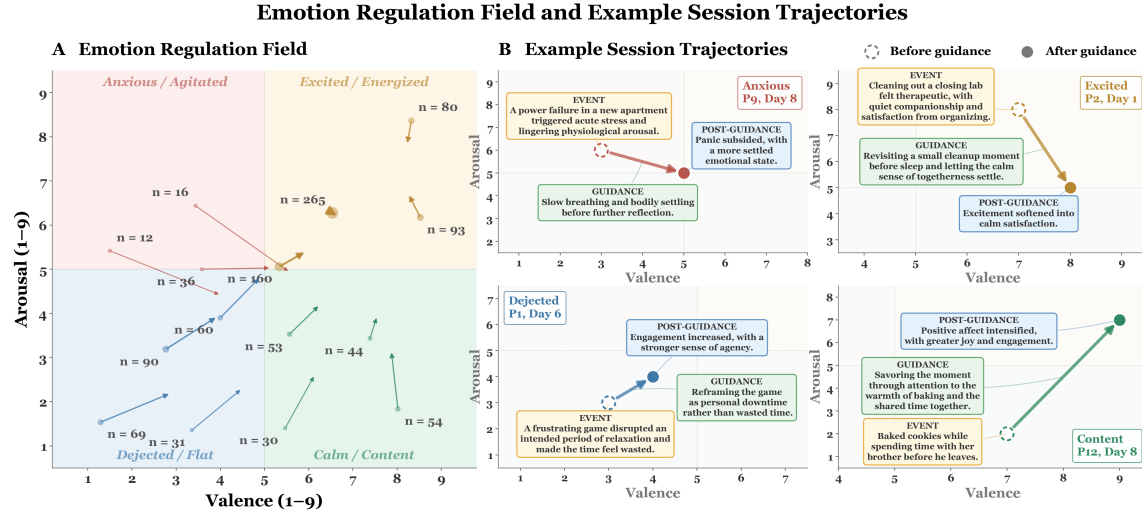


Fig. 8. Guided conversations were followed by state-dependent affective movement. (A) Arrows connect clustered mean self-reported valence and arousal before guidance to the corresponding means after guidance; labels show the number of sessions in each cluster. Anxious and dejected trajectories generally moved toward higher valence, with arousal decreasing for anxious states and increasing for dejected states. Calm trajectories also became more activated, while excited trajectories changed least. (B) Example session from each starting-state quadrant illustrates these directions by connecting the narrated event and guidance with the self-reported state after guidance.

reporting help in coping with their emotions ($M = 5.26$, $SD = 1.05$) and managing emotions in everyday life ($M = 4.89$, $SD = 1.05$; Figure 7). Participants also reported support for understanding their own emotions ($M = 4.63$, $SD = 0.96$). P10 entered feeling “so drained” and left much better, while P18 said the companion helped her reconsider a “really emotional” situation and decide she did not need to be “this upset.” P16 similarly found the conversation most helpful when she was “exhausted” or needed someone to talk to in the moment. Positive or already-settled states changed less. P9 said such conversations helped “feel more positive afterward,” while P12 “rarely ever mentioned a problem because I just didn’t have any” and P11 described her state as already normal and clear. These accounts complement the larger gains observed at low valence by distinguishing relief from distress from the maintenance of an already positive state.

6.3.2 Contextual Exploration and Emotional Acknowledgment Before Advice Supported High Perceived Understanding. Participants described feeling understood through contextual exploration and emotional acknowledgment before advice. Per-session understanding averaged 5.45 of 7, with 96.0% of ratings at 5 or above, while correlations with absolute estimation error were small ($\rho = -.014$ for valence, .029 for arousal, and .101 for dominance). Because participants never saw the numerical estimates, their judgments reflected how understanding was conveyed through the interaction. The following findings describe how participants experienced understanding and how it related to emotional benefit. First, participants experienced understanding through an ECA sequence that elicited context before offering guidance. Fourteen participants described clarification and follow-up questions as helping the system reach an accurate interpretation. P3 said that “after going back and forth three, four times,” the system had a good grasp of the situation. Logged ECA stages showed the same progression, with 99.9% of sessions beginning in Exploration, an expected dwell time of 1.8 companion turns, and Comforting transitioning to Action in 72.2% of adjacent stage pairs, compared with 16.1% for direct Exploration-to-Action transitions. This sequence created space to clarify the event

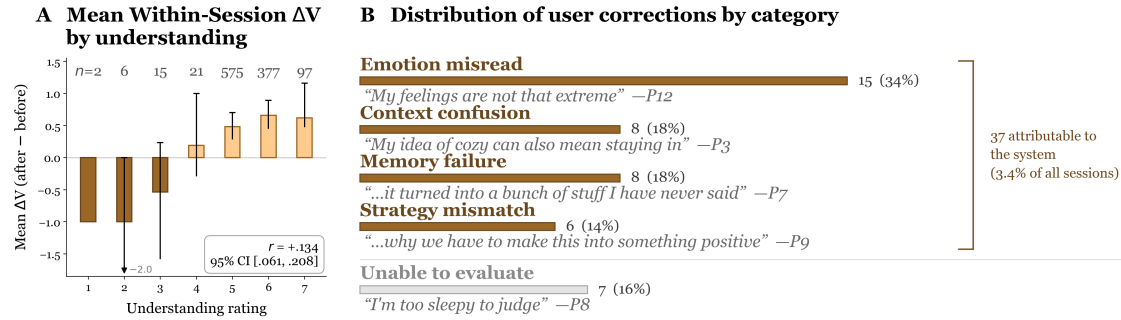


Fig. 9. Perceived understanding was associated with within-session valence change, and participant corrections identified recurring system mismatches. (A) Mean valence change (after minus before) by understanding rating. Bars show means, error bars show 95% confidence intervals, and labels show session counts; higher ratings corresponded to greater valence gains ($r = .134$, 95% CI [.061, .208]). (B) The 44 correction texts by coded category, each paired with a representative correction. The four system-attributable categories contained 37 corrections, representing 3.4% of all sessions; the remaining seven indicated that the participant was unable to evaluate the system.

and respond to emotion before proposing action. When questions followed the wrong interpretation, however, the same structure could feel misaligned, as P8 described receiving questions that were not on “*the right track*.” Second, perceived understanding remained consistently high but did not increase reliably over time. A linear mixed-effects model estimated only a small day slope ($b = +0.0121$, $p_{FDR} = .087$), equivalent to approximately +0.16 points from Day 1 to Day 14. Eighteen participants described the system as generally accurate, including P10, who felt understood “*right from day one*.” Participants also linked this accuracy to disclosure. P2 evaluated the system relative to “*how much information I gave it*,” while P5 sometimes needed to describe a situation in substantial detail before the system understood it. Stable understanding therefore emerged from an iterative exchange in which the system elicited relevant detail and participants supplied enough context for it to interpret.

Third, participants’ corrections showed that perceived understanding extended beyond emotion labels to broader interactional alignment. Only 4.0% of sessions included correction text, all following understanding ratings of 4 or below. Among 37 attributable corrections, 15 concerned emotion direction, intensity, or tone, 8 concerned the situation, 8 involved memory, and 6 concerned the desired support strategy (Figure 9B). P1 never felt a need to correct the system “*on the part where it defines my emotions*,” whereas P2’s corrections targeted “*the wrong suggestion*.” Participants’ sense of being understood therefore depended not only on emotion recognition, but also on contextual fit and appropriate support. Fourth, perceived understanding was positively associated with within-session valence gain ($r = .134$, 95% participant-bootstrap CI [.061, .208]). Mean valence change was negative at understanding ratings of 1–3 and positive at ratings of 5–7 (+0.48 to +0.66; Figure 9A). Sessions with correction text also showed lower valence change than uncorrected sessions (−0.27 vs. +0.56; paired Wilcoxon $W = 5$, $p = .005$, $n = 12$ participants), despite similar starting valence. Perceived misunderstanding was therefore associated with reduced emotional benefit.

6.3.3 Contextual Strategy Selection Produced Distinct Regulatory Actions Without Detectable Differences in Immediate Outcomes. Contextual strategy selection distributed guidance across the four Gross-informed strategy families, with no detectable differences in immediate valence gains (Friedman $\chi^2(3) = 0.613$, permutation $p = .897$) or subjective evaluations across families. Mean within-session valence gain was highest for situation modification (+0.672, $n = 317$), followed by cognitive change (+0.585, $n = 246$), response modulation (+0.453, $n = 256$), and attentional deployment

(+0.369, $n = 274$). Immediate evaluations were also similar across the four families, with mean helpfulness, feasibility, and willingness to act ratings ranging from 5.00 to 5.28 on the 7-point scales. Mixed-effects likelihood-ratio tests detected no significant differences by selected strategy (all $p > .29$, all $p_{\text{FDR}} = .395$; Table 7). No strategy family showed a clear advantage in immediate outcomes under context-dependent selection. Interviews confirmed that these strategy selections reached participants as different regulatory moves. Participants rated the companion's provision of actionable emotion-management ideas at $M = 4.84$ ($SD = 0.96$) and new perspectives at $M = 4.79$ ($SD = 1.13$; Figure 7). For situation modification, P1 recalled being encouraged to start with the easiest part of a stalled task. Cognitive change used participants' own facts to revise an unfavorable interpretation. P19 described the companion countering self-blame by recalling what had already been completed that day, saying, "*You already accomplished quite a lot.*" P12 adopted attentional suggestions such as getting a snack or taking a break. For response modulation, P16 drank warm water and breathed deeply before an exam, reporting that she "*actually tried that, and it worked.*" The range of strategies did not ensure that every suggestion suited its context. Six participants described a fixed pool of suggestions, including P1, whose solitary leisure suggestion was applied to a relationship conflict where it "*would only make things worse.*" These accounts show how the four families translated into concrete regulatory actions whose usefulness depended on their fit to the immediate situation.

6.3.4 Practical Feasibility and Alignment with Immediate Support Needs Encouraged Greater Guidance Acceptance and Uptake. Participants carried guidance into subsequent action when it identified a concrete step that fit their immediate resources and constraints. Exit ratings of relevance averaged 5.26 ($SD = 1.19$; Figure 7), while immediate feasibility and willingness-to-act ratings averaged 5.18 ($SD = 1.15$) and 5.13 ($SD = 1.23$), respectively (Table 8). Evening probes documented selective uptake beyond the conversation. Among 44 categorized and manually checked responses, 24 described trying the guidance with a positive result and five described partial or alternative enactment, or mixed results. Four reported actions that coincided with what participants already intended to do, eight reported not trying the suggestion, and three could not recall it. Reported benefits included relief from a headache after taking a screen break and feeling better prepared after choosing an interview outfit.

Participants acted on suggestions that fit their available time and resources, while distinguishing a reasonable proposal from one they wanted to follow. P6, P16, and P1 described adopting suggestions they judged feasible or reasonable. P7 explained that "*You can always take one minute to the side,*" while a walk could require time she did not have. The same practical criterion explained rejection. P3 could not obtain a suggested matcha at 5 a.m., and P12 would not interrupt a conversation with a friend to follow the advice. P18 articulated the distinction between willingness and reasonableness, saying, "*Sometimes I did not really want to follow the suggestion, but I still thought the suggestion itself was reasonable.*" Meanwhile, support was also judged by whether the moment called for reassurance, encouragement, or further listening. P2 sometimes wanted to hear that "*it'll be okay,*" and reassurance was sufficient for that need. P18 instead described moments when she wanted the companion to "*push me a little*" and help motivate action. Advice itself could be unwelcome when participants still wanted to talk. P7 explained that "*sometimes you just want something to listen to you*" and that solutions could feel overwhelming. P11 likewise preferred the earlier conversation to the later advice. Useful guidance therefore required both a response suited to the current problem and recognition of the support the participant was ready to receive.

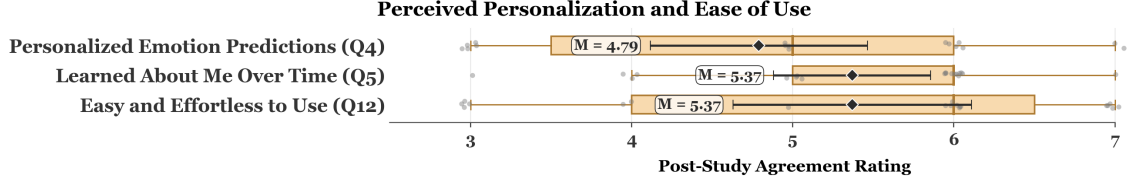


Fig. 10. Exit ratings of perceived personalization and ease of use over the 14-day deployment. Feeling increasingly known and ease of use received higher mean ratings than personalized emotion estimates. Items were rated on a 7-point scale (1 = strongly disagree, 7 = strongly agree).

6.4 RQ3: Conversational Familiarity Supported Engagement, While Cross-Session Memory Contributed to Personalization

To address RQ3, we examine changes in engagement and perceived helpfulness alongside participants’ adjustments to their narration practices and support expectations. We then examine how remembered context, changing circumstances, and the persistence of corrections shaped perceived personalization. We combine longitudinal session logs and feedback with exit ratings and interviews across the 14-day deployment.

6.4.1 Familiarity with Response Patterns Guided Narration Practices and Support Expectations. Across two weeks, participants continued to use PAIR at a similar frequency while writing less and rating its immediate helpfulness more highly. Daily session frequency showed no reliable trend ($p_{FDR} = .963$), and the small decline in conversation turns did not remain significant after correction ($p_{FDR} = .091$). Log-transformed mixed-effects models estimated daily decreases of 3.37% in initial narration length and 4.30% in user input length per turn, both measured in characters (both $p_{FDR} < .001$). Helpfulness ratings increased over the two weeks ($b = +0.029$ points per day, $p_{FDR} = .005$). Unadjusted means rose from 4.94 in Week 1 to 5.14 in Week 2 ($b = +0.216$ for the week contrast, $p_{FDR} = .011$). Feasibility and willingness to act showed no significant day trends after FDR correction ($p_{FDR} = .182$ and $.214$, respectively). Table 8 reports both models for all nine session measures.

Interviews revealed how participants adjusted the content, timing, and effort of their disclosure over time. Nine of nineteen participants described adapting their wording or conversational approach to how the companion responded. P5 spent increasing time composing initial narrations to “*use as few words as possible*” while retaining relevant context. P15 similarly learned “*how it understood messages over time*”, while P18 supplied additional detail and learned to “*explicitly describe my current emotion*” to improve understanding. P8 deliberately shortened openings to invite follow-up questions, explaining that “*I would know to wait*”, and left elaboration for subsequent turns. P14 described how the companion “*finds a way to expand it*” after a short answer and follows existing details after a longer one. After receiving repetitive advice, P16 moved from detailed narration to sharing less, saying, “*as time passed, I didn’t share that much.*” These accounts distinguished deliberate adjustments to expression from reduced willingness to elaborate after repetitive advice.

Participants’ expectations and evaluations of the companion also evolved over the two weeks. Three participants connected their evaluations to changing goals, familiarity, or their own interaction skills. P5 shifted from seeking empathy toward “*practical suggestions*” and reported that approaching conversations with the goal of “*looking for an answer or solution*” tended to leave him feeling better afterward. When explaining an emotional-understanding rating, P8 described “*almost identical*” responses becoming more acceptable through “*a basic kind of natural acceptance towards the AI*”. P12 could not distinguish whether the model had improved or “*I interacted with it better*”. Participants thus

evaluated later interactions through changing support goals, familiarity with the responses, and their own experience of communicating with the system.

6.4.2 Relevant Recall Supported Personalization, While Outdated Context and Forgotten Corrections Disrupted Continuity. Participants judged personalization by whether the companion retained relevant personal context and used it appropriately in later interactions. Post-study ratings for the companion’s ability to develop an understanding of the participant over time averaged 5.37 ($SD = 1.01$), while ratings of personalized emotion estimates averaged 4.79 ($SD = 1.40$; Figure 10). Interviews reflected this variation. Ten of nineteen participants described personalized responses, while fourteen also encountered generic or scripted ones. Participants distinguished between simply remembering personal information and using it meaningfully in later responses. P14 treated recall of hobbies, a pet, and work preferences as evidence that *“It has a memory.”* P1 acknowledged that the companion retained prior statements but described the resulting response as *“a fairly linear answer,”* expecting it to connect remembered information to personal tendencies and likely reactions. Personalization therefore depended on how remembered context was interpreted and applied, not only on whether it was retained.

Continuity became most visible when the companion connected earlier experiences to later developments. P5 began feeling better understood *“around halfway through the study,”* while P14 was surprised when the companion recalled *“a small thing I’d mentioned only once”* one or two weeks earlier. P19 described a stronger form of continuity. After discussing a routine of going straight home and feeling inactive after work, P19 began visiting a café instead. The companion later recalled this change and recognized it as *“a very good change.”* These cases demonstrated that remembered context supported personalization when it carried specific experiences forward and incorporated subsequent changes.

Personalization weakened when remembered context was missing, temporally misplaced, or treated as fixed. Twelve of nineteen participants reported instances in which the companion forgot prior context, and six described confusion about when events occurred or how they were related. P7 found that a later check-in no longer retained the context of a presentation discussed earlier that day. P10 similarly had to reconstruct information provided that morning, saying, *“I have to list it all over again.”* P2 described an event from three days earlier being treated as current even after correction. Context could also persist too rigidly. P15 felt the companion *“relied too heavily on what I said in the onboarding session,”* turning occasional activities into recurring assumptions despite *“I don’t do them like every day.”* These failures required participants to re-establish context or correct an outdated interpretation. Moreover, participants also judged the incorporation of corrections by whether later responses reflected their feedback. Six participants reported that feedback was incorporated, while nine described corrections that did not persist or eventually stopped correcting the system. P12 noticed that feedback about exaggerated emotional intensity led to more frequent use of *“calmer”* language. P18 also observed improvement, but only after having to *“correct it two or three times.”* P4 described the opposite pattern, where the companion responded appropriately to a clarification in the moment but failed to retain it later. These experiences made the persistence of revised interpretations across encounters a distinct criterion for perceived personalization.

7 DISCUSSION

7.1 Summary of Results

Our fourteen-day deployment connects momentary emotion understanding and regulation with the cross-session demands of sustained support. For **RQ1 (Emotion Understanding)**, appraisal-guided estimates aligned more closely with self-reported valence and dominance than arousal. The scaffold also improved valence accuracy over both baselines. For **RQ2 (Regulation Guidance)**, guided conversations were followed by higher valence, lower arousal in anxious

sessions, and higher arousal in dejected sessions. Participants felt understood through contextual exploration and emotional acknowledgment, and adopted guidance suited to their immediate needs and practical constraints. No strategy family showed a clear advantage in immediate outcomes under contextual selection. For **RQ3 (Cross-Session Continuity)**, inputs shortened and helpfulness ratings increased, while session frequency showed no reliable trend. Participants adjusted their narration and support expectations over time. Relevant recall and context updates reinforced perceived personalization, while outdated assumptions and forgotten corrections required users to repeat information or feedback. These findings motivate systems that adapt guidance to current needs and prior attempts while maintaining inspectable, revisable memory across changing circumstances.

7.2 Study Limitations and Open Challenges for Sustained Emotional Support

7.2.1 Study Limitations. Our study has several limitations concerning the scope of the deployment and the strength of the resulting evidence. The fourteen-day study included 19 adults aged 18–33 who were generally familiar with AI, limiting the populations represented. The fixed check-in schedule also constrained naturalistic use and does not reflect how often the system would be used voluntarily. Beyond deployment scope, affect ratings rely on participants’ own use of the scale, and concentrated understanding ratings reduce sensitivity to variation across sessions and participants. Although initial ratings were collected before system feedback, post-guidance reports may reflect how the conversation framed the experience [45]. Comparisons across regulation strategies are similarly observational, since strategy selection depended on the participant’s state. Controlled comparisons with unguided journaling and longer follow-up could strengthen these inferences and establish whether momentary effects carried into everyday well-being.

7.2.2 Open Challenges for Sustained Emotional Support. Our findings identify reliable interpretation of emotion and interactional context, along with memory-based adaptation, as two open challenges for sustained emotional support. Text-based estimates showed lower agreement for arousal than for valence and dominance, while interpreting subtle or recurring emotions also requires context beyond the immediate event. Access to this context depends on what users choose to disclose. Privacy and judgment concerns can restrict disclosure [13], while the companion’s questions and interpretations can shape how users describe subsequent experiences. This challenge extends beyond affect estimation to relational judgment. Models can generate fluent emotional context in supportive exchanges while still misjudging whether it fits the emotional moment [70]. Reliable support therefore requires interpreting evidence shaped by both selective disclosure and the interaction itself. Multimodal cues such as speech prosody could supplement text-based inference, while brief confirmation could let users refine interpretations before they inform guidance. These mechanisms should preserve users’ control over disclosure without making increasingly detailed accounts a prerequisite for support. The second challenge is translating memory into guidance that responds to prior attempts and changing circumstances. Participants encountered suggestions they had already considered and corrections that did not persist across encounters. These patterns show that recalling personal information does not ensure that subsequent support adapts to it. Maintaining and revising temporal context remains difficult for long-context systems [91]. The challenge is determining which prior experiences remain relevant and how they should change the next recommendation.

7.3 Toward Sustained Emotional Support

The challenges identified above highlight the capabilities required to sustain emotional support as users’ goals and circumstances evolve. Momentary support addresses immediate emotional needs through acknowledgment and contextual guidance. Sustained support additionally requires tracking changes in support preferences, incorporating the outcomes

DESIGN FOCUS	DESIGN IMPLICATION	DESIGN RATIONALE	SYSTEM REQUIREMENT
Evolving Support Goals	 DI1 Support goal alignment	Support goals change across encounters.	Align listening, advice, and session completion with current goals.
Guidance Continuity	 DI2 Adaptive guidance from longitudinal feedback	Failed suggestions recur; initial validation invites later doubts.	Adapt guidance to prior attempts, outcomes, and constraints.
Inspectable and Revisable Memory	 DI3 Memory transparency and correction	Assumptions persist while corrections are forgotten.	Expose memory sources and preserve corrections across sessions.
	 DI4 Retention and contextual reuse control	Personal disclosures remain sensitive across contexts.	Keep stored details, retention periods, and reuse under user control.

Fig. 11. Four design implications for sustained emotional support, organized around evolving support goals, guidance continuity, and user-controlled memory. Each implication connects its rationale to a system requirement.

of prior guidance, and updating personal context across interactions. Support preferences can shift between emotional expression and practical assistance, while recurring concerns require reassessing guidance in light of prior attempts and their outcomes. Cross-session personalization also raises questions of informational control, including how users inspect and correct stored information and govern its retention and reuse. Building on our findings, we derive several design implications for 1) aligning support with evolving goals, 2) adapting guidance through longitudinal feedback, and 3) supporting personalization and privacy through user-controlled memory (Figure 11).

7.3.1 Changing Support Goals Require Flexible Transitions Between Listening and Guidance. Sustained emotional support requires alignment between users’ current goals and the form and timing of assistance. Our findings indicate that these goals vary across encounters and can change with continued use. P5 shifted from seeking empathy toward “practical suggestions,” whereas P7 explained that “sometimes you just want something to listen to you” and found solutions overwhelming. These accounts extend research on how users define chatbot support [84] by highlighting both changing preferences over time and the purpose of individual encounters. In PAIR, emotional intensity and controllability inform strategy selection, while dialogue progresses toward action within a turn budget. A strategy suited to the inferred emotion may still be premature when the user wants to continue expressing an experience. The Exploration–Comforting–Action framework permits flexible transitions [50], requiring systems to determine whether advice is appropriate at a given moment [69]. Yet LLMs can favor particular support strategies in ways that impair appropriate selection [42]. These findings motivate using the current support goal alongside emotional state to determine when to continue listening or introduce guidance.

Design Implication 1: Support Goal Alignment through Intent Elicitation and Dialogue Adaptation. Systems should identify current support goals from users’ spontaneous requests or lightweight choices between continued expression and practical assistance. When intent remains unclear, targeted clarification questions can elicit conversational preferences [48]. These goals should remain revisable, with subsequent requests taking precedence over prior preferences. Before introducing advice, the system should confirm readiness if users have not already expressed it. Guidance should remain available on demand [83], while declining advice should return the conversation to listening. Users should also be able to conclude the interaction once their needs are met. These controls align dialogue transitions and session completion with users’ current goals across repeated interactions.

7.3.2 Prior Attempts and Outcomes Should Inform the Next Step in Regulation Guidance. Useful regulation guidance requires continuity between prior attempts, their outcomes, and subsequent recommendations. P18 reported receiving suggestions she had “*already tried*” that “*had not worked*,” while P3 valued familiar ideas made more salient through conversation. These accounts suggest that usefulness depends on how guidance builds on prior experience, including when to repeat, modify, or replace a suggestion. Regulatory flexibility theory similarly emphasizes responsiveness to feedback when maintaining or changing a strategy [5]. Research on LLM support for procrastination further highlights the need for structured steps and adaptation to users’ practical constraints [3]. Evaluating this adaptation also requires distinguishing immediate reassurance from subsequent usefulness. P8 described responses as “*validating what I really wanted to hear*” but later questioned the behaviors they affirmed. These findings motivate preserving what users attempted and how they evaluated its consequences, so that later guidance can respond to both practical outcomes and retrospective judgments.

Design Implication 2: Adaptive Guidance through Longitudinal Feedback on Attempts and Outcomes. Systems should maintain a support history linking concerns and recommendations to attempted actions, reported outcomes, and barriers. Personalized guidance can draw on both historical behavioral data and contextual information elicited through dialogue [40]. Brief follow-up could distinguish helpful attempts, unsuccessful attempts, and suggestions not yet tried. Consistent with adaptive interventions that use changing states and contexts to inform support decisions [60], these records could guide subsequent recommendations. Time constraints might prompt a shorter action, while an unsuccessful attempt could motivate an alternative or an adjustment to its implementation. Evaluation should track immediate emotional response, later uptake and outcomes, and whether subsequent guidance incorporates this feedback.

7.3.3 Inspectable and Revisable Memory Should Give Users Control over Subsequent Personalization. Sustained personalization depends on retaining personal context that remains accurate and that users are willing to retain and reuse. P15 found that the companion “*relied too heavily on what I said in the onboarding session*,” treating occasional activities as recurring habits. P4 reported that a correction was understood immediately but “*later it would forget again*.” These accounts highlight the need to revise both the interpretation of remembered information and its continued application. Research on conversational memory similarly documents users’ concerns about retained information and their desire for greater visibility and control [9, 96]. Accuracy also intersects with privacy: users weigh the benefits of personalized responses against the risks of disclosing personal information [98]. Information shared to explain one experience may remain sensitive when recalled in another. Memory control therefore needs to address both whether the system’s understanding remains valid and whether retaining and reusing that information remains acceptable to the user.

Design Implication 3: Memory Transparency and User Control for Accurate Personalization. Systems should expose memories as data objects that users can inspect, edit, and selectively reuse across conversations [33]. Each entry should show its source conversation and relevant dates. Given that conversational memories can contain psychological inferences [18], this view should distinguish users’ explicit statements from system interpretations, alongside temporary states and recurring preferences. Users should be able to edit entries, mark events as completed, and remove inaccurate assumptions. Responses could link to the memories actually retrieved, making the basis for personalization traceable [96]. Changes should propagate to related summaries and retrieval records so that subsequent guidance uses the revised context. Evaluation should check whether later conversations incorporate corrections and stop repeating outdated assumptions.

Design Implication 4: Privacy Control over Memory Retention and Contextual Reuse. Systems should let users choose session-only use or selective retention, review the details to be stored, and set retention periods. Memory selection should follow the support purpose [37]; unnecessary identifying details could be removed or generalized through user-reviewed suggestions [20]. Retained information should also have adjustable conditions for reuse. For example, a user could restrict a workplace disclosure to discussions of that concern while permitting selected scheduling details to inform sleep-related guidance. Findings that LLMs struggle with contextual privacy judgments [58] motivate checking both relevance and user permissions before retrieving a memory. Broader use should require permission, and expiration or deletion should remove the retained content from associated summaries and retrieval stores.

7.4 Conclusion

Sustained emotional support requires generative agents to connect emotion understanding with guidance that adapts to users' changing needs and experiences. To evaluate these capabilities in everyday use, we deployed PAIR for fourteen days with 19 participants across 1,093 sessions. The companion integrated appraisal-guided emotion inference, regulation guidance, and cross-session memory. We paired system estimates with participants' self-reports and examined emotional changes alongside guidance uptake and experiences of continued use. Estimates corresponded more closely to self-reported valence and dominance than arousal. Perceived understanding showed little association with estimation error and instead emerged through contextual exploration and emotional acknowledgment. Guided conversations were followed by state-dependent emotional changes, while guidance uptake reflected immediate support needs and practical feasibility. Across repeated interactions, personalization relied on applying remembered experiences and user corrections as circumstances and expectations changed. These findings characterize sustained emotional support as a process in which agents continually interpret experiences, learn from prior interactions, and adjust their responses over time. This work offers a path toward affective intelligent agents that support everyday emotional life through effective momentary support and meaningful continuity over time.

REFERENCES

- [1] Adnan Abbas, Caleb Wohn, Arnav Jagtap, Eugenia H Rho, Young-Ho Kim, and Sang Won Lee. 2026. "Having Lunch Now": Understanding How Users Engage with a Proactive Agent for Daily Planning and Self-Reflection. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 1332, 23 pages. <https://doi.org/10.1145/3772318.3790957>
- [2] Anthropic. 2026. Introducing Claude Opus 4.8. <https://www.anthropic.com/news/claude-opus-4-8>. Published May 28, 2026. Accessed September 6, 2026.
- [3] Ananya Bhattacharjee, Yuchen Zeng, Sarah Yi Xu, Dana Kulzhabayeva, Minyi Ma, Rachel Kornfield, Syed Ishtiaque Ahmed, Alex Mariakakis, Mary P Czerwinski, Anastasia Kuzminykh, Michael Liut, and Joseph Jay Williams. 2024. Understanding the Role of Large Language Models in Personalizing and Scaffolding Strategies to Combat Academic Procrastination. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 15, 18 pages. <https://doi.org/10.1145/3613904.3642081>
- [4] Niall Bolger, Angelina Davis, and Eshkol Rafaeli. 2003. Diary methods: Capturing life as it is lived. *Annual Review of Psychology* 54, 1 (2003), 579–616. <https://doi.org/10.1146/annurev.psych.54.101601.145030>
- [5] George A. Bonanno and Charles L. Burton. 2013. Regulatory Flexibility: An Individual Differences Perspective on Coping and Emotion Regulation. *Perspectives on Psychological Science* 8, 6 (2013), 591–612. <https://doi.org/10.1177/1745691613504116>
- [6] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9)
- [7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- [8] Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-Grained Affective Processing Capabilities Emerging from Large Language Models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Piscataway, NJ, USA, 1–8. <https://doi.org/10.1109/ACII59096.2023.10388177>

- [9] Cheng Chen, Maria D. Molina, Mengqi Liao, and Eugene Cho Snyder. 2026. Relational Gains, Privacy Strains: Exploring Users' Perceptions and Experiences with ChatGPT's Memory Feature. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 912, 17 pages. <https://doi.org/10.1145/3772318.3791635>
- [10] Huili Chen, Yubin Kim, Kejia Patterson, Cynthia Breazeal, and Hae Won Park. 2025. Social Robots as Conversational Catalysts: Enhancing Long-Term Human–Human Interaction at Home. *Science Robotics* 10, 100 (2025), eadk3307. <https://doi.org/10.1126/scirobotics.adk3307>
- [11] Tiantian Chen, Jiaqi Lu, Ying Shen, and Lin Zhang. 2026. ES-MemEval: Benchmarking Conversational Agents on Personalized Long-Term Emotional Support. In *Proceedings of the ACM Web Conference 2026 (United Arab Emirates) (WWW '26)*. Association for Computing Machinery, New York, NY, USA, 5810–5821. <https://doi.org/10.1145/3774904.3792143>
- [12] Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3014–3026. <https://doi.org/10.18653/v1/2022.emnlp-main.195>
- [13] Hsuen-Chi Chiu and Jeremy Foote. 2026. Chatting with Confidants or Corporations? Privacy Management with AI Companions. *Journal of Computer-Mediated Communication* 31, 4 (2026), zmag014. <https://doi.org/10.1093/jcmc/zmag014>
- [14] Kyung Yun Choi and Hiroshi Ishii. 2020. ambienBeat: Wrist-worn Mobile Tactile Biofeedback for Heart Rate Rhythmic Regulation. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction (Sydney NSW, Australia) (TEI '20)*. Association for Computing Machinery, New York, NY, USA, 17–30. <https://doi.org/10.1145/3374920.3374938>
- [15] Jean Costa, Alexander T. Adams, Malte F. Jung, François Guimbretière, and Tanzeem Choudhury. 2016. EmotionCheck: leveraging bodily signals and false feedback to regulate our emotions. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Heidelberg, Germany) (UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 758–769. <https://doi.org/10.1145/2971648.2971752>
- [16] Emmelyn A. J. Croes and Marjolijn L. Antheunis. 2021. Can We Be Friends with Mitsuku? A Longitudinal Study on the Process of Relationship Formation between Humans and a Social Chatbot. *Journal of Social and Personal Relationships* 38, 1 (2021), 279–300. <https://doi.org/10.1177/0265407520959463>
- [17] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F. Jung, Nicola Dell, Deborah Estrin, and James A. Landay. 2024. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 446, 18 pages. <https://doi.org/10.1145/3613904.3642336>
- [18] Abhisek Dash, Soumi Das, Elisabeth Kirsten, Qinyuan Wu, Sai Keerthana Karnam, Krishna P. Gummadi, Thorsten Holz, Muhammad Bilal Zafar, and Savvas Zannettou. 2026. The Algorithmic Self-Portrait: Deconstructing Memory in ChatGPT. In *Proceedings of the ACM Web Conference 2026 (United Arab Emirates) (WWW '26)*. Association for Computing Machinery, New York, NY, USA, 3471–3482. <https://doi.org/10.1145/3774904.3792671>
- [19] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13, 3 (1989), 319–340. <https://doi.org/10.2307/249008>
- [20] Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing Privacy Risks in Online Self-Disclosures with Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 13732–13754. <https://doi.org/10.18653/v1/2024.acl-long.741>
- [21] Barry L. Duncan, Scott D. Miller, Jacqueline A. Sparks, David A. Claud, Lisa Rene Reynolds, Jeb Brown, and Lynn D. Johnson. 2003. The Session Rating Scale: Preliminary psychometric properties of a “working” alliance measure. *Journal of Brief Therapy* 3, 1 (2003), 3–12.
- [22] Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. ChatGPT Outperforms Humans in Emotional Awareness Evaluations. *Frontiers in Psychology* 14 (2023), 1199058. <https://doi.org/10.3389/fpsyg.2023.1199058>
- [23] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4, 2 (2017), e19. <https://doi.org/10.2196/mental.7785>
- [24] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
- [25] James J. Gross. 1998. Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. *Journal of Personality and Social Psychology* 74, 1 (1998), 224–237. <https://doi.org/10.1037/0022-3514.74.1.224>
- [26] James J. Gross. 2015. Emotion Regulation: Current Status and Future Prospects. *Psychological Inquiry* 26, 1 (2015), 1–26. <https://doi.org/10.1080/1047840X.2014.940781>
- [27] James J. Gross and Oliver P. John. 2003. Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology* 85, 2 (2003), 348–362. <https://doi.org/10.1037/0022-3514.85.2.348>
- [28] Yuhao He, Li Yang, Chunlian Qian, Tong Li, Zhengyuan Su, Qiang Zhang, and Xiangqing Hou. 2023. Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-Analysis of Randomized Controlled Trials. *Journal of Medical Internet Research* 25 (2023), e43862. <https://doi.org/10.2196/43862>
- [29] 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. 2025. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI* 2, 4 (2025). <https://doi.org/10.1056/AIoa2400802>

- [30] Philip Held, Sarah A. Pridgen, Daniel R. Szoke, Yaozhong Chen, Zuhair Akhtar, and Darpan Amin. 2025. AI-Facilitated Cognitive Reappraisal via Socrates 2.0: Mixed Methods Feasibility Study. *JMIR Mental Health* 12 (2025), e80461. <https://doi.org/10.2196/80461>
- [31] Annabell Ho, Jeff Hancock, and Adam S. Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication* 68, 4 (2018), 712–733. <https://doi.org/10.1093/joc/jqy026>
- [32] Yuxuan Hu, Minghuan Tan, Chenwei Zhang, Zixuan Li, Xiaodan Liang, Min Yang, Chengming Li, and Xiping Hu. 2024. APTNESS: Incorporating Appraisal Theory and Emotion Support Strategies for Empathetic Response Generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) (CIKM '24). Association for Computing Machinery, New York, NY, USA, 900–909. <https://doi.org/10.1145/3627673.3679687>
- [33] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen Macneil. 2023. Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (UIST '23 Adjunct). Association for Computing Machinery, New York, NY, USA, Article 97, 3 pages. <https://doi.org/10.1145/3586182.3615796>
- [34] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth* 6, 11 (2018), e12106. <https://doi.org/10.2196/12106>
- [35] Sooyeon Jeong, Laura Aymerich-Franch, Kika Arias, Sharifa Alghowinem, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. 2023. Deploying a Robotic Positive Psychology Coach to Improve College Students' Psychological Well-Being. *User Modeling and User-Adapted Interaction* 33, 2 (2023), 571–615. <https://doi.org/10.1007/s11257-022-09337-8>
- [36] Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale. In *Second Conference on Language Modeling*. <https://openreview.net/forum?id=6ox8XZGOqP>
- [37] Eunkyung Jo, Yui Jeong, Sohyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 440, 21 pages. <https://doi.org/10.1145/3613904.3642420>
- [38] Domicile Jonauskaitė, Ahmad Abu-Akel, Nele Dael, Daniel Oberfeld, Ahmed M. Abdel-Khalek, Abdulrahman S. Al-Rasheed, Jean-Philippe Antonietti, Victoria Bogushevskaya, Amer Chamseddine, Eka Chkonja, Violeta Corona, Eduardo Fonseca-Pedrero, Yulia A. Griber, Gina Grimshaw, Aya Ahmed Hasan, Jelena Havelka, Marco Hirnstein, Bodil S. A. Karlsson, Eric Laurent, Marjaana Lindeman, Lynn Marquardt, Philip Mefoh, Marietta Papadatou-Pastou, Alicia Pérez-Albéniz, Niloufar Pouyan, Maya Roinishvili, Lyudmyla Romanyuk, Alejandro Salgado Montejó, Yann Schrag, Aygun Sultanova, Mari Uusküla, Suvi Vainio, Grażyna Wąsowicz, Sunčica Zdravković, Meng Zhang, and Christine Mohr. 2020. Universal Patterns in Color-Emotion Associations Are Further Shaped by Linguistic and Geographic Proximity. *Psychological Science* 31, 10 (2020), 1245–1260. <https://doi.org/10.1177/0956797620948810>
- [39] Matthew Jörke, Defne Genç, Valentin Teutschbein, Shardul Sapkota, Sarah Chung, Paul Schmiedmayer, Maria Ines Campero, Abby C. King, Emma Brunskill, and James A. Landay. 2026. Bloom: Designing for LLM-Augmented Behavior Change Interactions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (CHI '26). Association for Computing Machinery, New York, NY, USA, Article 1167, 27 pages. <https://doi.org/10.1145/3772318.3790506>
- [40] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. 2025. GPTCoach: Towards LLM-Based Physical Activity Coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 993, 46 pages. <https://doi.org/10.1145/3706598.3713819>
- [41] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. 2004. Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Human-Computer Interaction* 19, 1–2 (2004), 61–84. https://doi.org/10.1207/s15327051hci1901&2_4
- [42] Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can Large Language Models be Good Emotional Supporter? Mitigating Preference Bias on Emotional Support Conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 15232–15261. <https://doi.org/10.18653/v1/2024.acl-long.813>
- [43] Minseo Kim, Taemin Kim, Thu Hoang Anh Vo, Yugyeong Jung, and Uichin Lee. 2025. Exploring Modular Prompt Design for Emotion and Mental Health Recognition. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 1181, 18 pages. <https://doi.org/10.1145/3706598.3713888>
- [44] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 701, 20 pages. <https://doi.org/10.1145/3613904.3642937>
- [45] Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1046, 15 pages. <https://doi.org/10.1145/3613904.3642693>
- [46] Richard S. Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press, New York, NY, USA. <https://doi.org/10.1093/oso/9780195069945.001.0001>

- [47] Joseph E. LeDoux. 2000. Emotion Circuits in the Brain. *Annual Review of Neuroscience* 23, 1 (2000), 155–184. <https://doi.org/10.1146/annurev.neuro.23.1.155>
- [48] Belinda Z. Li, Alex Tamkin, Noah D. Goodman, and Jacob Andreas. 2025. Eliciting Human Preferences with Language Models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/c9867d5a22653ce98b02595061e40f12-Abstract-Conference.html arXiv:2310.11589.
- [49] Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E. Kraut, and David C. Mohr. 2023. Systematic Review and Meta-Analysis of AI-Based Conversational Agents for Promoting Mental Health and Well-Being. *npj Digital Medicine* 6, 1, Article 236 (2023). <https://doi.org/10.1038/s41746-023-00979-5>
- [50] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. In *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)*. Association for Computational Linguistics, Online, 3469–3483. <https://doi.org/10.18653/v1/2021.acl-long.269>
- [51] Tianjian Liu, Hongzheng Zhao, Yuheng Liu, Xingbo Wang, and Zhenhui Peng. 2024. ComPeer: A Generative Conversational Agent for Proactive Peer Support. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (Pittsburgh, PA, USA) (UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 117, 22 pages. <https://doi.org/10.1145/3654777.3676430>
- [52] Valeria Lopez Torres. 2023. Before and After Lockdown: A Longitudinal Study of Long-Term Human–AI Relationships. In *Proceedings of the AHFE 2023 International Conference on Applied Human Factors and Ergonomics*. AHFE International, Orlando, FL, USA. <https://doi.org/10.54941/ahfe1004188>
- [53] Kate Loveys, Mark Sagar, Xueyuan Zhang, Gregory Frichione, and Elizabeth Broadbent. 2021. Effects of Emotional Expressiveness of a Female Digital Human on Loneliness, Stress, Perceived Support, and Closeness Across Genders: Randomized Controlled Trial. *Journal of Medical Internet Research* 23, 11 (2021), e30624. <https://doi.org/10.2196/30624>
- [54] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 13851–13870. <https://doi.org/10.18653/v1/2024.acl-long.747>
- [55] James L. McGaugh. 2004. The Amygdala Modulates the Consolidation of Memories of Emotionally Arousing Experiences. *Annual Review of Neuroscience* 27, 1 (2004), 1–28. <https://doi.org/10.1146/annurev.neuro.27.070203.144157>
- [56] Albert Mehrabian. 1996. Pleasure-Arousal-Dominance: A General Framework for Describing and Measuring Individual Differences in Temperament. *Current Psychology* 14, 4 (1996), 261–292. <https://doi.org/10.1007/BF02686918>
- [57] Jingbo Meng and Yue Dai. 2021. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not? *Journal of Computer-Mediated Communication* 26, 4 (2021), 207–222. <https://doi.org/10.1093/jcmc/zmab005>
- [58] Niloofar Miresghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. arXiv:2310.17884.
- [59] Saif M. Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Melbourne, Australia, 174–184. <https://doi.org/10.18653/v1/P18-1017>
- [60] Inbal Nahum-Shani, Shawna N. Smith, Bonnie J. Spring, Linda M. Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A. Murphy. 2018. Just-in-Time Adaptive Interventions (JITAs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine* 52, 6 (2018), 446–462. <https://doi.org/10.1007/s12160-016-9830-8>
- [61] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V. Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F. Huckins, Jason Holden, Sarah M. Preum, Colin Depp, Nicholas Jacobson, Mary P. Czerwinski, Eric Granholm, and Andrew T. Campbell. 2024. MindScape Study: Integrating LLM and Behavioral Sensing for Personalized AI-Driven Journaling Experiences. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 186 (Nov. 2024), 44 pages. <https://doi.org/10.1145/3699761>
- [62] Kevin N. Ochsner and James J. Gross. 2005. The Cognitive Control of Emotion. *Trends in Cognitive Sciences* 9, 5 (2005), 242–249. <https://doi.org/10.1016/j.tics.2005.03.010>
- [63] OpenAI. 2026. GPT-4o Mini TTS Model. <https://developers.openai.com/api/docs/models/gpt-4o-mini-tts>. Accessed September 3, 2026.
- [64] OpenAI. 2026. GPT-5 Mini Model. <https://developers.openai.com/api/docs/models/gpt-5-mini>. Accessed September 3, 2026.
- [65] Dariya Ovsyannikova, Victoria Oldemburgo de Mello, and Michael Inzlicht. 2025. Third-Party Evaluators Perceive AI as More Compassionate than Expert Humans. *Communications Psychology* 3, 1 (2025), 4. <https://doi.org/10.1038/s44271-024-00182-6>
- [66] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. 2014. PopTherapy: Coping with Stress through Pop-Culture. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Oldenburg, Germany, 109–117. <https://doi.org/10.4108/icst.pervasivehealth.2014.255070>
- [67] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a Multimodal Sensor Dataset for Continuous Emotion Recognition in Naturalistic Conversations. *Scientific Data* 7, 1 (2020), 293. <https://doi.org/10.1038/s41597-020-00630-y>
- [68] Luiz Pessoa. 2008. On the Relationship Between Emotion and Cognition. *Nature Reviews Neuroscience* 9, 2 (2008), 148–158. <https://doi.org/10.1038/nrn2317>

- [69] Kexin Quan, Dina Albassam, Mengke Wu, Zijian Ding, and Jessie Chin. 2026. Towards AI as Colleagues: Multi-Agent System Improves Structured Ideation Processes. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 817, 26 pages. <https://doi.org/10.1145/3772318.3790375>
- [70] Kexin Quan, Pavithra Ramakrishnan, and Jessie Chin. 2025. Can AI Take a Joke—Or Make One? A Study of Humor Generation and Recognition in LLMs. In *Proceedings of the 2025 Conference on Creativity and Cognition (C&C '25)*. Association for Computing Machinery, New York, NY, USA, 431–437. <https://doi.org/10.1145/3698061.3734388>
- [71] James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. <https://doi.org/10.1037/h0077714>
- [72] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 5986–6004. <https://doi.org/10.18653/v1/2024.acl-long.326>
- [73] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, and Corina Sas. 2019. HCI and Affective Health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–17. <https://doi.org/10.1145/3290605.3300475>
- [74] Klaus R. Scherer. 2001. Appraisal Considered as a Process of Multilevel Sequential Checking. In *Appraisal Processes in Emotion: Theory, Methods, Research*, Klaus R. Scherer, Angela Schorr, and Tom Johnstone (Eds.). Oxford University Press, New York, NY, USA, 92–120. <https://doi.org/10.1093/oso/9780195130072.003.0005>
- [75] Matthias Schmidmaier, Jonathan Rupp, Darina Cvetanova, and Sven Mayer. 2024. Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 456, 18 pages. <https://doi.org/10.1145/3613904.3642035>
- [76] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 903, 20 pages. <https://doi.org/10.1145/3613904.3642152>
- [77] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating Self-Guided Mental Health Interventions Through Human-Language Model Interaction: A Case Study of Cognitive Restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 700, 29 pages. <https://doi.org/10.1145/3613904.3642761>
- [78] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 9977–10000. <https://doi.org/10.18653/v1/2023.acl-long.555>
- [79] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. A Longitudinal Study of Self-Disclosure in Human–Chatbot Relationships. *Interacting with Computers* 35, 1 (2023), 24–39. <https://doi.org/10.1093/iwc/iwad022>
- [80] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A Longitudinal Study of Human–Chatbot Relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903. <https://doi.org/10.1016/j.ijhcs.2022.102903>
- [81] Petr Slovak, Alissa Antle, Nikki Theofanopoulou, Claudia Daudén Roquet, James Gross, and Katherine Isbister. 2023. Designing for Emotion Regulation Interventions: An Agenda for HCI Theory and Research. *ACM Trans. Comput.-Hum. Interact.* 30, 1, Article 13 (March 2023), 51 pages. <https://doi.org/10.1145/3569898>
- [82] Wally Smith, Greg Wadley, Sarah Webber, Benjamin Tag, Vassilis Kostakos, Peter Koval, and James J. Gross. 2022. Digital Emotion Regulation in Everyday Life. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 444, 15 pages. <https://doi.org/10.1145/3491102.3517573>
- [83] Inhwa Song, SoHyun Park, Sachin R Pendse, Jessica Lee Schleider, Munmun De Choudhury, and Young-Ho Kim. 2025. ExploreSelf: Fostering User-driven Exploration and Reflection on Personal Challenges with Adaptive Guidance by Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 306, 22 pages. <https://doi.org/10.1145/3706598.3713883>
- [84] Inhwa Song, Sachin R Pendse, Neha Kumar, and Munmun De Choudhury. 2025. The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW249 (Oct. 2025), 29 pages. <https://doi.org/10.1145/3757430>
- [85] Elena Spiridon and Stephen H. Fairclough. 2017. The Effects of Ambient Blue Light on Anger Levels: Applications in the Design of Unmanned Aircraft GCS. *International Journal of Unmanned Systems Engineering* 5, 3 (2017), 53–69. <https://physiologicalcomputing.org/wp-content/uploads/2017/11/IJUSEng-Ambient-Blue-Light-GCS-Design.pdf>
- [86] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2025. VITA: A Multi-Modal LLM-Based System for Longitudinal, Autonomous and Adaptive Robotic Mental Well-Being Coaching. *J. Hum.-Robot Interact.* 14, 2, Article 38 (March 2025), 28 pages. <https://doi.org/10.1145/3712265>
- [87] Ala N Tak and Jonathan Gratch. 2023. Is GPT a Computational Model of Emotion?. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, IEEE, Piscataway, NJ, USA, 1–8. <https://doi.org/10.1109/ACII59096.2023.10388119>

- [88] Patricia Valdez and Albert Mehrabian. 1994. Effects of Color on Emotions. *Journal of Experimental Psychology: General* 123, 4 (1994), 394–409. <https://doi.org/10.1037/0096-3445.123.4.394>
- [89] Greg Wadley, Wally Smith, Peter Koval, and James J. Gross. 2020. Digital Emotion Regulation. *Current Directions in Psychological Science* 29, 4 (2020), 412–418. <https://doi.org/10.1177/0963721420920592>
- [90] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 87 (April 2024), 35 pages. <https://doi.org/10.1145/3637364>
- [91] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=pZiyCaVuti>
- [92] Anna Xygykou, Panote Siriaraya, Wan-Jou She, Alexandra Covaci, and Chee Siang Ang. 2024. “Can I Be More Social with a Chatbot?”: Social Connectedness through Interactions of Autistic Adults with a Conversational Virtual Human. *International Journal of Human-Computer Interaction* 40, 24 (2024), 8937–8954. <https://doi.org/10.1080/10447318.2023.2292880>
- [93] Nutchanon Yongsatianchot, Parisa Ghanad Torshizi, and Stacy Marsella. 2023. Investigating Large Language Models’ Perception of Emotion Using Appraisal Theory. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, Piscataway, NJ, USA, 1–8. <https://doi.org/10.1109/ACIIW59127.2023.10388194>
- [94] Bin Yu, Jun Hu, Mathias Funk, and Loe Feijs. 2018. DeLight: Biofeedback Through Ambient Light for Stress Intervention and Relaxation Assistance. *Personal and Ubiquitous Computing* 22, 4 (2018), 787–805. <https://doi.org/10.1007/s00779-018-1141-6>
- [95] Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond C. Ong. 2024. Large Language Models are Capable of Offering Cognitive Reappraisal, if Guided. In *First Conference on Language Modeling*. <https://openreview.net/forum?id=yK8MT91dQY>
- [96] Shuning Zhang, Rongjun Ma, Ying Ma, Shixuan Li, Yiqun Xu, Xin Yi, and Hewu Li. 2025. Understanding Users’ Privacy Perceptions Towards LLM’s RAG-based Memory. In *Proceedings of the 2025 Workshop on Human-Centered AI Privacy and Security (HAIPS ’25)*. Association for Computing Machinery, New York, NY, USA, 10–19. <https://doi.org/10.1145/3733816.3760750>
- [97] Tengan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. ESCoT: Towards Interpretable Emotional Support Dialogue Systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13395–13412. <https://doi.org/10.18653/v1/2024.acl-long.723>
- [98] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. “It’s a Fair Game”, or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI ’24)*. Association for Computing Machinery, New York, NY, USA, Article 156, 26 pages. <https://doi.org/10.1145/3613904.3642385>
- [99] Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhua Luo. 2025. Customizing Emotional Support: How Do Individuals Construct and Interact With LLM-Powered Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI ’25)*. Association for Computing Machinery, New York, NY, USA, Article 376, 20 pages. <https://doi.org/10.1145/3706598.3713453>
- [100] Qingxiaoyang Zhu, Austin Chau, Michelle Cohn, Kai-Hui Liang, Hao-Chuan Wang, Georgia Zellou, and Zhou Yu. 2022. Effects of Emotional Expressiveness on Voice Chatbot Interactions. In *Proceedings of the 4th Conference on Conversational User Interfaces (Glasgow, United Kingdom) (CUI ’22)*. Association for Computing Machinery, New York, NY, USA, Article 22, 11 pages. <https://doi.org/10.1145/3543829.3543840>

A RUNTIME ORCHESTRATION

Figure 12 summarizes how the system routes event narrations, delivers guidance, estimates emotion after the conversation, and updates memory for subsequent sessions.

B PROMPT EVALUATION DETAILS

Table 4 reports the full ranking distribution from the formative human evaluation (Section 3.1). Cog placed in the top two in 72% of rankings and last in only 7%, and won 84% of its direct matchups against Baseline; Combined earned the most first-place finishes after Cog (34 vs. 39 of 90) and won 78% of its matchups against Baseline. Neuro lost 76% of its direct matchups against Cog, and Baseline ranked last in 59% of rankings.

For each scenario, we determined whether each rater ranked Cog or Combined higher and used majority vote across the three raters to identify the preferred prompt, following Section 3.1. Preferences split 15:15 across the 30 scenarios. Cog was preferred in 14 of 17 single-component scenarios, and Combined in 12 of 13 multi-component scenarios. Selecting Cog for single-component events and Combined for multi-component events therefore matched the majority preference in 26 of these 30 formative scenarios (87%; Figure 3). The four mismatches involved rater disagreement.

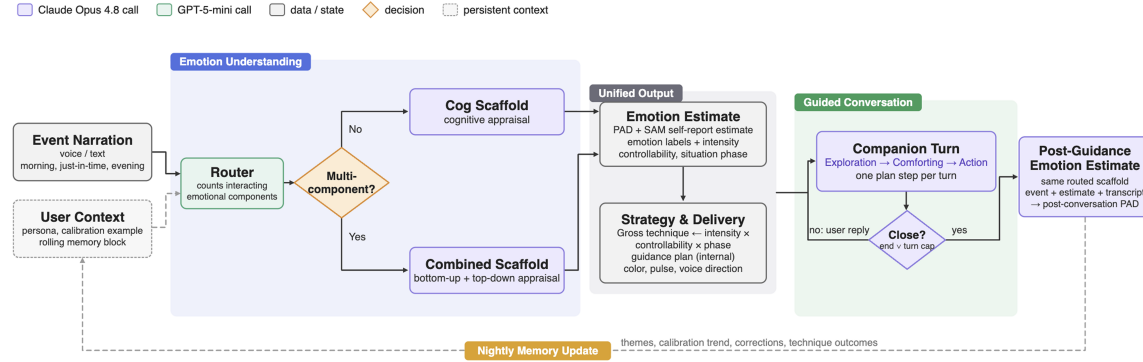


Fig. 12. Runtime orchestration for a single session. An event narration and the persisted user context enter a router that counts interacting emotional components and sends the narration to the cognitive-appraisal scaffold or, where several components interact, to the combined scaffold. Both produce one unified output: a dimensional emotion estimate with a matching SAM reading, emotion labels, appraised intensity, controllability, and situation phase, from which the regulation strategy and all delivery parameters are derived. The guided conversation then advances one plan step per turn along the exploration, comforting, and action arc until it closes. The scaffold selected by the router then re-runs over the event, the estimate, and the transcript to produce a post-guidance estimate. A nightly pass folds themes, calibration trend, corrections, and technique outcomes back into the user context.

Scenario-level overall preferences coincided with those based on reasoning fidelity in all 30 scenarios and outcome plausibility in 27.

Table 4. Ranking distribution of the four emotion-inference prompts across 90 scenario-rater rankings (30 scenarios $\times$ 3 raters, overall-preference dimension). Each cell counts how often a prompt placed at that rank in the blind bracket tournament; lower mean rank is better. Results were consistent across the outcome-plausibility and reasoning-fidelity dimensions, which are omitted for brevity.

Prompt	#1	#2	#3	#4	Mean rank
Cog	39	26	19	6	1.91
Combined	34	31	8	17	2.09
Neuro	13	12	51	14	2.73
Baseline	4	21	12	53	3.27

C PERSONA INTERVIEW QUESTIONS

The persona interview was administered conversationally by the companion during onboarding. Participants answered by voice, with the transcript editable as text, and each question included optional hint prompts. The seven questions, in order, with the construct each targets:

- (1) (*self-description*) “Alright, no more checkboxes. Let’s just talk for a bit. Want to give me a quick intro of yourself, so I can get to know you better?”
- (2) (*typical day*) “I love that. Now, tell me about your life right now — what does a typical day look like?”
- (3) (*coping style*) “I can picture that. When things feel heavy or tiring, what usually helps you cope?”
- (4) (*support sources*) “That sounds like something important for you. When life feels hard, who or what do you usually lean on?”

- (5) (*current preoccupations*) “I’m glad to know what helps hold you up. What’s been taking up a lot of space in your mind lately?”
- (6) (*good day*) “I hear you. And on the other side — what does a genuinely good day feel like for you?”
- (7) (*mood contagion*) “Do other people’s moods tend to rub off on you easily?”

D POST-STUDY QUESTIONNAIRE

All items rated on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree).

Section 1: Perceived Emotional Understanding (RQ1)

- (1) The companion understood how I was feeling in the moment, based on what I described.
- (2) The companion accurately anticipated how I would feel after our conversations.
- (3) The companion picked up on the nuances of how I was feeling, not just the obvious emotions.
- (4) The companion’s emotion predictions felt personalized to me rather than generic.
- (5) The companion seemed to get to know me better over the two weeks.

Q1 and Q3 adapted from PETS [75]; Q2, Q4, and Q5 developed for this study.

Section 2: Perceived Emotional Guidance (RQ2)

- (6) The companion helped me better understand my own emotions.
- (7) The companion’s guidance was relevant to my emotional situation.
- (8) The companion’s guidance gave me concrete ideas for managing my emotions.
- (9) The companion’s guidance helped me see my situation from a different perspective.
- (10) The companion supported me in coping with emotional situations.

Q7 adapted from the Session Rating Scale [21]; Q9 derived from the cognitive-reappraisal construct [25]; Q10 adapted from PETS [75]; Q6 and Q8 developed for this study.

Section 3: Technology Acceptance

- (11) Using the companion helped me manage my emotions more effectively in everyday life.
- (12) The companion was easy to use, and interacting with it did not require much effort.

Q11 and Q12 adapted from the Technology Acceptance Model [19].

E SEMI-STRUCTURED INTERVIEW PROTOCOL

The interview lasted approximately 20–25 minutes and was audio-recorded. Participants first explained their ratings on the post-study questionnaire, then took part in an open-ended interview reflecting on their experiences and responding to the following questions.

- (1) **Overall experience.** How would you describe your experience over the past two weeks?
- (2) **Surprise moments.** Was there a specific moment when the companion’s response surprised you? What made it right or wrong?
- (3) **Emotion understanding.** How accurate did the system’s emotion interpretations feel? Did accuracy change over time? Were the explanations meaningful and specific?

- (4) **Suggestion adoption.** Which suggestions did you try? What worked, what didn't? Any suggestions that felt reasonable but you didn't act on?
- (5) **Emotional awareness.** Did using the companion change how you notice or think about your emotions? Any new patterns you became aware of?
- (6) **Most/least useful.** What was most useful? Least useful or most frustrating? *Probes:* the colors, expressions, and movement; the burden of the daily check-ins; desired changes.
- (7) **Open.** Anything else you'd like to share?

F PARTICIPANT CHARACTERISTICS

Table 5 reports the demographic, AI-background, and baseline expressive-suppression measures collected in the onboarding questionnaires (Section 5.1).

Table 5. Participant characteristics from the onboarding questionnaires ($N = 19$). AI familiarity was rated on a 5-point scale from not familiar at all to extremely familiar. The three ERQ-S expressive-suppression items were rated on a 7-point scale from strongly disagree to strongly agree; the composite is their mean (Cronbach's $\alpha = .74$).

Characteristic	n (%) or M (SD)
<i>Demographics</i>	
Age in years	25.1 (5.1)
Age range	18–33
Woman	13 (68%)
Man	6 (32%)
Asian	15 (79%)
White	2 (11%)
Black or African American	2 (11%)
High school or equivalent	5 (26%)
Some college	1 (5%)
Bachelor's degree	2 (11%)
Master's degree	9 (47%)
Doctoral degree	2 (11%)
<i>AI background</i>	
Familiarity with AI tools (1–5)	3.95 (0.91)
Very or extremely familiar	13 (68%)
Uses AI tools sometimes	4 (21%)
Uses AI tools often	6 (32%)
Uses AI tools very often	9 (47%)
Prior use of AI for emotional support	8 (42%)
<i>ERQ-S expressive suppression (1–7)</i>	
I keep my emotions to myself	4.53 (1.65)
I control my emotions by not expressing them	3.84 (1.61)
When I am feeling negative emotions, I make sure not to express them	3.79 (1.62)
Composite	4.05 (1.32)

G DATA COLLECTED

Table 6 summarizes the data collected per session and study phase.

Table 6. Data collected per session and per study phase.

Source	Data
Onboarding (Day 0)	Demographics; AI background; ERQ suppression; persona interview responses; calibration narration + SAM ratings
Self-report (per session)	Before/after SAM ratings (valence, arousal, dominance); three feedback items; understanding rating + correction text if ≤ 4 ; rotating-question responses
System logs (per session)	Narration text and modality; initial and post-guidance affective estimates (PAD, SAM, emotion labels, intensity); selected regulation strategy; per-turn conversational stage and support-strategy labels; full transcript with timestamps; delivery-channel state (color, pulse speed, expression sequence, speech on/off, language); prompt version and persona snapshot
Engagement telemetry	App opens; just-in-time logging frequency and timing; missed sessions; voice vs. text ratio; slider interaction dynamics
Post-study (Day 14+)	12-item questionnaire; semi-structured interview recording

H WITHIN-SESSION CHANGES IN SELF-REPORTED AFFECT

Figure 13 reports the marginal change in each SAM dimension, pooling sessions across the entry-state quadrants examined in Section 6.3.1. Descriptive means are session-weighted; confidence intervals use 10,000 whole-participant bootstrap resamples. Inferential tests compare the 19 participants’ mean changes against zero using exact two-sided Wilcoxon signed-rank tests, with Holm correction across the three dimensions.

For the regulation-field visualization (Figure 8), we clustered nearby pre-guidance coordinates within each entry-state quadrant and drew a vector from each cluster’s mean starting state to its mean post-guidance state. We used three clusters for the smallest quadrant and four for each of the other quadrants to retain local variation without plotting every trajectory. One session from each quadrant illustrates a direction visible in the corresponding field and connects the reported event with the guidance and post-guidance state.

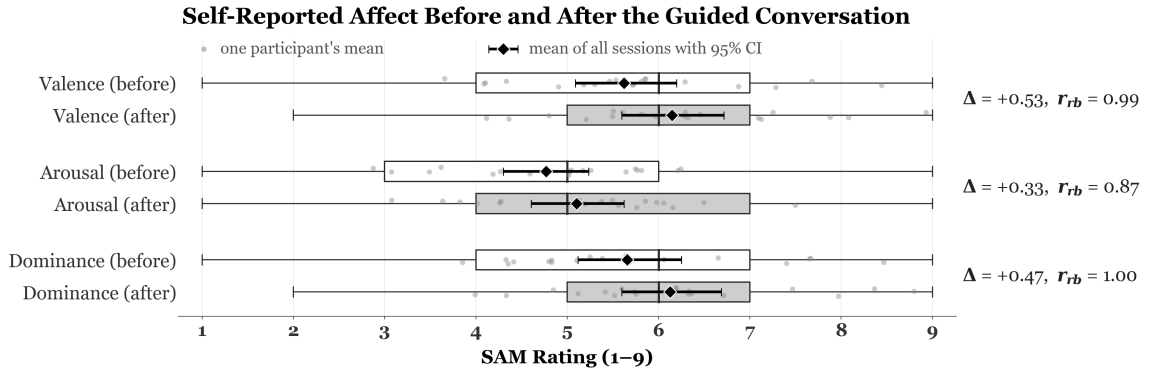


Fig. 13. Self-reported affect before and after guided conversation. Boxes show session-level distributions, grey points show each participant’s mean, and diamonds with error bars show the session-weighted mean across sessions with 95% participant-bootstrap confidence intervals. Δ denotes mean within-session change and r_{rb} the rank-biserial correlation from participant-level Wilcoxon signed-rank tests. All three contrasts remained significant after Holm correction ($p_{\text{Holm}} < .001$).

I IMMEDIATE FEEDBACK BY REGULATION STRATEGY

Table 7 reports immediate feedback by selected strategy. For each rating, we fitted a linear mixed-effects model with strategy as a fixed effect and random intercepts for participant and participant-day. Likelihood-ratio tests compared models with and without the strategy term, and Benjamini–Hochberg correction controlled the false discovery rate across the three tests.

Table 7. Mean immediate feedback on 7-point scales by selected regulation strategy. SM denotes situation modification ($n = 317$), CC cognitive change ($n = 246$), RM response modulation ($n = 256$), and AD attentional deployment ($n = 274$). Means are unadjusted session averages. Test statistics refer to the strategy term in the mixed-effects models.

Rating	SM	CC	RM	AD	$\chi^2(3)$	p	p_{FDR}
Helpfulness	5.03	5.16	5.01	5.00	3.00	.391	.395
Feasibility	5.28	5.18	5.20	5.04	2.98	.395	.395
Willingness	5.24	5.09	5.09	5.10	3.72	.293	.395

J SESSION MEASURES OVER THE DEPLOYMENT

Table 8 reports the full longitudinal results for the nine session measures summarized in Sections 6.2 and 6.4. Each measure was modelled twice with the same random-effects structure, random intercepts for participant and for participant-day. The primary model treats study day as a continuous predictor and estimates a per-day slope. The secondary model replaces day with a week factor (W1 = days 1–7, W2 = days 8–14) and estimates the W2 minus W1 contrast. Sessions per day is a property of the participant-day and was modelled on daily aggregates; the other eight measures were modelled at the session level. Counts used a negative binomial family with a log link, narration length and characters per turn a Gaussian family on the log scale, and absolute valence error a Gaussian family on the log1p scale, so their estimates are on the link scale and correspond to -0.50% , -3.37% , -4.30% per day for conversation turns, narration length, and characters per turn. For log-linked outcomes, the table reports $100b$ on the link scale, while percentage changes in the main text are obtained as $100(\exp(b) - 1)$. The remaining measures, all 7-point ratings, used a Gaussian family on their observed scale. Benjamini–Hochberg correction controlled the false discovery rate across the nine tests within each model. Separate exploratory models added participant-specific random slopes for study day to assess individual variation in change.

Table 8. Longitudinal models for the nine session measures across the 14-day deployment. M (SD) are unadjusted means over all sessions and within Week 1 (days 1–7) and Week 2 (days 8–14). b_{day} is the per-day slope from the day-continuous mixed-effects model (reported $\times 100$), and Δ_{week} is the Week 2 minus Week 1 contrast from the week-binned model. Estimates for count and log-transformed outcomes are on the link scale.

Measure	M (SD)			Day-continuous model		Week-binned model	
	Overall	Week 1	Week 2	$b_{\text{day}} \times 100$ [95% CI]	p_{FDR}	Δ_{week} [95% CI]	p_{FDR}
<i>Counts (negative binomial)</i>							
Sessions per day	4.13 (1.06)	4.16 (1.13)	4.10 (1.00)	−0.04 [−1.52, 1.45]	.963	−0.019 [−0.137, 0.098]	.747
Conversation turns	10.02 (2.80)	10.23 (2.93)	9.87 (2.66)	−0.50 [−0.98, −0.02]	.091	−0.037 [−0.075, 0.001]	.109
<i>Log-transformed (lengths in characters, e_V in SAM points with $\log 1p$)</i>							
Narration length	59.5 (105.9)	69.9 (140.7)	48.6 (45.0)	−3.43 [−4.75, −2.10]	< .001	−0.234 [−0.341, −0.127]	< .001
Characters per turn	33.3 (32.0)	38.6 (37.0)	27.7 (24.6)	−4.39 [−5.48, −3.30]	< .001	−0.284 [−0.375, −0.194]	< .001
Valence error $ e_V $	1.21 (1.08)	1.27 (1.08)	1.15 (1.07)	−0.69 [−1.42, 0.03]	.087	−0.055 [−0.113, 0.002]	.109
<i>Ratings (1–7)</i>							
Understanding	5.45 (0.81)	5.41 (0.77)	5.49 (0.85)	+1.21 [−0.04, 2.45]	.087	+0.068 [−0.031, 0.168]	.227
Helpfulness	5.04 (1.04)	4.94 (1.07)	5.14 (1.00)	+2.94 [1.12, 4.76]	.005	+0.216 [0.070, 0.361]	.011
Feasibility	5.18 (1.15)	5.12 (1.17)	5.23 (1.12)	+1.34 [−0.45, 3.13]	.182	+0.110 [−0.033, 0.252]	.196
Willingness to try	5.13 (1.23)	5.11 (1.27)	5.16 (1.19)	+1.23 [−0.61, 3.06]	.214	+0.076 [−0.070, 0.222]	.348