Hear You in Silence: Designing for Active Listening in Human Interaction with Conversational Agents Using Context-Aware Pacing

Zhihan Jiang
zj2445@cumc.columbia.edu
The University of Hong Kong
Hong Kong, SAR, China
Columbia University
New York, United States

Qianhui Chen
2023103597@ruc.edu.cn
Renmin University of China
Beijing, China

Chu Zhang
zhangchu0908@outlook.com
City University of Hong Kong
Studio for Narrative Spaces
Hong Kong, SAR, China
Guangdong University of Technology
Guangzhou, China

Yanheng Li
lydia.yh-li@my.cityu.edu.hk
City University of Hong Kong
Studio for Narrative Spaces
Hong Kong, SAR, China

RAY LC*
ray.lc@cityu.edu.hk
City University of Hong Kong
Studio for Narrative Spaces
Hong Kong, SAR, China

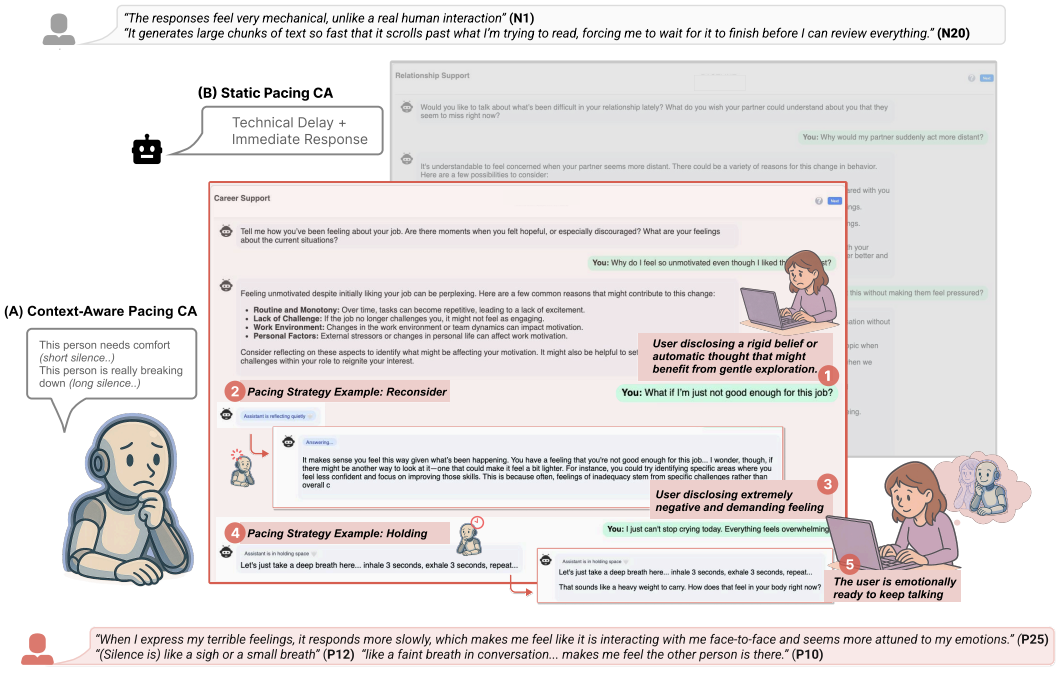


Figure 1: Comparison of two conversational pacing models. (A) A Conversational Agent (CA) with Context-Aware Pacing adapts its response timing to the situation. For example, when a user shares a rigid belief ①, the agent employs a `Reconsider` strategy ②, briefly pausing to reflect before responding. When encountering intense negative emotions ③, it uses a `Holding` strategy ④, creating a prolonged, supportive silence that allows the user to become emotionally ready to continue. (B) In contrast, a CA with Static Pacing defaults to technical delays and immediate responses regardless of context.

*Correspondences can be addressed to ray.lc@cityu.edu.hk.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791238

Abstract

In human conversation, empathic dialogue requires nuanced temporal cues indicating whether the conversational partner is paying attention. This type of "active listening" is overlooked in the design of Conversational Agents (CAs), which use the same pacing for one conversation. To model the temporal cues in human conversation, we need CAs that dynamically adjust response pacing according to user input. We qualitatively analyzed ten cases of active listening to distill five context-aware pacing strategies: *Reflective Silence*, *Facilitative Silence*, *Empathic Silence*, *Holding Space*, and *Immediate Response*. In a between-subjects study (N=50) with two conversational scenarios (relationship and career-support), the context-aware agent scored higher than static-pacing control on perceived human-likeness, smoothness, and interactivity, supporting deeper self-disclosure and higher engagement. In the career-support scenario, the CA yielded higher perceived listening quality and affective trust. This work[1] shows how insights from human conversation like context-aware pacing can empower the design of more empathic human-AI communication.

CCS Concepts

• **Human-centered computing → Collaborative and social computing**.

Keywords

Active Listening, Conversational Pacing, Context-Awareness, Conversational Agents

ACM Reference Format:

Zhihan Jiang, Qianhui Chen, Chu Zhang, Yanheng Li, and RAY LC. 2026. *Hear You in Silence*: Designing for Active Listening in Human Interaction with Conversational Agents Using Context-Aware Pacing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3772318.3791238

1 Introduction

Human connection relies on more than just words; it is profoundly shaped by the subtle, temporal rhythms of communication [17]. In supportive conversations, "active listening" requires the listener to use nuanced pacing to signal empathy, attention, and shared understanding [117]. Silence, in this context, is not an empty void but a powerful communicative tool to show attention and interest [17]. It can be used strategically to hold space for a speaker, encourage deeper self-disclosure, or convey empathy [8, 61]. Understanding the function of this pacing is fundamental to understanding how people build trust and rapport.

Despite the increasing integration of Conversational Agents (CAs) into socially significant roles like emotional support, their design largely ignores subtle but crucial forms of human communication such as pacing [30]. Current systems typically adopt a static and mechanistic interaction style, prioritizing efficiency and immediate responses [77, 81]. This efficiency-first paradigm creates a critical disconnect, resulting in interactions that feel robotic and superficial [30]. The core problem is not just technological, but a failure to model the interpersonal functions of conversational pacing that are essential for human-centric communication [105, 106]. However, historically in the field of CA design, dynamic pacing with respect to textual content has been largely ignored.

Previous research attempting to bridge this gap has faced two major limitations, stemming from an incomplete understanding of the dynamic pacing in human communication:

- **Superficial Models of Timing:** Attempts to manage conversation timing in CAs have often treated pauses as simple, static delays—technical signals to simulate "thinking" or reduce perceived automation [83, 130]. This perspective overlooks the potential for silence to serve as a dynamic, relational tool embedded within the conversation's context.
- **Over-emphasis on Verbal Content:** Research on implementing active listening in CAs has focused almost exclusively on verbal strategies, such as paraphrasing and asking follow-up questions [24]. While valuable, these efforts ignore the non-verbal dimension of pacing, which communication psychology has long established as an intentional communicative act essential for building rapport [8].

What existing works fundamentally overlook is that human communication is highly context-dependent [106]. The appropriateness of a particular pacing strategy shifts based on relational goals [63]. For instance, while immediate responses suit task-oriented goals, supportive contexts demand more patient and reflective pacing to foster connection [50]. Therefore, to design more human-centric and supportive CAs, there is a need to deepen our understanding of how people use pacing strategies in different social contexts.

To address this gap, our study first seeks to understand and systematize human pacing strategies before translating them into computational design. We conducted a qualitative analysis of ten exemplary video recordings of human-to-human active listening to distill the functional uses of silence and response timing. This analysis yielded a taxonomy of five context-aware pacing strategies: *Reflective Silence*, *Facilitative Silence*, *Empathic Silence*, *Holding Space*, and *Immediate Response*, corresponding to eight specific pacing strategies (`Recognize`, `Reconfirm`, `Re-engage`, `Reposition`, `Reconsider`, `Resonate`, `Holding`, and `Resolve`). Based on these findings, we designed and built an LLM-based CA [124, 132] that uses context-aware pacing embodying these strategies. With this CA, our work investigates the value of operationalizing the relational functions of human silence into a mechanistic, implementable framework of context-aware pacing. We hypothesized that users would perceive this pacing as empathetic social cues, facilitating more empathic human-AI communication. The CA dynamically adjusts the response pacing strategies according to user input. More specifically, this work explores these two research questions:

- **RQ1:** How do CAs using context-aware pacing strategies impact users' perceived quality of interaction and experience, specifically in terms of listening quality, affective trust, cognitive trust, human-likeness, smoothness, and interactivity?
- **RQ2:** How does context-aware pacing influence user interaction behaviors in text-based supportive conversations with CAs, specifically in terms of depth of self-disclosure and level of engagement?

[1] Project website: https://zhihanjiang.com/pacing/.

To answer these questions, we conducted a between-subject study (N=50) comparing our context-aware CA against a static-pacing baseline CA across two supportive scenarios (career and relationship advice). These contexts were designed to contrast a scenario blending practical advice with emotional needs (career support) against one centered on interpersonal and affective processing (relationship support). Our work makes the following contributions:

- We challenge the prevailing focus on verbal output in CA design. To the best of our knowledge, we are the first to systematically introduce conversational pacing into CA design, providing empirical evidence that the strategic use of silence for context-aware pacing is critical for developing empathic human-AI relationships analogous to active listening.
- We designed and evaluated a context-aware pacing CA under two supportive scenarios (career and relationship). Our study systematically investigates the effects of context-aware pacing on perceived interaction quality and experience (i.e., listening quality, affective trust, cognitive trust, human-likeness, smoothness, and interactivity) and interaction behaviors (i.e., depth of self-disclosure and level of engagement).
- We propose a new interaction design space centered on dynamic pacing modulation and introduce five concrete, implementable mechanisms: *Reflective Silence*, *Facilitative Silence*, *Empathic Silence*, *Holding Space*, and *Immediate Response*. Our work provides practical pathways and design implications for more empathic human-AI communication.

2 Related Work

2.1 Active Listening in Conversational Agents

Active listening is a dynamic process that combines attentiveness, understanding, and constructive intention [52]. It conveys unconditional acceptance and unbiased attitudes towards speakers' experiences [94]. Rooted in the principle of empathic listening, psychotherapists have developed several active listening strategies, such as eye contact, head-nodding, backchanneling, paraphrasing, and summarizing [66, 116]. Despite being subtle, these cues for listening could produce positive interaction outcomes [117], such as reducing loneliness [41], improving trustworthiness [90], responsiveness [40], and friendliness [14].

One of the primary benefits of active listening is its power to encourage deeper self-disclosure. Guiding individuals to express themselves more deeply has been a central topic in HCI studies [125]. In human-to-human communication, people tend to share more about themselves when they perceive clear benefits or feel emotionally safe [26, 33, 129]. One key barrier to deeper disclosure lies in the anxiety of social evaluation, which refers to the distress or fear experienced in situations where an individual anticipates being judged by others [25]. In contrast, active listening demonstrates a non-judgmental and open attitude towards speakers, thereby encouraging more expression.

Inspired by these benefits, HCI researchers have explored integrating active listening capabilities into CAs. In voice-based CAs, backchannel cues (*"hmm," "yeah"*) could significantly promote a user's sense of being "heard" [21, 84, 120]. Similarly, verbal strategies, such as paraphrasing, verbalizing emotions, summarizing, and encouraging disclosure, also have positive effects in conveying attentiveness and improving user experience with CAs [120]. Social Penetration Theory suggests that the exchange of personal information catalyzes relationship development through reciprocity [3]. Similar to interpersonal communication [78], text-based CAs that disclose their feelings, emotions or preferences could create an equal and affable conversational atmosphere [74], thereby eliciting a sense of being "heard" and enhancing people's perceived intimacy [60]. More recently, researchers have also proposed displaying the agent's "internal empathic resonance" via a secondary text channel to explicitly visualize its understanding of the user [97].

Despite this progress, existing studies primarily focus on *what* an agent says, overlooking a fundamental component of human communication: *how* and *when* it responds. In human dialogue, the non-verbal, temporal dynamics of a conversation (pacing) are powerful communicative signals that convey empathy and presence [17, 117]. By treating active listening primarily as a content-generation problem, existing CA designs have neglected the crucial role of conversational pacing. Our work addresses this critical gap by investigating how strategic silence and timing can be designed to foster deeper connection in human-AI interaction.

2.2 Conversational Pacing for Active Listening

In human conversation, appropriate conversational pacing is a key non-verbal component of active listening. The optimal pace is context-dependent: a fast response could signal proactivity in simple requests [105, 106], whereas slower pacing could provide more space for people to reflect and process complex issues, enacting feelings of being understood and connected [10]. In slower pacing mode, silence is used strategically to convey rich emotional and informational meanings [106]. Unlike unconscious and simple linguistic pauses within conversations, interactive silence is a deliberate tool used to convey intimacy and deep thought [17]. In this context, silence contributes to creating a space for softening people's fierce emotions, leading to shared understanding and nonjudgmental attitudes [37]. For example, silence is especially emphasized in patient-clinician encounters [53, 61]. In end-of-life care, invitational and compassionate silence given by doctors could be consciously trained through contemplative practice, contributing not only to empathy and even "mutual wisdom" towards life [8]. While conversational pacing is a critical component of active listening in human communication, its systematic design and evaluation within CAs remains underexplored [49].

High-quality listening is not associated with a uniformly fast or slow pace. An inappropriately fast reply may feel mechanical and shallow [30], while a persistently slow response can cause confusion, reduce naturalness, and irritate the user [15, 55, 61]. However, within this spectrum, deliberate pacing is often more than just delays in response time. Slower pacing can potentially enhance social presence [30, 38, 39]. This perspective closely aligns with the CASA (Computer As Social Actor) framework, which suggests that people will mindlessly perceive machines demonstrating enough social cues as if they were human [78, 91]. Even minimal social cues (such as gender or ethnicity labels) could trigger unconscious social responses to machines, regardless of the specific demographic or background of the users [78]. Under this framework,

silence, as a fundamental element in interpersonal communication, is not merely a delay in response time but a communicative signal. Similar to human communication, slower pacing could lead to a more natural and smooth experience [5] and trigger the effort heuristic [18, 54], thereby enhancing social presence perceptions [30]. However, slower pacing is also frequently associated with unresponsiveness, especially in situations where a fast response is expected [30]. This trade-off highlights the need for an adaptive approach to pacing [35]. The effectiveness of a CA's pacing is highly dependent on the context, as users have different expectations for task-oriented conversations (prioritizing efficiency) versus socio-emotional conversations (prioritizing support) [63].

Building on this, we introduce context-aware pacing: a strategy where a CA flexibly adjusts its response timing to a user's situational needs on a turn-by-turn basis. Instead of adopting a static pace, such a context-aware pacing CA could use strategic silence to better convey attentiveness and empathy.

2.3 The Role of Delay in Human-AI Interaction

Existing works relevant to the temporal aspects of CAs mainly focus on investigating the technical delay before AI responses, implicitly treating the response delay as a technical flaw, i.e., a system artifact to be mitigated or concealed [72]. By taking on this position, various strategies were explored to reduce negative user perceptions of delayed responses. One common strategy is using textual or graphic indicators to show progress. Some interactive indicators aim at actively involving users in other activities during waiting, shifting users' attention away from slow pacing, and decreasing their perceived waiting time [110]. More recent studies used human-like behaviors such as backchannel fillers to mitigate latency during interaction, which would enhance AI's interactivity [15, 22, 42, 69, 119]. Current LLM-based conversation tools also use some visual cues to indicate progress. For example, ChatGPT[2] uses a pulsating black dot to indicate ongoing processing, displaying messages in real time as they are generated "with great effort". Beyond traditional progress indicators, some studies also call for a more "meaningful" design to demonstrate the uniqueness of a Generative AI system, such as presenting AI's reasoning steps [58] (e.g., Gemini-2.5-pro[3] and GPT-5 Thinking).

Considering the potential trade-offs of slow pacing design, these strategies inform when exploring the communicative potential of machine's "silence", how to minimize its downsides on user experience. However, this perspective overlooks evidence that a slower response is not always detrimental [121]. For example, Park et al. [83] found that when an algorithm provided accurate advice, slower response times actually led to increased user trust and adherence to its suggestions. This finding suggests that "silence" is not inherently negative; its interpretation is highly context-dependent [15]. Moreover, as discussed in Sections 2.1 and 2.2, existing studies overlooked the temporal cues to convey communicative and socio-emotional meaning [106]. Silence can be perceived not just as a system delay but as thoughtful deliberation. To bridge this gap, we designed and evaluated a context-aware pacing CA to provide empirical evidence on how context-aware pacing influences users' perceived interaction quality, experience, and interaction behaviors.

3 Design Space

3.1 Formative Study on Context-Aware Pacing for Active Listening

Most research on active listening in CAs focuses on verbal strategies like paraphrasing and questioning [109, 120], leaving the crucial role of pacing within conversations largely unaddressed. This overlooks how silence functions as a powerful communicative resource in human interaction for emotional articulation and self-awareness [32]. We reframe silence as context-aware pacing, i.e., the strategic use of silence to support a speaker's needs. To understand how to operationalize this concept, we conducted a formative study analyzing real-world active listening cases, focusing specifically on classifying the function of pacing strategies in response to the conversational context.

3.1.1 Active Listening Case Collection. To ensure the validity and diversity of cases, we sourced videos from YouTube[4] using the term "Active Listening Counseling." We selected videos that explicitly mentioned "active listening" in their title or description. The videos were further screened based on the following criteria: (1) content centered around real or simulated counseling dialogues; (2) duration of over 20 minutes, including complete conversational segments; (3) clear visibility of listener behavior, including both verbal and nonverbal strategies; and (4) English as the primary language. From this corpus, we selected 10 cases (ranging in length from 20 to 60 minutes each) covering diverse topics (e.g., exam anxiety, relationship issues, body image) to ensure the generalizability of our findings. The links to the 10 cases can be found in the Appendix A.1.

3.1.2 Data Analysis. We employed an open coding approach to analyze the function of pacing within active listening strategies employed by the counselors [75]. Three researchers independently coded the data, focusing on "strategic use of silence segments" as the unit of analysis. Rather than treating silence as an isolated phenomenon, we analyzed how it was embedded in context, responded to prior utterances, and guided the dialogue. As shown in Figure 2, our coding scheme captured dimensions such as the strategy type, contextual trigger, silence duration, timing, and surrounding verbal responses. Any disagreements were resolved through discussion to reach a consensus.

During the coding process, we reached saturation, meaning that no new categories emerged as we continued analyzing the data. The categories identified effectively covered all instances of silence and its strategic functions, indicating that we had fully captured the relevant behaviors within this context. To ensure consistency and reliability in the coding, we adopted a multi-coder approach. After each researcher independently coded the data, we compared their results and resolved any discrepancies through discussion. This discussion process helped us maintain consistency while ensuring the accuracy of the coding scheme.

[2]https://chatgpt.com/
[3]https://gemini.google.com/
[4]https://www.youtube.com/
[5]Dating Anxiety: https://www.youtube.com/watch?v=Xj3q96mCfC8

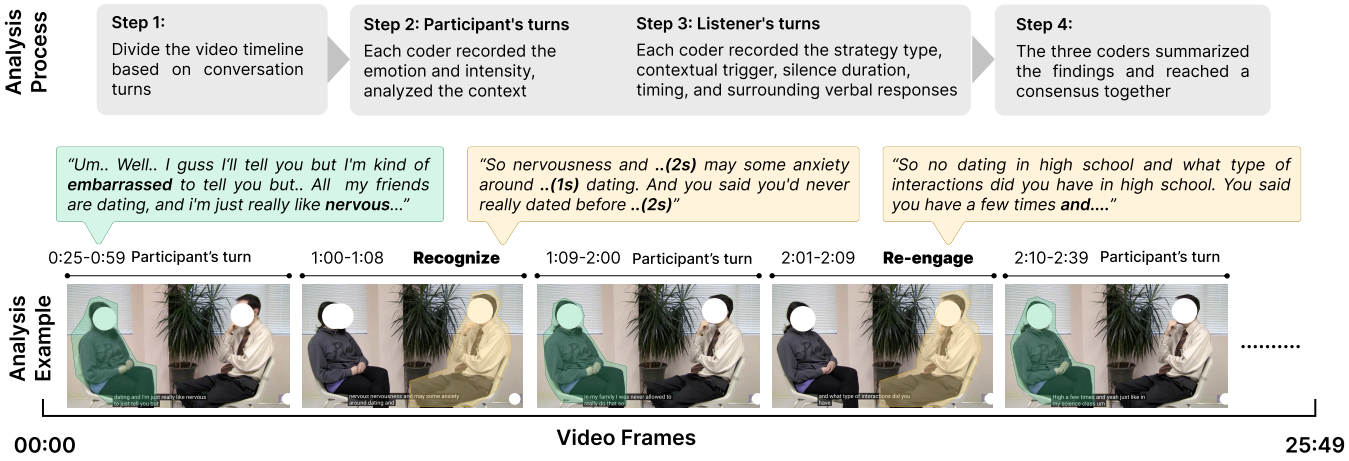


Figure 2: Data analysis process and an analysis example for a video where the speaker faced dating anxiety[5].

3.1.3 Pacing-Based Listening Strategies. Ultimately, we identified five key types of pacing-based strategies, each serving distinct communicative and psychological support functions within conversations: *Reflective Silence*, *Facilitative Silence*, *Empathic Silence*, *Holding Space*, and *Immediate Response*. *Reflective Silence* conveys the listener's thoughtful processing and confirmation of the speaker's message. Through brief silence and reflective responses, the listener signals that they are seriously attending to the content, thereby reinforcing trust and engagement. *Facilitative Silence* is employed in contexts of ambiguity, disruption, or topic transition. Here, silence acts as a subtle prompt, encouraging the speaker to clarify thoughts or re-engage in narration. *Empathic Silence* broadly applies when speakers articulate deep feelings, beliefs, or values. By remaining silent, the listener expresses acceptance and empathy without judgment, fostering self-expression and emotional integration. *Holding Space* emphasizes creating a non-directive, emotionally safe environment during moments of intense vulnerability. The listener uses silence to allow the speaker to fully experience and process their internal state without interruption or pressure. Finally, *Immediate Response*, which applies no deliberate silence, addresses explicit information-seeking needs efficiently. These strategies represent distinct modes of support in active listening that help respond to emotional disclosures, clarify narratives, and facilitate deeper conversational progression. More specifically, they are operationalized by eight specific, implementable strategies (e.g., `Recognize`, `Reconfirm`, etc.), which are detailed in Table 1.

3.1.4 Distribution of Pacing Strategies. To provide quantitative guidance for CA designers on prioritizing pacing implementations, we analyzed the distribution of the observed strategies across the ten analyzed cases (total $N = 289$ pacing instances). Based on the results, we computed the frequency of each strategy. As shown in Table 1, informational and clarification strategies were predominant: `Resolve` and `Reconfirm` appeared most frequently. Conversely, high-intensity affective strategies reserved for moments of peak emotional vulnerability like `Holding` were deployed sparingly. This distribution was topic-dependent: for instance, the case regarding the Downward Arrow Technique (Case 2) relied heavily on clarification (15 `Reconfirm`), whereas the emotionally intense *"I and Thou - touching pain and anger case"* (Case 6) prioritized deep support (5 `Holding`, 6 `Resonate`). A detailed breakdown of strategy distribution for each video is provided in Table 5 (Appendix A.1).

3.1.5 Pacing Transition Patterns. Additionally, beyond individual response triggers, we examined the relationship between strategies to describe natural pacing shifts. We computed a transition probability matrix [20, 108] that maps the likelihood of moving from one strategy to the next (ranging from 0 to 1) based on the strategy sequences observed in the video corpus (detailed in Appendix A.2). The analysis reveals two distinct patterns: Conversational Inertia and Supportive Arcs. First, we observed strong continuity (inertia) in informational and clarification contexts. As shown in Figure 10, `Resolve` and `Reconfirm` have the highest self-transition probability (0.51, 0.41), suggesting that information seeking or clarification often requires multiple turns to resolve ambiguity. Second, we identified specific "Supportive Arcs" for emotional regulation. Notably, the high-intensity `Holding` strategy rarely persists; instead, it frequently transitions into `Recognize` (0.40) or `Resonate` (0.40). This maps a natural therapeutic trajectory: the listener first holds space for the outburst, and then pivots to validation or emotional resonance once the intensity subsides. Conversely, `Reposition` frequently transitions to `Resolve` (0.33), suggesting that once a user's perspective is successfully reframed, they become ready to move toward solution-oriented action.

3.2 The Design Framework for Context-Aware Pacing in Conversational Agents for Active Listening

3.2.1 Reflective Silence. *Reflective Silence* is mainly applicable when people share their experiences or feelings, or when they indirectly seek advice. For example, words like *"I've been unsure if I'm..."* or *"I just don't know if I can keep going with..."* often indicate such situations. In these cases, the listener should focus on providing emotional support and validation, helping people feel understood

Table 1: Summary of Context-Aware Pacing Strategies in Active Listening

Type	Strategy	Context Trigger	Context Example	Description of Strategy	Strategy Example	Silence Duration	Timing	Frequency
Reflective Silence	`Recognize`	When users need their experiences or feelings to be acknowledged or validated, or when they are seeking advice.	People: *"I just don't know if I can keep going with..."*	Provide emotional validation and reflect it back to them, with the goal of making the user feel understood.	Listener: *"OK, I see. Maybe... (silence) You might be because...(silence) Is that your feelings, right?"*	1–2s	After transition words like *"but"*, *"maybe"*, etc.	21.5%
Facilitative Silence	`Reconfirm`	When the user says something vague, contradictory, or unclear, prompting you to reconfirm.	People: *"I can't quite explain it. It's just... complicated."*	Rephrase the key information shared by the user in a gentle manner and invite them to expand on their narrative.	Listener: *"(silence) Can you tell me a bit more about what you mean?"*	2–3s	Before response	27.3%
	`Re-engage`	The user's story fades out, they pause awkwardly, or stop elaborating, needing a gentle nudge to continue.	People: *"Hmm... (long silence)"*	Offer short, incomplete connecting phrases to encourage the user to continue adding to their story, ending with ellipses to maintain a sense of pause.	Listener: *"So... And because..."*	2-3s	Before response	4.2%
Empathic Silence	`Reposition`	The user seems stuck in a rigid or negative perspective, expressing blame or hopelessness.	People: *"I'll never get anywhere. I'm always stuck."*	After reflecting the user's emotions, guide them to re-examine their feelings or situation, offering positive exploratory suggestions.	Listener: *"I hear you feel stuck, (silence) but have you thought about any small changes you could try?"*	5–6s	Before response	4.2%
	`Reconsider`	The user expresses a rigid belief or automatic thought that might benefit from gentle exploration.	People: *"People like me just don't succeed."*	Listener pauses to consider if the response may hurt.	Listener: *"(silence) It's difficult to face the (situations/problems). How did that experience make you feel?"*	2–3s	Before response	5.9%
	`Resonate`	Whether positive or negative, the user is still immersed in the emotion of their story	People: *"I had a... I was going through a lot... like I'm gonna be free for the rest of my life."*	Whether positive or negative, the listener should first pause, allowing the emotional flow to occur, and then respond emotionally.	Listener: *"(silence) I feel quite touched in a particular kind of way."*	3–15s	Before response	5.9%
Holding Space	`Holding`	The user repeatedly shares something intense, painful, or vulnerable.	People: *"It still feels like it's all happening, like a nightmare I can't wake up from, so overwhelming."*	When the user shares deeply negative emotions or pain, appropriately guide emotional release.	Listener: *"(silence) You seem nervous... You need to breathe."*	3–16s	Before response	2.1%
Immediate Response	`Resolve`	When users seek information and answers directly.	People: *"...What should I do?"*	Listener gives suggestions immediately.	Listener: *"Oh, I see. Maybe you can..."*	0s	Immediate	29.1%

and giving them space to process their emerging thoughts or emotions. This type is operationalized by the `Recognize` strategy.

Prior research shows that appropriate delays can be perceived as more attentive and empathetic [4], and that a "silence followed by feedback" rhythm gives users space to process emotions [21]. Our case analysis likewise found that listeners often keep silent briefly, through nods or short utterances like *"hmm"*, before offering a full response. Therefore, the `Recognize` strategy first offers a short emotional acknowledgment (e.g., *"I see..."*) followed by a natural silence. This structure aims to convey listening and acceptance before a substantive reply.

3.2.2 Facilitative Silence. *Facilitative Silence* applies when people express vague, contradictory, or incomplete thoughts. It is also useful when their narration fades or pauses awkwardly. Typical examples include statements like *"I can't quite explain it. It's just... complicated,"* or moments of extended silence where people fail to continue. We operationalized this through two strategies: `Reconfirm` and `Re-engage`.

Active listening research suggests that people in such states often have not yet formed clear emotions or intentions. Gentle guidance, such as non-verbal nods or verbal prompts in the form of encouraging questions, can help them continue expressing their thoughts and feelings [29, 73]. Our formative study observed two common listener tactics in this situation. First, we observed that listeners often repeat or summarize a user's words to help organize language and extend ideas. Based on this, we propose the `Reconfirm` strategy, which helps people clarify vague or contradictory content. It may repeat or rephrase a person's words to invite elaboration. A typical response might be: *"(restatement) (silence) Can you tell me a bit more about what you mean?"*

Second, when a user's narration trails off, listeners often rely on non-verbal cues (e.g., eye contact and body language), or short vocalizations (e.g., *"So..."* to signal encouragement. Professional listening practices describe this as "deep listening," where silence before a response promotes further expression [27]. Inspired by this, we design the `Re-engage` strategy to be applied when narration fades or pauses awkwardly. It uses short verbal prompts such as *"So..."* to create a guided silent space. For example: *"So... Could you please explicate your actions or feelings at that time?"* This silence allows people to fill in the silence and continue, gently encouraging them to elaborate.

3.2.3 Empathic Silence. Whereas *Reflective Silence* validates the speaker's message, *Empathic Silence* addresses deeper emotional undercurrents, particularly when individuals express vulnerability or negative self-perceptions, such as *"People like me just don't succeed."* The goal here shifts from acknowledgment to active emotional co-regulation and perspective scaffolding, fostering a deeper connection. This type is operationalized by three distinct strategies: `Reposition`, `Reconsider`, and `Resonate`.

Empathic listening requires sensitivity to emotional cues and appropriate responses [13]. Studies also suggest that reflective questioning can deepen emotional involvement or help people step back from overwhelming feelings [12]. Our analysis aligned with this, showing two common patterns for supporting users. First, when users recalled negative experiences, listeners would confirm the emotion, briefly pause, and then redirect positively. Based on this,

we introduce the `Reposition` strategy. When people share negative experiences, `Reposition` prompts them to reconsider their situation from a positive angle by combining silence with a positive reframe. For example, if a user says: *"I'm always stuck,"* the listener first acknowledges the emotion, remains silent for 5–6 seconds, and then responds: *"I hear you feel stuck, (silence) but have you thought about any small changes you could try?"* Second, when users expressed self-blame, listeners instead would guide them with reflective questioning. `Reconsider` encourages them to reflect gently by combining silence with exploratory self-reflection. The listener confirms emotions, keeps silence for 2–3 seconds, and asks: *"(silence) How did that experience make you feel?"*, which avoids pressure while opening new perspectives.

Moreover, when people are immersed in long stories or detailed emotional disclosure, listeners often share a moment of silence before responding with resonance, such as expressing deep understanding or joy. Prior research suggests a correlation between the use of extended silence and increased perceptions of emotional depth and resonance [6]. To simulate this, we propose the `Resonate` strategy, which uses longer silences to create space for emotional connection. When people deliver long emotional narratives, `Resonate` emphasizes longer silences of 3–15 seconds, followed by responses like: *"(silence) I feel quite touched in a particular kind of way."* This rhythm signals genuine empathy and fosters resonance.

3.2.4 Holding Space. *Holding Space* is most applicable when people express intense negative emotions, especially intense vulnerability or distress. For instance, a user may say: *"It still feels like it's all happening, like a nightmare I can't wake up from."* Our case analysis found that continuous disclosure in such states may overwhelm people, sometimes leading to breakdowns such as crying or sobbing. If listeners continue probing, the situation may worsen. Instead, listeners often guide people to breathe deeply to reduce immediate emotional pressure. Research also shows that breathing practices, such as deep breathing, help restore calm and reduce anxiety during emotionally heavy interactions [48]. Based on this, we propose the `Holding` strategy. The listener invites reflection and silence, guiding people to regulate emotions through breathing. For example: *"(silence) You seem nervous... You need to breathe."* This strategy involves pauses of 3–16 seconds, providing time and space for processing emotions. The aim is to create a non-judgmental environment where people can keep silent, reflect, and experience their feelings safely.

3.2.5 Immediate Response. *Immediate Response* is applied when users are not seeking emotional support, but are simply seeking information or advice (e.g., *"So, what should I do now?"*). While rushing to a solution may be counter-productive in supportive contexts, for task-oriented questions, an immediate response is often expected for efficiency and professionalism. This type is operationalized by the `Resolve` strategy, which delivers an answer with no deliberate silence. This approach meets users' expectations for efficiency and avoids the frustration that a redundant or delayed acknowledgment may cause in this specific, non-emotional context.

3.3 Conversational Agent Design and Implementation

To operationalize the findings from our formative study, we designed and implemented a context-aware pacing CA. It incorporates the five pacing types via the eight concrete strategies as listed in Table 1. The CA combines a user-facing front-end with a sophisticated Python backend. The backend utilizes Flask as the web server, LangChain to orchestrate conversational logic, and the OpenAI GPT-4o API[6] for context-aware response generation. Its pipeline, as illustrated in Figure 3, is composed of three core backend modules (*Context Analysis*, *Response Generation*, and *Conversational Memory Modules*) that work in concert to manage conversational context and generate paced responses. All prompts and pacing parameters are provided in Appendix B.

3.3.1 User Interface. The user interface was implemented in JavaScript as a standard chat window. A central design goal was to ensure the agent's context-aware pacing felt natural and non-disruptive. To avoid explicitly revealing the underlying strategies (e.g., displaying "Empathic Silence strategy activated") that could potentially break immersion and make the agent appear overly mechanical, we designed the visual feedback to subtly communicate the agent's processing state without exposing the mechanism. As illustrated in Fig. 3, this was primarily achieved through a dynamic status indicator, which is a status bar that visualizes the agent's activity during silence, signaling attention rather than system lag. The specific text varied to match the agent's internal task and the function of the silence. For example, "Assistant is reflecting quietly" was used for Empathic Silence strategies (e.g., `Resonate`) to signal agent-side deliberation before a thoughtful response. In contrast, "Waiting..." was used for Facilitative Silence strategies (e.g., `Reconfirm`) as a more neutral, brief pause before offering a simple clarifying prompt. The `Holding` strategy used "Assistant is in holding space" both before and during the response to signal that the silence itself was the supportive action, while the `Recognize` strategy uniquely used "Assistant is reflecting and answering..." to show simultaneous processing and responding. These were distinct from the generic "Thinking..." (initial processing) and "Answering..." (response generation) states.

3.3.2 Context Analysis Module. When a user sends a message, this module first classifies the user's intent and emotional state based on the conversational triggers identified in our formative study (see Table 1). This module functions as a prompt-based classifier that selects exactly one of the eight strategies (e.g., information-seeking triggers `Resolve`; high emotional valence triggers `Resonate`) shown in Table 1. These eight strategies are the implementable operations that map to our five high-level conceptual types (e.g., `Resolve` is the implementation of the *Immediate Response* type). Based on the selected strategy, the module generates a control signal containing two key pieces of information: (1) the strategy label for response generation and (2) the appropriate silence duration. This silence simulates natural cognitive processing time before or during "answering."

[6] https://platform.openai.com/docs/models/gpt-4o

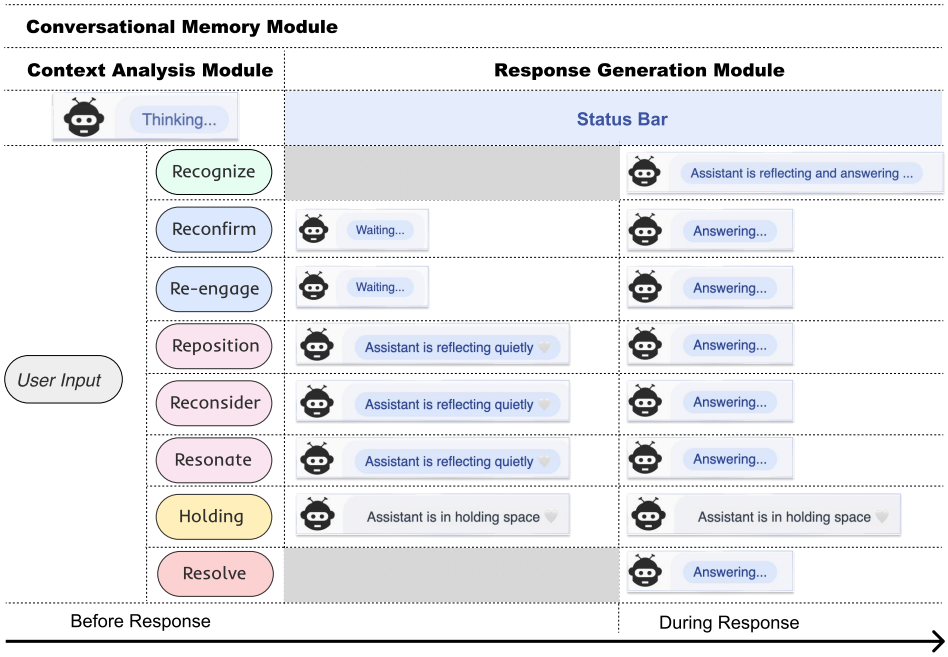


Figure 3: Pipeline and visual elements of the context-aware pacing CA. It consists of three core backend modules (*Context Analysis, Response Generation,* and *Conversational Memory Module*). After receiving user input, the *Context Analysis Module* would first select the most appropriate pacing strategy when the status bar of the CA shows "Thinking...". Then its output and user input would be fed into the *Response Generation Module* to generate responses and apply corresponding pacing strategies. Both modules are supported by the *Conversational Memory Module* to manage the conversational contexts.

3.3.3 Response Generation Module. The control signal from the *Context Analysis Module* then conditions the *Response Generation Module*. This module dynamically generates responses and adjusts behaviors based on the chosen strategy to create a natural conversational rhythm. We implement the context-aware pacing through (1) punctuation-aware micro-pauses (e.g., brief silence at commas and longer silence at ellipses) and (2) applying silence duration calculated by the *Context Analysis Module* according to the corresponding timing as illustrated in Table 1.

Specifically, for `Holding`, the stream begins with a grounding instruction (*"Let's just take a deep breath here... inhale 3 seconds, exhale 3 seconds, repeat..."*) before the silence, and then continues with the generated content. Besides, when the user is inactive for 60 seconds, a separate code path triggers a brief, strategy-labeled `proactive engagement` check-in based on the chat history, such as *"I'm still here if you want to continue."*

3.3.4 Conversational Memory Module. To enable coherent, long-running interactions, the *Conversational Memory Module* manages the dialogue history. We employ a summarization technique where a token budgeter reserves space for future replies. When the context window nears its limit, older turns are summarized into a condensed memory string, while the most recent turns are retained verbatim (e.g., keeping the last 8 turns). This ensures that the *Context Analysis* and *Response Generation Modules* have access to relevant historical context for making accurate, history-aware decisions.

4 User Study

To evaluate how the context-aware pacing CA impacts users' perceived quality of interaction and experience (**RQ1**) and how the context-aware pacing influences users' texting behaviors (**RQ2**), we conducted a between-subjects study under two supportive scenarios (career and relationship) compared to a static-pacing CA baseline. The study has been approved by the Institutional Review Board (IRB) and conducted ethically with university-approved IRB regulations. Participants gave consent to allow us to use their data anonymously with the understanding that the data would be deleted after the analysis.

4.1 Study Design

4.1.1 Task Scenario Description. Participants were asked to engage in two scenarios within supportive communication. One was the career-support scenario, where participants were instructed to adopt the perspective of a young adult experiencing difficulties in career development and workplace communication. They were encouraged to seek advice from the CA on career direction uncertainty, interview anxiety, or professional skill development. Another was the relationship-support scenario, where participants sought guidance on romantic or intimate relationship difficulties, such as communication breakdowns, trust issues, or conflict resolution. These two scenarios simulated common, yet stylistically different, conversational contexts [88, 96], thereby increasing realism.

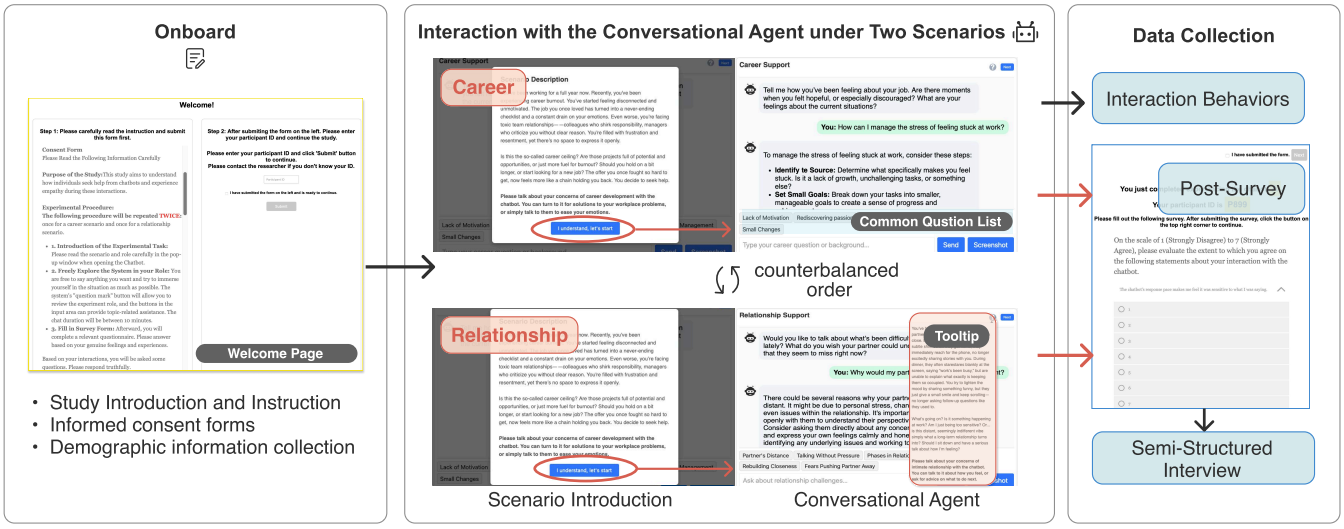


Figure 4: Overview of the user study procedure. The user first landed on the welcome page, and then interacted with the conversational agent under two scenarios, career and relationship support, in a counterbalanced sequence. After completing every scenario, the user completed a survey evaluating the corresponding experience. Finally, a semi-structured interview was conducted.

Aligned with previous HCI research [85, 112], we utilized vignette stories to facilitate participant immersion and familiarity with the assigned topics. Prior work demonstrates that vignettes help participants adopt a protagonist's perspective and elicit real-world perceptions [70, 79, 114], while creating a safer space to discuss private topics such as social comparisons [112] or help-seeking [67]. Following this approach, all participants were invited to read two well-crafted vignette stories before starting conversations with the CA. Both stories employed second-person pronouns and offered detailed accounts of the backgrounds, problems, and feelings that participants were about to discuss. After reading the vignettes, participants were then asked to engage in the dialogues.

In the conversational stage, to ensure the diversity of topics and avoid running out of things to say, different scenarios provided participants with a "common question list." This list included common topics for conversation, but participants were free to choose whether to use them to guide their discussion.

The detailed vignette stories about scenario backgrounds and role setup, CA prompts for each scenario, and the specific common question lists can be found in Appendix C.

4.1.2 Experimental Design. We conducted a between-subjects experiment where 50 participants were randomly assigned to either a context-aware pacing condition or a baseline control condition. Each participant completed two supportive dialogue scenarios, with the order counterbalanced.

Given the free-form nature of user input, it was infeasible to strictly control turn-by-turn content in either condition. Instead, to ensure fair comparison, both conditions used an identical core response procedure: the same underlying model (i.e., GPT-4o), session memory mechanism, streaming transport, and visual interface. The only methodological difference was the high-level prompt strategy. While the specific strategies differed, both prompts leveraged the same high-capability base model with the shared goal of producing high-quality, supportive content.

- **Control group (N=25)**: Participants interacted with a static-pacing baseline CA that responded immediately, using only inherent API latency with no context-aware timing adjustments. The status bar in this condition displayed only generic states ("Thinking..." and "Answering..."). The baseline CA used a generic supportive prompt (Appendix B.2.3).
- **Experimental group (N=25)**: Participants interacted with the designed context-aware pacing CA that dynamically adjusted response pacing according to the user input for active listening. The context-aware CA built upon this generic supportive prompt by adding specific, dynamic pacing strategies (Appendix B.2.2).

As shown in Fig. 4, before using the CA, participants viewed a Consent Form that included the study introduction and demographic information collection. After submitting the consent, they entered the CA interface to begin the experiment. To eliminate sequence effects, the presentation order of the two scenarios was counter-balanced. Approximately half of the participants followed the "Career-Relationship" sequence, while the remainder followed the "Relationship-Career" sequence. A scenario-related vignette story appeared when the interface opened and remained available via the top-right help tooltip. Additionally, the participant's typing area included a 'common question list,' where they could click various questions to add them to their input box as references for topics when they were lacking inspiration. Each scenario dialogue lasted at least 10 minutes.

4.1.3 Participants. We recruited 50 participants by word of mouth and online recruitment via social media. All participants were proficient in reading and writing English, and the study was conducted

Table 2: Demographic Distribution and Chi-Square Analysis ($N = 50$). The analysis confirms no significant differences between groups.

Demographic Variable	Experimental ($N = 25$)	Control ($N = 25$)	χ^2 Statistic	p-value	Result
Age (18-24 vs. 25-34)	14 (56%)	17 (68%)	0.3396	0.5601	Not significant
Gender (Female vs. Other)	15 (60%)	18 (72%)	0.3565	0.5505	Not significant
Employment (Student vs. Other)	19 (76%)	13 (52%)	2.1701	0.1407	Not significant
CA Experience (Daily vs. Non-Daily)	17 (68%)	12 (48%)	1.3136	0.2517	Not significant

in English. We performed the data quality check based on the following exclusion criteria: (1) technical malfunctions where the CA failed to trigger or respond; and (2) insufficient effort (e.g., participants who provided highly homogeneous or largely blank responses). All recruited participants successfully completed the study and passed data quality checks; therefore, no data was excluded.

Among them, 15 identified as males (30%), and 33 identified as females (66%). The remaining participants identified as non-binary or preferred not to disclose their gender. Regarding age, 31 participants (62%) were between 18 and 24 years old, and 19 participants (38%) were between 25 and 34 years old. As for their employment, 64% were students, while 32% were employed. More than half (58%) used CAs daily. Participants were randomly assigned to either the context-aware experimental group or the static-pacing control group. This random assignment ensures that demographic factors and unmeasured preferences (such as personal resonance with a specific scenario) were randomly, rather than systematically, distributed. A post-hoc Chi-Square analysis confirmed that the groups were statistically equivalent: regarding age ($p = 0.56$), gender ($p = 0.55$), and employment status ($p = 0.14$). Specifically, the experimental group ($N = 25$) had 14 participants aged 18-24 (vs. 17 in control), 15 females (vs. 18 in control), and 19 students (vs. 13 in control). Prior CA experience was also broadly distributed, with only one participant in each group reporting no previous use. As shown in Table 2, given that user characteristics are statistically balanced across conditions, we inferred that the baseline propensity to resonate with the topics was equally distributed. Since the career-support and relationship-support scenarios represent common life stressors relevant to this demographic [88, 96], the likelihood to engage deeply with either topic did not differ systematically between groups. Consequently, this allowed us to attribute significant differences in the following sections to our pacing intervention rather than demographic confounds. More details on participant demographics can be found in the Appendix D.

- **Questionnaire**. After completing dialogue under each scenario, participants would fill out a questionnaire to test user experiences. We also conducted manipulation checks. All measures were rated on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). The variables include perceived listening [21], human-likeness [11], cognitive and affective trust [47], smoothness [104], interactivity [57]. Details of all measurement items are provided in Table 3.
- **Interview**. After completing both scenarios, to better understand how participants interpret CA's response pacing during the study, we conducted semi-structured interviews with all participants [2]. We also analyzed participants' responses to the three open-ended questions in the post-survey. We specifically asked how and why participants perceived CA's response pace differently (See Appendix E for the interview outline).
- **Interaction Behaviors**. Participants' interactions with the CA during the experiment, including the conversation content, timestamp of each conversation, and pacing strategies applied (if applicable), were recorded.

4.2 Data Analysis Approach

4.2.1 Questionnaire. We collected participants' perception of (a) perceived listening, (b) human-likeness, (c) trust (cognitive and affective), (d) smoothness, and (e) interactivity. We calculated Cronbach's α to assess the internal consistency of the measurement items (Table 3). Mann-Whitney U tests were conducted to compare the responses from the Experimental group and the Control group.

4.2.2 Interview Thematic Analysis. We employed an iterative, consensus-based thematic analysis [34] on data from post-experiment interviews and open-ended questionnaire responses. Following transcription, four researchers independently performed open coding to identify themes regarding context-aware pacing. Through collaborative discussions and preliminary exercises, we developed an initial codebook, regularly refining codes to ensure reliability until theoretical saturation was achieved. The initial set of codes covered perceptions of context-aware pacing ("response speed," "active listening content," etc.), conversational behaviors in different scenarios (e.g., "seeking emotional support."); emotional experiences during conversations (e.g., "need satisfaction"), etc. Finally, we iteratively refined these into core themes by examining contextual meanings and eliminating irrelevant codes.

4.2.3 Interaction Behaviors. To evaluate how the CA's context-aware pacing influenced participants' texting behaviors, we quantified their depth of self-disclosure and engagement levels based on their conversation contents. The depth of self-disclosure was measured using the total emotion words, first-person pronoun usage, and user turn length:

- Total Emotion Words. We leveraged the NRC Word-Emotion Association Lexicon, which contains over 14,000 English words and associates them with eight basic emotions based on Plutchik's wheel, as well as positive and negative sentiment [86]. We used this library to sum the total count of word occurrences in the user input that are associated with an emotion or sentiment. This metric reflects affective self-disclosure, which is the act of sharing personal feelings and

Table 3: Questionnaire used for collecting participant feedback. All perceptual variables were rated on 7-point Likert scales.

Variable	Item	Cronbach's alpha
Perceived listening	The chatbot's response pace makes me feel it was sensitive to what I was saying.	$\alpha = 0.85$
	The chatbot delivered a sense of agreement for what I was saying when appropriate.	
	The chatbot's response pace kept track of points I made.	
	The chatbot assured me that it got what I said.	
	The chatbot assured me that it was receptive to my ideas.	
	The chatbot's response speed effectively conveyed a good listening attitude.	
Human-likeness	fake/natural	$\alpha = 0.855$
	mechanic/humanlike	
	unconscious/conscious	
	artificial/lifelike	
	rigid/elegant	
Cognitive Trust	I have no reservations about acting on the chatbot's advice.	$\alpha = 0.873$
	I have good reason to trust the chatbot's competence.	
	The chatbot handles my concerns carefully, I can confidently depend on it.	
	I felt my questions were properly analyzed and addressed by the chatbot.	
Affective Trust	The chatbot's response pace makes me feel it was responding to me caringly.	$\alpha = 0.803$
	The chatbot displays a kind and caring attitude toward me.	
	I can talk freely with the chatbot about my problems and know that it will want to listen.	
Smoothness	The conversation flow was smooth.	$\alpha = 0.879$
	The response pace of the chatbot makes me feel that the conversation was natural.	
	The response pace of the chatbot makes me feel comfortable.	
	The response pace of the chatbot makes me feel that the conversation was like an in-person conversation.	
Interactivity	I feel that the chatbot's response pace was interactive.	
Manipulation Check	The chatbot pauses intentionally when responding to me.	
	The chatbot would slow down appropriately when responding to me.	

emotional states [68, 93, 100]. We posit that a higher total emotion word count signifies a greater volume of emotional disclosure.

- Total First-Person Pronoun Usage. We defined lists of first-person singular pronouns (i.e., "I", "me", "my", "myself", and "mine") and plural pronouns (i.e., "we", "us", "our", "ours", and "ourselves") and counted their usage in the user's input. This count serves as a linguistic marker of self-focused attention and self-disclosure, indicating that a speaker is referencing their own experiences, thoughts, or feelings [16, 93].
- User Input Length (in Words). Measuring the word count of user messages is a common method for estimating the depth of self-disclosure, based on the assumption that longer, more detailed responses reflect greater engagement and willingness to share personal information [9, 59, 82].

The user engagement level is measured using the number of conversation turns and pacing strategy distribution:

- Total Conversation Turns [98]. We counted the total number of conversational exchanges, with one turn defined as a user input followed by a CA response. We used this as our primary metric for engagement, based on the assumption that more engaged users will continue the dialogue for a greater number of turns [43, 113, 131].
- Pacing Strategy Distribution. For the experimental group, we further analyzed the frequency distribution of the different pacing strategies applied by the CA. This distribution serves as a proxy for the nature and diversity of user engagement. A more diverse distribution indicates a deeper, multifaceted engagement.

Mann-Whitney U tests were conducted to compare the depth of self-disclosure and engagement levels of the Experimental group and the Control group.

5 Results

In the following sections, we present the results of our quantitative and qualitative analyses. For all Mann-Whitney U tests, we report the U-statistic (U) and p-value (p), and the rank-biserial correlation coefficient (r) which serves as a measure of effect size.

5.1 Manipulation Check

To better examine whether our manipulations of context-aware pacing were successfully designed, we asked two questions about their perceptions of pacing and the appropriateness of pacing (see the survey questionnaire in Table 3). The two-sided Mann-Whitney U Test showed that our manipulations were successful across the two scenarios (career: $U = 56$, $p < 0.001$, $r = 0.71$; relation: $U = 64$, $p < 0.001$, $r = 0.7$). Participants in the experimental group perceived a significant pacing change (career: $M = 5.68$, $SD = 1.25$; relation: $M = 5.4$, $SD = 1.00$) compared to those in the control group (career: $M = 2.92$, $SD = 1.5$; relation: $M = 3.24$, $SD = 1.39$). The experimental group also rated these changes (career: $M = 5.12$, $SD = 1.05$; relation: $M = 5.20$, $SD = 1.08$) as significantly more appropriate (career: $U = 78.5$, $p < 0.001$, $r = 0.65$; relation: $U = 78.5$, $p < 0.001$, $r = 0.68$) than the control group (career: $M = 2.96$, $SD = 1.46$; relation: $M = 3.04$, $SD = 1.46$).

5.2 Semantic Content Analysis for Confound Check

As discussed in the previous section, the inherent variability in the turn-by-turn content generated by CAs made it infeasible to strictly control for content. Therefore, to determine whether there are any underlying differences in the semantic content of the agents' responses between the experimental and control conditions that could confound our findings, we performed a semantic content analysis on the full corpus of agent responses from both groups.

Table 4: The accuracy and distribution of pausing strategies. Specifically, the Career/Relationship-Accuracy represents the number (and percentage) of strategies that were classified correctly by the CA.

Pausing Strategy	All	Career	Career-Accuracy	Relationship	Relationship-Accuracy
`Holding`	8	5 (2.3%)	5 (100%)	3 (1.4%)	3 (100%)
`Resolve`	94	49 (22.7%)	46 (93.9%)	45 (21.4%)	39 (86.7%)
`Recognize`	70	33 (15.3%)	30 (90.9%)	37 (17.6%)	28 (75.7%)
`Reconfirm`	103	47 (21.8%)	41 (87.2%)	56 (26.7%)	44 (78.6%)
`Reconsider`	35	16 (7.4%)	13 (81.3%)	19 (9.0%)	14 (73.7%)
`Reposition`	60	38 (17.6%)	36 (94.7%)	22 (10.5%)	18 (81.8%)
`Resonate`	15	5 (2.3%)	4 (80.0%)	10 (4.8%)	7 (70.0%)
`Re-engage`	41	23 (10.6%)	21 (91.3%)	18 (8.6%)	18 (100%)
Overall	426	216 (50.7%)	196 (90.7%)	210 (49.3%)	171 (81.4%)

We employed the all-MiniLM-L6-v2 sentence-transformer model[7] to generate dense vector embeddings for all agent responses, enabling quantitative comparison of semantic content. For each scenario, we computed pairwise cosine similarities within each condition (intra-group) and between conditions (inter-group) [92]. To assess potential content shifts, we adopted a distributional analysis approach analogous to cluster validity estimation [95]. We posited that if pacing strategies fundamentally altered the content of the advice, the semantic similarity between conditions (inter-group) would be significantly lower than the natural semantic variation observed within a single condition (intra-group). We computed pairwise cosine similarities for each scenario and tested for divergence using the Mann-Whitney U test and Jensen-Shannon divergence for distribution overlap [64].

Results demonstrate strong semantic similarity across both scenarios, with no significant differences found between conditions. For the career scenario, there was no significant difference between the mean within-experimental similarity (0.203) and between-group similarity (0.199), as indicated by a Mann-Whitney U test ($p = 0.440$) and a negligible effect size ($r = 0.004$). The relationship domain showed a similar pattern, with a mean within-experimental similarity of 0.250 and between-group similarity of 0.255. Statistical analysis showed no significant divergence ($p = 0.494$, $r = 0.004$). In both cases, low Jensen-Shannon divergence values (career: 0.033; relationship: 0.085) confirmed that the semantic content distributions were highly overlapping.

These findings suggest that, despite variations, the experimental and control groups exhibit similar semantic content. This check for confounds demonstrates that the perceptual differences reported in the following sections can be more reliably attributed to our primary manipulation of pacing strategies.

5.3 Context Classification Performance

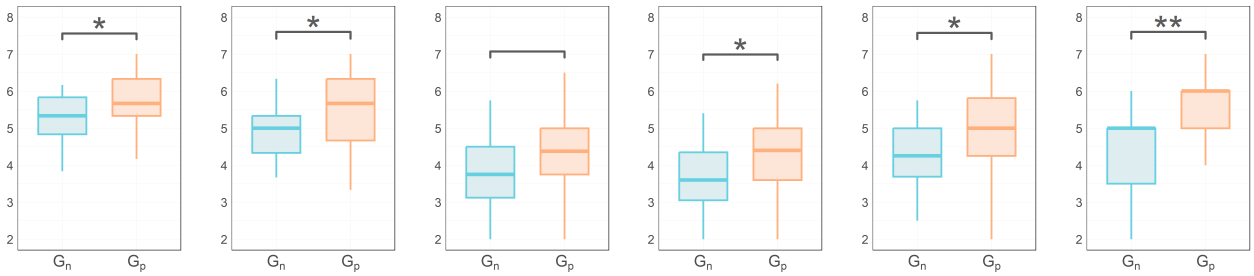
Since our architecture implicitly synthesizes user intent and emotional cues to drive strategy selection, we evaluated the module's reliability based on the accuracy of the final deployed strategy. This metric serves as a holistic proxy for the system's understanding of the conversational context. We conducted a post-hoc evaluation using the interaction logs collected during the user study. Two researchers independently annotated these turns based on the coding scheme defined in Table 1, categorizing the ideal pacing strategy (e.g., `Resolve`, `Recognize`, `Holding`) for each context. Inter-rater reliability was calculated using Cohen's Kappa (career: $\kappa = 0.9$; relation: $\kappa = 0.93$), indicating strong agreement between the two coders. Any discrepancies between the two coders were resolved through discussion to establish a final consensus-based ground truth. We then compared the ground-truth annotations against the actual strategies selected by the Context Analysis Module during the study. The system achieved a high overall accuracy of 86.2%. The accuracy of each strategy can be found in Table 4. Performance was robust across both scenarios, though higher in the Career scenario (90.7%) compared to the Relationship scenario (81.4%). Specifically, the module showed strong precision in identifying task-oriented intents requiring Immediate Response (`Resolve` accuracy: 90.4%) and conversational stalls requiring `Re-engage` (95.1%). Strategies for critical emotional states also performed well, with `Holding` strategy achieving 100% accuracy and `Reposition` achieving 90.0%. However, performance was relatively lower for strategies requiring deeper semantic interpretation of subtle emotional cues, such as `Resonate` (73.3%) and `Reconsider` (77.1%). A qualitative review of these error cases suggests they were primarily benign misclassifications between semantically similar supportive strategies (e.g., substituting `Reconsider` for `Resonate`), rather than critical failures. The results confirm that the pacing strategies were deployed appropriately according to the experimental design.

5.4 The Impact of Context-Aware Pacing on Perceived Interaction Quality and User Experiences (RQ1)

To evaluate how context-aware pacing influenced user perception, we analyzed survey data using two-sided Mann-Whitney U tests and situated these findings within qualitative interview feedback. We compared outcomes for participants using the context-aware agent (experimental group) with those using a static-pacing agent (control group) across two scenarios: career support and relationship support.

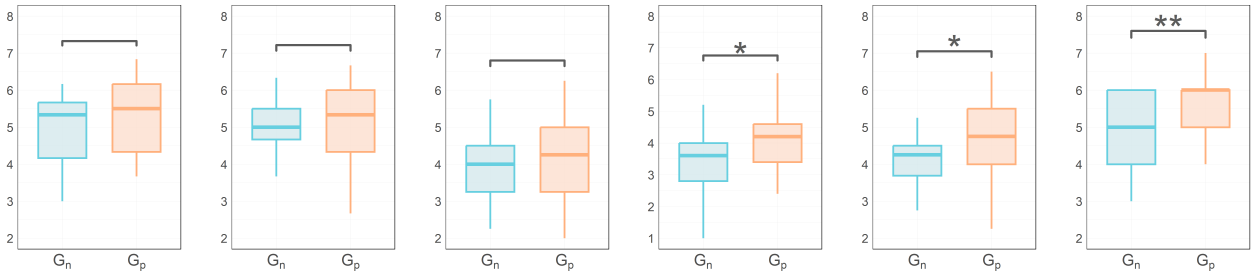
5.4.1 Pacing Increased Affective Trust and Listening Quality in Career-Support Scenario by Simulating Deliberation. A key finding was that context-aware pacing significantly improved relational metrics, though this effect was most pronounced in the career-support scenario. Participants rated the context-aware agent significantly

[7] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2



(a) Listening (b) Affective Trust (c) Cognitive Trust (d) Human-likeness (e) Smoothness (f) Interactivity

Figure 5: Statistical analysis on questionnaire results in career-support scenario, with G_n referring to the control group, and the context-aware pacing group G_p. These figures show the significant difference between the control and experimental group in terms of perceived listening quality, affective trust, human-likeness, smoothness, and interactivity.



(a) Listening (b) Affective Trust (c) Cognitive Trust (d) Human-likeness (e) Smoothness (f) Interactivity

Figure 6: Statistical analysis on questionnaire results in relation-support scenario, with G_n referring to the control group, and the context-aware pacing group G_p. These figures show the significant difference between the control and experimental group in terms of perceived human-likeness, smoothness, and interactivity.

higher on perceived listening quality in the career scenario (M = 5.74, SD = 0.89; U = 188, p = 0.0158, r = 0.343) compared to the control agent (M = 5.05, SD = 1.06), as shown in Figure 5(a). The context-aware pacing also significantly improved affective trust in the career-support scenario (Figure 5(b); M = 5.45 vs. M = 4.73; U = 197, p = 0.024, r = 0.32). These effects were not statistically significant in the relationship scenario (Figure 6), nor was cognitive trust significantly affected in either scenario.

Qualitative feedback explains how context-aware pacing built affective trust: participants interpreted silence as evidence of cognitive effort and care, as shown in Figure 7. This perception of "thinking" enhanced the credibility of the agent's advice and made it feel more supportive. Participants frequently used terms like *"serious thinking"* (P21), *"deliberation"* (P14), and *"special attention"* (P6) to describe this. P21 drew a direct analogy to human behavior: *"If I reply slowly, it also means I am thinking carefully... When I asked whether I could check my partner's phone, the chatbot took longer to respond and corrected my behavior. I felt that its advice was more convincing than if it had responded quickly."* Conversely, rapid responses from the control agent were sometimes perceived as superficial, undermining trust. P14 noted, *"It just picked up on a few of my words, and I wouldn't take its response very seriously."*

This feedback, which links perceived deliberation directly to trust, also helps explain the specificity of our quantitative findings. We interpret this scenario-specific effect as being driven by the nature of the task. The career-support scenario is a hybrid task that blends the need for practical, actionable advice with emotional validation. In this context, the agent's "deliberation" signaled that it was "seriously thinking" about the user's problem, which enhanced the affective trust and listening quality. This cue was less relevant in the more purely affective relationship scenario. This also explains the lack of effect on cognitive trust, which relates to perceived competence and accuracy. The pacing didn't make the agent seem smarter (it was the same GPT-4o model), but it did make it seem more caring and less superficial, thus boosting affective trust specifically in the context where thoughtful advice was most valued.

5.4.2 Pacing Enhanced Human-Likeness by Creating a Sense of Social Presence. Context-aware pacing could enhance participants' human-likeness perception in both career-support scenario (U = 182, p = 0.011, r = 0.36) and relationship-support (U = 206, p = 0.039, r = 0.30), as shown in Figure 5(d) and Figure 6(d). Participants explained that the temporal variations made the agent feel less like a machine and more like an attentive entity, as shown in Figure 7. The silence successfully simulated conversational backchannels

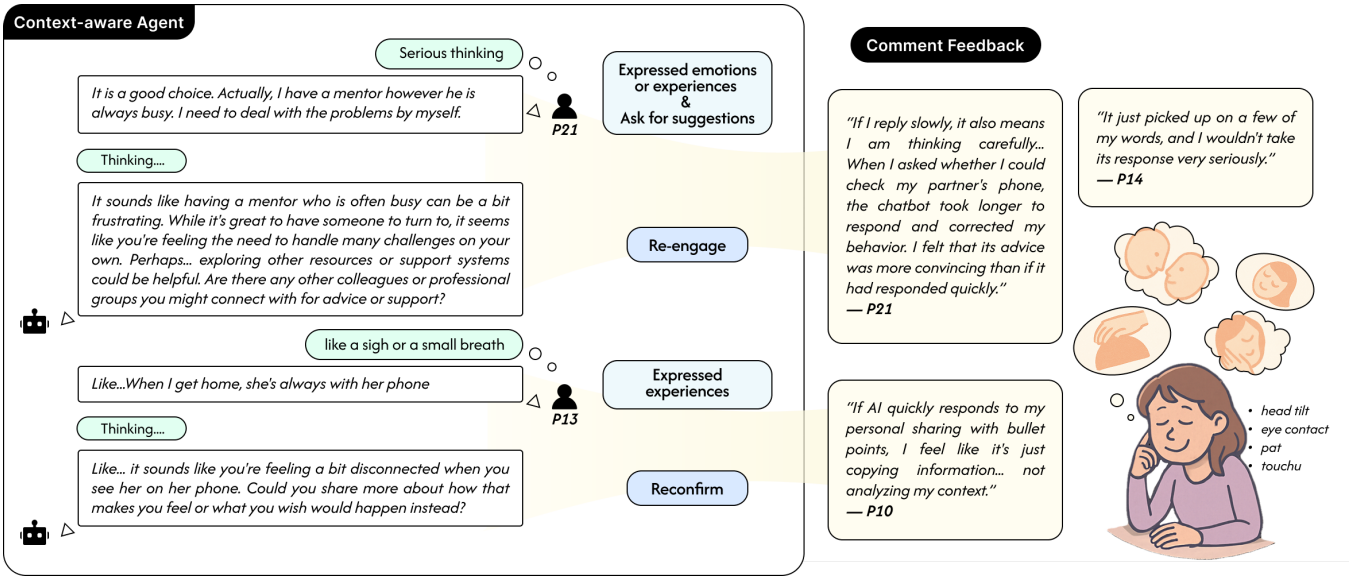


Figure 7: Pacing enhanced both persuasion and human-likeness by simulating a more deliberate interaction, boosting trust and listening quality, while creating a stronger sense of social presence. This made the interaction feel more natural and human-like.

and even physical presence. For example, participants described silence as being *"like a sigh or a small breath"* (P12) or *"like a faint breath in conversation... makes me feel the other person is there"* (P10). Remarkably, participants extended this feeling of presence to imagined physical embodiment. P8 described silence as resembling a friend's comforting *"pat"* or *"touch"*, while P10 imagined a *"head tilt"* or *"eye contact"*. In contrast, the static control agent's speed was often described as distancing and mechanical: *"If AI quickly responds to my personal sharing with bullet points, I feel like it's just copying information... not analyzing my context."* (P10)

5.4.3 Pacing Improved Overall Interaction Quality. The intervention significantly improved general metrics of interaction quality across both scenarios. As illustrated in Figure 5(e) and Figure 6(e), smoothness ratings were higher for the context-aware agent in both the career ($U = 205$, $p = 0.024$, $r = 0.3$) and relationship scenarios ($U = 180$, $p = 0.010$, $r = 0.36$). Similarly, perceived interactivity saw a strong improvement (Figure 5(f) and Figure 6(f)) in both career ($U = 141$, $p = 0.001$, $r = 0.48$) and relationship scenarios ($U = 162$, $p = 0.002$, $r = 0.43$). This suggests that dynamic pacing made the conversation flow feel more natural overall.

5.4.4 The Double-Edged Sword of Pacing: Violating Machine Efficiency Heuristics. Despite its widespread benefits, pacing was not universally positive. For a notable minority of participants, slowness violated the machine heuristic, namely the expectation that AI should be faster and more efficient than humans [102] (N11, P18). This led to frustration, particularly when participants felt the agent's response quality did not justify the wait. Participants expected high utility from AI, and slow responses could be perceived as poor performance or system malfunction. *"AI's unique advantage is that it could provide instant feedback... If you are an AI and still slow as such, it means you didn't even put in the effort."* (P11) Some participants interpreted the slower pacing as poor model capability, leading to a sense of disengagement (P8) and negative expectancy violation, as shown in Figure 8. *"If it is an intelligent robot, it should be able to respond quickly." (P23)* Their default expectation is that the robot will respond quickly and give lengthy answers. *"(slowing down) makes me feel irritated and wonder if the system is malfunctioning."* (P13) Even though our status bar indicates that the slower pacing means the chatbot "thinking" or "reflecting," some participants still automatically assumed it was merely an excuse for loading issues (P23). These negative interpretations suggest that while slower pacing can convey listening and connection, it may also evoke users' concerns of inefficiency, asynchronicity, and uncertainty.

In summary, the results demonstrate that context-aware pacing significantly enhances users' interaction quality and experience with CAs, with consistent improvements across both supportive scenarios in perceived human-likeness, smoothness, and interactivity. Affective trust and perceived listening quality were also significantly enhanced, although this effect was primarily concentrated in the goal-oriented career-support context. This positive perception was driven by interpreting pacing as thoughtful deliberation. However, this effect was in tension with some users' expectations of machine efficiency, creating a design tension between perceived empathy and perceived performance.

5.5 The Impact of Context-Aware Pacing on Users' Text-based Expression (RQ2)

To answer RQ2, we analyzed how context-aware pacing influenced participants' texting behaviors. We combined qualitative feedback on user experience with quantitative analysis of interaction logs to measure changes in self-disclosure and engagement.

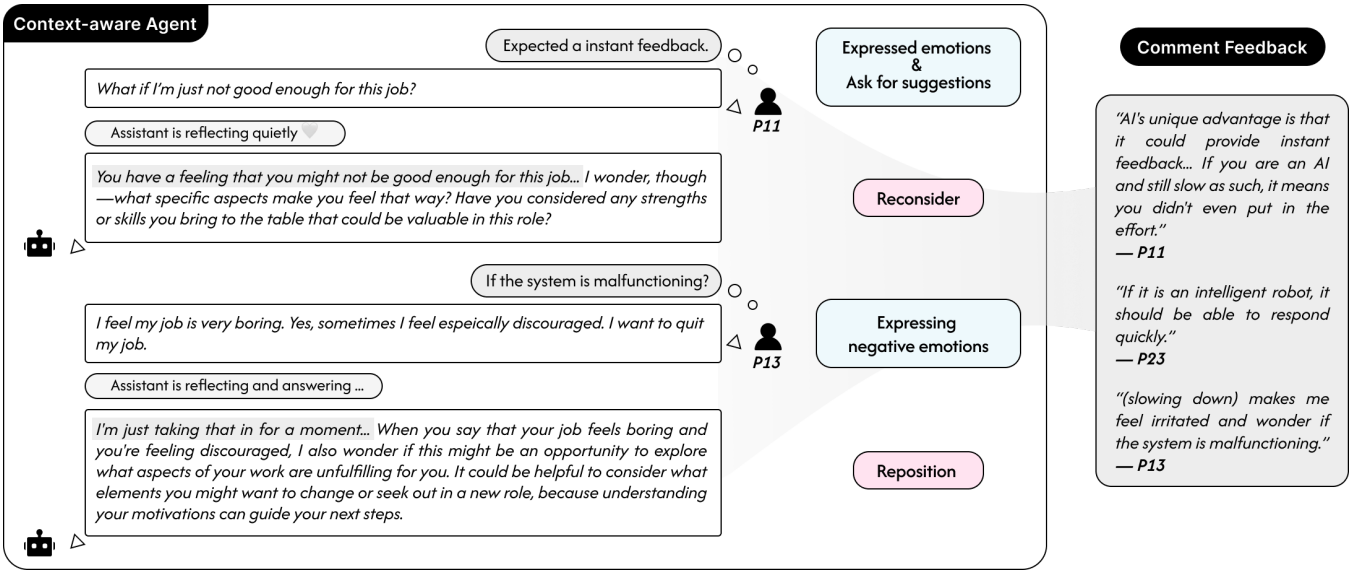


Figure 8: Some participants expressed that they would prefer a quick response over slow-paced emotional support. Otherwise, it feels more like a malfunction of the AI than an empathetic strategy.

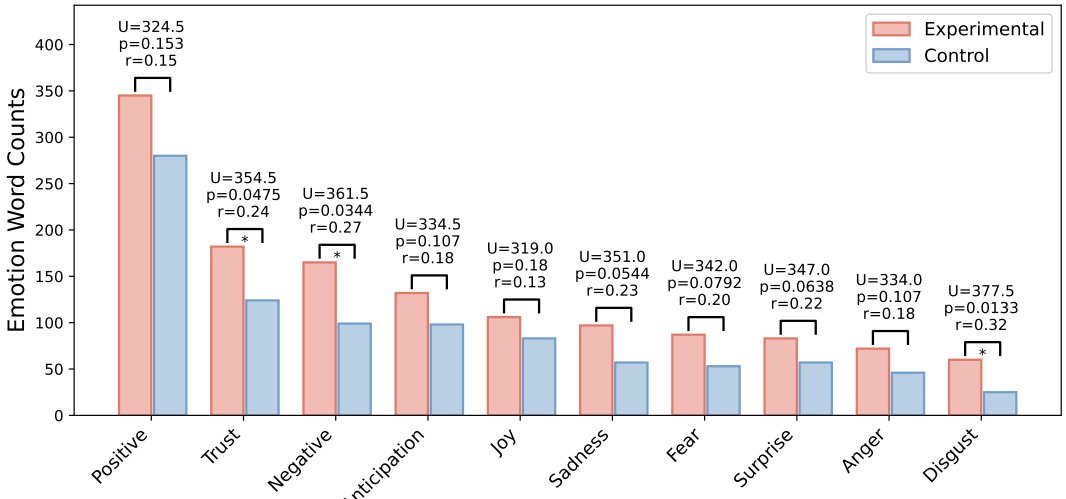


Figure 9: The emotion word count distribution across different groups. U is the test statistic. p is the p-value. r is the effect size.

5.5.1 Context-Aware Pacing Led to Deeper Self-Disclosure. There was a significant increase in participants' self-disclosure. Quantitative analysis of chat logs showed that participants interacting with the context-aware agent used significantly more total emotional words ($U = 359.0$, $p = 0.04$, $r = 0.301$) compared to the control group. This indicates deeper affective self-disclosure, as users shared more about their internal feeling states [68, 100]. Beyond the total emotion words, we analyzed each type of emotion expressed. As shown in Figure 9, the participants' inputs consistently contain more emotion words across different emotions in the experimental group than those in the control group with (marginal) significance (at 0.1 level) in the trust, negative, sadness, and disgust dimensions. This suggests the agent provided a safe space for users to express a wider range of emotions, including negative feelings, which can be forms of valuable emotional release. This finding was complemented by a significant increase in first-person pronouns (e.g., "I," "me," "my") ($U = 382.0$, $p = .001$, $r = 0.384$). This suggests users adopted a higher degree of self-focus and personal ownership, more frequently referencing their own experiences, thoughts, and feelings [16, 93].

Furthermore, we observed a trend toward longer user input length (in words) in the experimental group ($M = 19.91$ words) versus the control group ($M = 18.15$ words) with marginal significance at the 0.1 level ($p = 0.06$, $r = 0.268$). Participants felt that the context-aware pacing CA provided a safe space that encouraged

deeper reflection. P9 stated that the conversation *"resembled psychological counseling,"* which was *"inspiring and guiding me to say more."* P2 noted that the agent seemed to *"consider whether I was willing to continue sharing..."* rather than issuing commands like other systems. This perceived attentiveness created an environment conducive to vulnerability and personal expression.

5.5.2 Pacing Increased Engagement by Improving Interaction Flow. Context-aware pacing also led to significantly higher user engagement, measured by the total conversation turns ($U = 520.0, p < 0.01$, $r = 0.884$). Within each group, there was no significant difference in the total conversation turns between the two scenarios (Control: $p = 0.656$, Experimental: $p = 0.634$). For the experimental group, the number of pausing strategies is shown in Table 4. We conducted a Chi-Squared Test, and the result showed that the scenarios were not significantly correlated with the pausing strategies ($\chi^2 = 8.40$, $df = 7$, $p = 0.30$). This confirms that behavioral differences are likely attributable to user reactions rather than variations in agent behavior across scenarios.

Participants explained that the slower, more deliberate pacing helped them better process the AI's replies, creating a more immersive interaction and reducing information overload (P10, N9). P10 contrasted this with typical rapid-fire CAs: *"It prevents me from scrolling back to catch up with the popping messages, presenting the information in a more layered way."* Another participant highlighted how pacing supported parallel processing, similar to human dialogue: *"I have more time to carefully read line by line. While it's still outputting, I can start typing my response, which makes the interaction feel more like a real dialogue."* (P8)

6 Discussion

6.1 Context-Aware Pacing for Active Listening in CA

To understand conversational pacing in human-CA interaction, we implemented and evaluated a system featuring context-aware pacing. We designed the agent to operationalize different pacing strategies ranging from *Immediate Response* for task-oriented queries to deliberate silence (e.g., *Holding Space*) for emotional vulnerability, assessing effects on perceived interaction quality and user experience.

Our study found that context-aware pacing improved perceived listening quality, aligning with previous interpersonal communication works where appropriate response pacing is suggested to enhance responsiveness and high-quality listening [106]. Unlike prior HCI studies focusing on verbal content [21, 46, 56, 60, 66, 120], our results suggested that compared to a static fast response, CAs' adaptively slow response or even silence can also lead to higher-quality listening perception, even when the content of the response is similar.

Crucially, positive user perception relied on the CA's high accuracy (86.2%) in selecting suitable strategies. This reliability ensured appropriate pacing for distinct contexts like factual queries (`Immediate Response`) versus intensive distress (`Holding Space`), preventing frustration from mismatched pacing (e.g., a long, awkward silence during a simple information exchange). While precision was slightly lower for subtle emotional nuances, these were largely "benign misclassifications" (e.g., substituting `Reconsider` for `Resonate`) that maintained a thoughtful tone. By avoiding the most disruptive errors, the CA preserved the overall perception of empathy and safety.

Consistent with self-disclosure research [21], context-aware pacing facilitated deeper disclosure in both scenarios. Our CA led users to provide longer responses with more emotional words and first-person pronouns. As an active listening strategy, context-aware pacing conveys a non-judgmental attitude, avoiding conversational dominance and prompting user expression [46]. Thus, in addition to delivering high-quality and warm verbal content, context-aware pacing itself can serve as a meaningful and constructive cue that fosters more open and engaged interactions.

Our analysis revealed that the CA's context-aware pacing successfully simulated the pacing in human conversations, making the dialogue more human-like. People linked CAs' slow pacing to human behaviors such as analyzing, deliberating, or comforting behaviors. This aligns with previous findings that slow response pacing could enhance social presence [30], reducing machine heuristic stimulated by static and mechanical fast response [5, 102]. While our current implementation utilizes a mechanistic "sort-and-hold" delay, our results indicate that users do not perceive this as empty latency. Instead, they anthropomorphized the pacing as a deliberate, human-like silence, interpreting it as "serious thinking" or "breathing". This suggests that perception of silence can be decoupled from the mechanism of generation, demonstrating the promise of using pacing strategies to regulate social presence in CAs.

Furthermore, contrary to research treating delays as problematic, our survey shows context-aware pacing enhanced smoothness and interactivity. Vocal turn-taking studies suggest natural responses often begin before the previous turn finishes [36]; our data reflects similar patterns. P8 described the slow pacing: *"When it replies slowly, it's still outputting on its side while I can type in response at the same time, which makes the conversation feel more interactive."* These indicate that naturally overlapping dialogue flow also occurs in slower-paced text-based interactions with LLM-backed CAs. Context-aware pacing with appropriate silence is not a breakdown or meaningless null signal in conversations, but can facilitate a more fluent and consistent interaction experience [72].

Our study shows that designing context-aware pacing requires a delicate balance between emotional support effects and users' impatience. This tension can be understood through the lens of Expectancy Violation Theory (EVT) [19]. EVT posits that violations of predictive expectations trigger an evaluation process. In our study, while most users interpreted the delay as a positive signal of thoughtfulness, a subset of users holding a "machine heuristic" maintained a cognitive schema of robots as omniscient and capable of delivering fast, stable replies [28]. This highlights the need for future work to analyze conversational contexts and users' expectations for response speed, enabling CAs to better align pacing with users' individual preferences. Two potential design paths could be explored. One path, aligning with user agency, would be to provide explicit controls, such as a slider to adjust the desired pacing style (e.g., ranging from "Efficient" to "Reflective"), allowing users to set their preferred pacing style. Another path involves developing implicit adaptation models, where the CA automatically learns and attunes to a user's pacing preferences over time. Both approaches

could help provide a sense of being heard and understood while mitigating the disruptive perceptions caused by a one-size-fits-all pacing strategy.

6.2 Design Implications for Embodying Pacing in CAs

Our findings offer a new lens for designing empathic human-CA interactions, shifting focus from *what* an agent says to *how* and *when* it says it. We propose several design implications based on our operationalization of context-aware pacing and findings.

6.2.1 Beyond Context: Learning Individual Pacing Personas. Our framework of five strategies (*Reflective Silence*, *Facilitative Silence*, *Empathic Silence*, *Holding Space*, and *Immediate Response*) provides a functional vocabulary for implementation. For example, rather than a generic "thinking" delay, designers can implement *Reflective Silence* specifically to signal active processing of user input, or *Empathic Silence* to create space for emotional resonance after a vulnerable disclosure. This approach moves beyond a simple fast or slow dichotomy to enable more nuanced, context-aware interaction timing.

However, our findings reveal a critical next step beyond situational context-awareness: personalization. We observed various user expectations and preferences for pacing. Some participants interpreted slow pacing as empathy, while others perceived it as inefficiency. This suggests that future work could not only adapt to the conversation's context but also to the user's individual communication style.

For example, future work could explore models where the agent learns a user's "pacing persona." Existing research has shown that user characteristics, such as personality traits, do impact their attitudes toward AI [51, 99]. While future systems would not (and likely should not) attempt to model a user's full psychological profile, they could learn to adapt to observable communication styles and preferences that may be associated with these traits.

Furthermore, just as users have different preferences, supportive "listener" styles are not monolithic [128]. Our framework provides a foundation for one effective, therapeutically-inspired pacing model. Future systems, however, could be designed with multiple, distinct "pacing personas" (e.g., a direct, solution-focused style vs. our more reflective, therapeutic style), allowing a user to select the interaction model they find most supportive. By observing how a user reacts to different pacing strategies over time (e.g., do they frequently interrupt, or do they show deeper reflection after providing space?), an advanced agent could infer which "pacing personas" best align with their communication style and current needs, thus deepening personalization.

6.2.2 Integrate Pacing with High-Quality Response. Our "double-edged sword" finding reveals a critical design contingency: pacing acts as a multiplier for user expectations. A deliberate silence signals thoughtful deliberation; if the subsequent response is generic or low-quality, the user's trust can be more significantly damaged than by an immediate low-quality response. This aligns with prior research [83] finding that when the advice was inaccurate, delayed responses led to lower trust and reliance. Our qualitative results offer a potential explanation for this finding: pacing functions as a "promissory note" to the user, implying that the upcoming content will be valuable and tailored. However, if the content fails to deliver on this promise, it creates a significant expectancy violation, leading to heightened frustration and a feeling that the agent's empathic signals are inauthentic.

Therefore, pacing strategies should not be treated as a simple add-on feature but must be deeply integrated with the generative model's confidence in its own output. This suggests a more sophisticated implementation where pacing strategies are dynamically selected based on risk assessment. For instance, high-effort pacing strategies (like long reflective silence and slower speed) should be reserved for turns where the system has high confidence in its ability to generate a relevant and insightful response. Conversely, when response quality is uncertain, a faster, more transactional pacing style may be safer, as it avoids setting expectations that cannot be met.

This "risk assessment" becomes paramount in high-stakes domains such as well-being chatbots, where trust and emotional safety are the foundation of the interaction. For instance, during a crisis moment (e.g., acute anxiety), a long, reflective silence could be misinterpreted as abandonment [76]. In this context, a visual cue that signals active processing (such as the status bar used in our design) can convey that the agent is still present and attending to their input, mitigating perceptions of abandonment [1]. Furthermore, if the CA delivers a generic or tone-deaf reply after a long silence, the resulting expectancy violation may be actively invalidating for a user in a sensitive emotional state, causing significant emotional harm. Therefore, for well-being agents, we suggest a more conservative approach: high-effort pacing should be reserved for high-confidence responses with transparency (e.g., a status bar showing "Reflecting..."). When system confidence is low, a safer alternative is to use a faster, simpler acknowledgment, combined with transparency (e.g., *"Let me take a moment to consider that..."*) to safely manage expectations.

6.2.3 Overcoming Interaction Inertia through Gradual Adaptation. Our findings (Section 5.4.4) show that some users' negative reactions to pacing stem from a strong interaction inertia [30]. This is tightly related to the Machine Heuristic [102], where impressions are conditioned by long-term AI experience, framing users to expect static, rapid pacing as granted. Any deviation from this learned norm violates their entrenched expectations. This raises a crucial question: *is their frustration a fundamental rejection of dynamic pacing, or is it a temporary adaptation cost due to the abrupt violation of a deeply learned script?* This suggests involving a dynamic process of user onboarding. A potential design implication is to introduce context-aware pacing gradually. For example, a CA could initially interact with a new user using a faster, more "machine-like" pace to align with their existing expectations. Over several conversations, as the user becomes accustomed to the agent, the system could progressively introduce more nuanced strategies as well as learn from the user's interaction patterns as discussed in Section 6.2.1.

Future research should investigate this gradual adaptation approach. A longitudinal study could test whether users who are slowly eased into a dynamic pacing model show higher long-term acceptance and satisfaction compared to those who experience it abruptly. This would help determine if the negative effects of

expectancy violation can be mitigated by "training" the user to appreciate a more context-aware conversational rhythm over time.

6.2.4 Balancing Routine Efficiency with Critical Affective Pacing. Our findings suggest that pacing design must be adaptive rather than uniform. Our formative findings reveal a functional asymmetry where informational strategies (`Resolve`, `Reconfirm`) are more frequently used by listeners, a pattern aligned with our user study results (Table 4). These frequent exchanges constitute the "hygiene factor" of the interaction; they must be efficient to establish baseline competence [7]. In contrast, affective strategies (`Holding`, `Resonate`) reside in the "long tail", which are statistically rare but contextually critical. These sparse moments represent high-stakes pivots of user vulnerability where errors are most damaging [7, 111]. However, standard AI optimization, which minimizes average error, risks under-performing in these sparse but critical regions [127]. Consequently, designers should not allocate resources solely proportional to frequency. Instead, the "long tail" demands disproportionate engineering effort to ensure safety and sensitivity [111], as these rare moments, such as the high-intensity breakthroughs observed in our formative cases, often define the ultimate success of the therapeutic alliance.

6.2.5 From Artificial Pacing to Real Cognition: The Active Listening Chain-of-Thought. Looking beyond the mechanistic implementation used in this study, our findings offer a blueprint for the next generation of reasoning models (e.g., Chain-of-Thought processing [118]). As models increasingly require actual computational time to perform complex reasoning, which is often visualized as "thinking" steps [58], our framework can be integrated to humanize this latency. By framing this processing time as context-aware silence in supportive scenarios, designers can transform a technical bottleneck into a relational asset, aligning the AI's actual "thinking" time with the user's expectation of human-like contemplation. Moreover, future systems could explicitly incorporate affective reflection into the model's reasoning chain itself to enable an *Active Listening Chain-of-Thought*, utilizing the silence to internally deliberate on the user's emotional state before generating a response, thereby shifting the pacing from a simulated effect to a functional component of empathetic cognition.

6.2.6 From Understanding to Design: Modeling Pacing as a State Transition System. Conversational contexts are inherently interrelated, leading to natural shifts in pacing. Our formative analysis (Section 3.1.5) reveals that pacing is not merely a reaction to the current input, but a function of the conversational state. This implies that future CAs should not treat pacing selection as an isolated classification task. Instead, pacing could be modeled as a State Transition System [122]. For instance, a CA detecting intense distress could enter a "Holding State." In this state, even if the user asks a minor factual question, the CA might prioritize maintaining a slower, supportive cadence rather than jarringly switching to a fast *Immediate Response*. By preserving these "Supportive Arcs", designers can prevent unnatural oscillations between fast and slow responses, ensuring the agent maintains a consistent and attentive persona throughout the interaction.

6.3 Pitfalls of Current CA Response Timing

Our research critiques two common pitfalls in contemporary CA design that stem from prioritizing efficiency and comprehensiveness.

- **The fallacy of uniform immediacy**. The prevailing design philosophy often treats any delay as system latency to be eliminated. Although contemporary generative models exhibit varied response delays based on computational complexity ("thinking" efforts), this technical latency is often decontextualized from the user's relational needs [87]. In other words, the system delays are due to complex queries, not the users' emotional space. While speed is crucial for purely task-oriented queries, our findings confirm that in supportive contexts where both advice and emotional comforting are needed, consistently immediate responses can be perceived as perfunctory, emotionally detached, and affectively untrustworthy. By responding instantly, the agent fails to perform the social function of "deliberation", signaling to the user that their input was merely processed rather than truly considered.
- **Conversational dominance through continuous output.** The current paradigm often encourages CAs to "talk too much" by providing lengthy, comprehensive answers as quickly as possible. This approach fails to recognize that silence and pauses are essential for user reflection and turn-taking. When an agent floods the interaction with continuous text, it creates a high cognitive load for the user and shifts the dynamic from a collaborative dialogue to a passive information dump. This disempowers the user and discourages self-disclosure. In contrast, our findings show that strategies like *Holding Space* and *Facilitative Silence*, where the agent strategically "shuts up and listens", empower users by giving them room to process emotions and elaborate on their thoughts, leading to deeper engagement.

6.4 On the Generalizability of Our Context-Aware Pacing Design and Results

The findings from this study have several implications for generalizability, which must be considered alongside the study's specific limitations (discussed in Section 6.6).

6.4.1 Generalizability Across Interaction Modalities. While our study's principle that response timing is a deliberate communicative act is highly relevant for voice and multimodal CAs, the implementation must be fundamentally re-conceptualized. Voice pacing is inextricably linked with prosody, intonation, and vocal fillers (e.g., *"hmm"*, *"uh-huh"*). In text-based CAs, a 3-second silence with a visual indicator can signal deliberation, while the same silence in a voice call is more prone to be interpreted as a connection error or social awkwardness.

Current voice agents often focus on minimizing or masking latency, for instance, by using streaming processing or pre-composed fillers (*"Let me check that for you"*) [55, 123]. Advanced neural text-to-speech systems also control prosody and micro-pauses to make speech sound natural [80, 101], but this is often in service of linguistic rather than relational goals. Our work directly complements

these technical approaches by reframing silence and timing as a design resource to be orchestrated intentionally. For example, our finding on using pauses to signal thoughtfulness could inspire an analogous, modality-specific voice strategy: not a problematic 3-second "dead air" silence, but a carefully timed combination of a filler ("*Well...*"), prosodic variation, and a short, deliberate pause. Future work can explore how to orchestrate these voice-specific elements to implement the functional, relational pacing strategies outlined in our framework.

6.4.2 Generalizability Across Different Scenarios. As we found in our results (Section 5.4.1), pacing strategies demonstrated the strongest positive impact in hybrid tasks like career support, which require blending informational advice with emotional comfort. The observed variance between scenarios indicates that even within seemingly similar supportive tasks, user needs may differ. For example, users in a relationship context may prioritize the content of the validation, while users in a career context place greater value on the perceived thought process behind the advice. Therefore, a critical generalizable takeaway for design is to move beyond broad task categories. Systems must adapt by identifying the immediate subcontext and the user's shifting goals between information-seeking and emotional validation.

6.4.3 Generalizability Across User Populations and Experiences. While the core principle of using pacing to convey social cues in CAs is likely generalizable, we caution readers against generalizing our specific results or five-strategy framework to all populations. Our findings suggest that users' backgrounds and traits are critical factors. For example, a user's AI experience (Section 5.4.4) influences perception; those conditioned to expect speed may be less receptive to deliberate pauses [30, 102]. This variance suggests that for context-aware pacing to be effective broadly, systems will likely require robust personalization (Section 6.2.1) or adaptive onboarding strategies (Section 6.2.3). Moreover, the generalizability of our findings is further constrained by factors like individual differences [99] (discussed in Section 6.6.2) and cultural background [65, 89] (discussed in Section 6.6.4). Therefore, the primary generalizable takeaway is not our specific implementation, but the principle that designers should account for these user-level variables when implementing pacing.

6.5 Broader Implications and Ethical Considerations

Our work points toward significant societal benefits, such as low-cost, scalable well-being support [44, 45]. By being perceived as more "caring" and "deliberate", such agents could improve access to mental health tools [107]. They may also serve as a non-judgmental gateway to care for individuals who fear the stigma of human therapy, offering a "safe space" for initial self-disclosure [23]. However, an agent that effectively promotes trust and self-disclosure also creates profound ethical risks. First, as discussed in Section 6.2.2, low-quality content following a deliberate silence can cause emotional harm and active invalidation, necessitating robust content safeguards. Potential solutions range from proactive strategies, such as linking pacing to a "risk assessment" or using emotion detection to avoid distressing silence [31], to adaptive safeguards, such as using pre-defined templates in moments of uncertainty, implementing real-time user feedback loops, or escalating to a hybrid human-AI model [103, 115]. Second, techniques that build trust can be exploited, creating a dual risk of data privacy violations and user manipulation. Privacy can be protected through technical safeguards like on-device processing and user-controlled data deletion [126]. Manipulation can be mitigated with policy safeguards like strict content guardrails and continuous auditing [71]. These solutions could be paired with transparency and user consent mechanisms to protect the user's vulnerabilities [126].

6.6 Limitations and Future Work

While our findings demonstrate the benefits of context-aware pacing, we identify four key areas for future research based on the study's limitations and nuanced results.

6.6.1 Limitations in Controlling for Generative Content. A limitation of our study is the inherent variability of generative content. Because participant conversations were free-flowing and agent responses were dynamically generated, it was infeasible to strictly control the turn-by-turn text. While the semantic content analysis (Section 5.2) confirmed strong semantic equivalence, this analysis cannot fully discount subtle, unmeasured differences. Furthermore, as we discussed in Section 6.2.2, response quality can influence the user's perception of pacing. In this study, our goal was to explore the effect of pacing. Therefore, to control the response quality, we used a single, high-quality generative model (GPT-4o) for both conditions. While this isolated the pacing effect from model quality, it makes it unclear how our framework would perform with less (or more) capable models. In the future, this interplay could be explicitly tested, for example, by evaluating how users perceive the same pacing strategies when delivered by models with different generative qualities.

6.6.2 The Role of Individual Differences and Interaction Inertia. Our results highlight significant variance in user preferences, underscoring the influence of different user backgrounds and experiences. This variance potentially explains why effects differed between scenarios. Some participants valued concrete, actionable advice as the primary form of support, while others prioritized emotional validation. For users focused on efficiency, reflective pauses sometimes caused frustration (*"I forgot what I was going to say next,"* P2), whereas for others, they created a necessary space for thought. These varying preferences may stem from two sources. First, as prior work suggests, personality traits can significantly impact attitudes towards AI [99]. For instance, people with extroversion personality traits have a higher tendency to anthropomorphize robots [51]. Second, as we discussed in Section 6.2.3, users' ingrained expectations and interaction inertia [30, 102] influence their perception. Our study did not systematically investigate the root cause of this variance. This suggests that a one-size-fits-all pacing model is insufficient. Future systems should strive for personalization, adapting pacing strategies based not only on users' individual preferences but also on their ingrained expectations and communication style.

6.6.3 Ambiguity of Silence in Text-Only, Minimal-Identity Interactions. This study was confined to text-based interaction, which

limits the richness of communication cues. As P12 noted, silence is easier to interpret when one can simultaneously *"see the other person's facial expressions or sense their tone of voice."* Without multimodal cues, text-based silence can be ambiguous and misinterpreted as system lag rather than thoughtful deliberation. This ambiguity is likely moderated by the agent's perceived social identity, a factor not explicitly controlled in our study. A stronger, more human-like persona could provide the social context that helps users interpret ambiguous pacing cues more charitably [62]. Future work should therefore investigate the interplay between pacing and identity, examining how a well-defined persona of CAs might scaffold user interpretation of timing in text-based chat, in parallel with exploring richer modalities like voice and video.

6.6.4 Cultural Specificity of Pacing Norms and Limited Demographic Diversity. Finally, the interpretation of silence and conversational timing is deeply rooted in cultural norms. The strategies and pause durations effective in this study reflect communication patterns specific to our participant cohort, which consisted entirely of individuals from an East Asian cultural background, where silence can often signify respect, thoughtfulness, or emotional regulation [89]. All participants were proficient in English and the study was conducted entirely in English; while this mitigated potential language confounds, it does not remove the underlying cultural lens. For instance, pause durations that signify respect in some cultures may indicate indifference, disagreement, or awkwardness in others [65]. These findings cannot be generalized globally. Therefore, cross-cultural validation is essential before deploying context-aware pacing models in diverse user populations. Future research should investigate how these pacing strategies can be adapted to meet the diverse conversational expectations of users from various cultural backgrounds.

6.6.5 Interdependence and Isolation of Pacing Strategies. Our study evaluated the holistic effect of the context-aware pacing model, which prevents us from isolating the specific impact of each of our five strategies (e.g., *Reflective Silence* vs. *Holding Space*). The observed benefits are likely the result of a synergistic effect, but some strategies may be more impactful than others. Furthermore, our analysis of sequential pacing patterns suggests that these strategies are not merely independent components but are sequentially interdependent. A limitation of our current system is that it selects strategies based on the classification of user's current input. Although our Conversational Memory Module retains dialogue history to ensure semantic coherence, it does not explicitly model pacing state transitions. Future work should use more controlled methods, such as A/B testing and sequential modeling, to distinguish the contribution of individual strategies from the transition logic that connects them.

6.6.6 Limitations of Formative Study Data. Our strategies were derived from a qualitative analysis of ten exemplary active listening videos, which presents limitations regarding generalizability. First, this small sample may not comprehensively capture all nuanced pacing strategies employed in human-human active listening. Second, our dataset, while covering diverse topics, did not control for the diversity of counselor styles and included multiple videos from a single creator. Different therapeutic approaches or individual counselor habits may feature different pacing norms, which remain underexplored in our framework. Finally, these videos were sourced exclusively from a therapeutic counseling domain. While this context was ideal for identifying strategies for the supportive conversations we tested, the generalizability of the resulting affective strategies to other domains (e.g., purely task-oriented technical support) should be further validated. Future work should therefore investigate pacing strategies across a wider variety of contexts to develop a more broadly applicable taxonomy of conversational pacing.

6.6.7 Limitation of Predefined Scenarios. Our reliance on predefined role-play scenarios introduces a limitation regarding ecological validity. While the two scenarios represent common life stressors and elicited positive effects, we acknowledge that varying levels of personal resonance may have influenced individual engagement. However, our randomized between-subjects design served to distribute these individual differences equally across conditions, ensuring that such variations remained random rather than systematic. Consequently, the observed effects are attributable to the pacing intervention rather than contextual confounds. Future work should validate these findings in a more naturalistic setting, for instance, through a longitudinal study where users discuss their own real-life problems to confirm the effects of pacing in real-world contexts.

7 Conclusion

This paper underscores the often-overlooked importance of temporal cues in designing CAs for active listening. Based on a qualitative analysis of human conversations, we derived a framework of five context-aware pacing strategies: *Reflective Silence*, *Facilitative Silence*, *Empathic Silence*, *Holding Space*, and *Immediate Response*. We implemented these into a prototype and conducted a between-subjects study ($N = 50$) comparing it against a static baseline across career and relationship scenarios. Results demonstrate that context-aware pacing significantly enhances perceived human-likeness, smoothness, and interactivity, while fostering deeper self-disclosure and engagement (evidenced by increased emotional language, first-person pronouns, and conversation turns). This work provides empirical evidence encouraging designers to move beyond optimizing response content and strategically incorporate the communicative power of silence and pacing. We hope this research serves as an exploratory step toward promoting more human-centric and emotionally attuned CAs by encouraging a design paradigm that values strategic listening as highly as articulate responding.

Acknowledgments

This work is supported by City University of Hong Kong Teaching Development Grant (6000901), City University of Hong Kong Booster Fund (7030021), and Hong Kong Research Grants Council Theme-based Research Scheme (T45-205/21-N). We sincerely thank all anonymous reviewers for their insightful comments and constructive feedback that helped improve this paper.

References

[1] Martin Adam, Michael Wessel, and Alexander Benlian. 2020. AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets* 31 (2020), 427 – 445. doi:10.1007/s12525-020-00414-7

[2] Omolola A Adeoye-Olatunde and Nicole L Olenik. 2021. Research and scholarly methods: Semi-structured interviews. *Journal of the american college of clinical pharmacy* 4, 10 (2021), 1358–1367.

[3] Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston.

[4] Naeimeh Anzabi and Hiroyuki Umemuro. 2023. Effect of different listening behaviors of social robots on perceived trust in human-robot interactions. *International Journal of Social Robotics* 15, 6 (2023), 931–951.

[5] Jana Appel, Astrid von der Pütten, Nicole C Krämer, and Jonathan Gratch. 2012. Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. *Advances in Human-Computer Interaction* 2012, 1 (2012), 324694.

[6] Bagus Tris Atmaja and Masato Akagi. 2020. The effect of silence feature in dimensional speech emotion recognition. *arXiv preprint arXiv:2003.01277* (2020).

[7] Emmanuel Ayedoun, Yuki Hayashi, and Kazuhisa Seta. 2018. Adding Communicative and Affective Strategies to an Embodied Conversational Agent to Enhance Second Language Learners' Willingness to Communicate. *International Journal of Artificial Intelligence in Education* 29 (2018), 29 – 57. doi:10.1007/s40593-018-0171-6

[8] Anthony L Back, Susan M Bauer-Wu, Cynda H Rushton, and Joan Halifax. 2009. Compassionate silence in the patient–clinician encounter: a contemplative approach. *Journal of palliative medicine* 12, 12 (2009), 1113–1117.

[9] A. Barak and Orit Gluck-Ofri. 2007. Degree and Reciprocity of Self-Disclosure in Online Forums. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society* 10 3 (2007), 407–17. doi:10.1089/cpb.2006.9938

[10] Josef Bartels, Rachel Rodenbach, Katherine Ciesinski, Robert Gramling, Kevin Fiscella, and Ronald Epstein. 2016. Eloquent silences: A musical and lexical analysis of conversation between oncologists and their patients. *Patient education and counseling* 99, 10 (2016), 1584–1594.

[11] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[12] Ananya Bhattacharjee, Pan Chen, Abhijoy Mandal, Anne Hsu, Katie O'Leary, Alex Mariakakis, Joseph Jay Williams, et al. 2024. Exploring user perspectives on brief reflective questioning activities for stress management: Mixed methods study. *JMIR Formative Research* 8, 1 (2024), e47360.

[13] Graham Bodie, Andrea Vickery, Kaitlin Cannava, and Susanne Jones. 2015. The Role of "Active Listening" in Informal Helping Conversations: Impact on Perceptions of Listener Helpfulness, Sensitivity, and Supportiveness and Discloser Emotional Improvement. *Western Journal of Communication* 79 (01 2015), 151–173. doi:10.1080/10570314.2014.943429

[14] Graham D Bodie, Kellie St. Cyr, Michelle Pence, Michael Rold, and James Honeycutt. 2012. Listening competence in initial interactions I: Distinguishing between what listening is and what listeners do. *International Journal of Listening* 26, 1 (2012), 1–28.

[15] Halim-Antoine Boukaram, Micheline Ziadee, and Majd F Sakr. 2021. Mitigating the Effects of Delayed Virtual Agent Response Time Using Conversational Fillers. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (Virtual Event, Japan) *(HAI '21)*. Association for Computing Machinery, New York, NY, USA, 130–138. doi:10.1145/3472307.3484181

[16] T. Brockmeyer, Johannes Zimmermann, D. Kulessa, M. Hautzinger, H. Bents, H. Friederich, W. Herzog, and M. Backenstrass. 2015. Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Frontiers in Psychology* 6 (2015). doi:10.3389/fpsyg.2015.01564

[17] Thomas J Bruneau. 1973. Communicative silences: Forms and functions. *Journal of communication* 23, 1 (1973), 17–46.

[18] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[19] Judee K Burgoon. 1978. A communication model of personal space violations: Explication and an initial test. *Human communication research* 4, 2 (1978), 129–142.

[20] Joseph N Cappella. 1980. Talk and silence sequences in informal conversations II. *Human Communication Research* 6, 2 (1980), 130–145.

[21] Eugene Cho, Nasim Motalebi, S. Shyam Sundar, and Saeed Abdullah. 2022. Alexa as an Active Listener: How Backchanneling Can Elicit Self-Disclosure and Promote User Experience. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 273 (Nov. 2022), 23 pages. doi:10.1145/3555164

[22] Nicholas Christenfeld. 1995. Does it hurt to say um? *Journal of Nonverbal Behavior* 19, 3 (1995), 171–186.

[23] Liz L. Chung and Jeannie Kang. 2023. "I'm Hurt Too": The Effect of a Chatbot's Reciprocal Self-Disclosures on Users' Painful Experiences. *Archives of Design Research* (2023). doi:10.15187/adr.2023.11.36.4.67

[24] Hanne K Collins. 2022. When listening is spoken. *Current Opinion in Psychology* 47 (2022), 101402.

[25] Valerian J Derlaga and John H Berg. 1987. *Self-disclosure: Theory, research, and therapy.* Springer Science & Business Media.

[26] Valerian J Derlega, Barbara A Winstead, Alicia Mathews, and Abby L Braitman. 2008. Why does someone reveal highly personal information? Attributions for and against self-disclosure in close relationships. *Communication Research Reports* 25, 2 (2008), 115–130.

[27] Ronald M Epstein and Mary Catherine Beach. 2023. "I don't need your pills, I need your attention:" Steps toward deep listening in medical encounters. *Current Opinion in Psychology* 53 (2023), 101685.

[28] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What makes users trust a chatbot for customer service? An exploratory interview study. In *International conference on internet science*. Springer, 194–208.

[29] Phillip Glenn. 2024. " SO YOU'RE TELLING ME...": PARAPHRASING (FORMULATING), AFFECTIVE STANCE, AND ACTIVE LISTENING. *International Journal of Listening* 38, 1 (2024), 28–40.

[30] Ulrich Gnewuch, Stefan Morana, Marc TP Adam, and Alexander Maedche. 2022. Opposing effects of response time in human–chatbot interaction: The moderating role of prior experience. *Business & Information Systems Engineering* 64, 6 (2022), 773–791.

[31] Diego Gosmar, Elena Peretto, and Oita Coleman. 2024. Insight AI Risk Detection Model - Vulnerable People Emotional Situation Support. *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering* (2024). doi:10.1145/3661167.3661235

[32] Ángel López Gutiérrez and Juan José Arroyo Paniagua. 2024. An exploration of silence in communication. *European Public & Social Innovation Review* 9 (2024), 1–18.

[33] Cory Hallam and Gianluca Zanella. 2017. Online self-disclosure: The privacy paradox explained as a temporally discounted balance between concerns and rewards. *Computers in Human Behavior* 68 (2017), 217–227.

[34] David Hammer and Leema K Berland. 2014. Confusing claims for data: A critique of common practices for presenting qualitative research on learning. *Journal of the Learning Sciences* 23, 1 (2014), 37–46.

[35] Allyson Hauptman, Beau Schelble, Christopher Flathmann, Rohit Mallick, and Nathan McNeese. 2025. Ethical adaptation: exploring the use of adaptive autonomy in the design of ethical AI teammates in healthcare. *AI and Ethics* (2025), 1–18.

[36] Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38, 4 (2010), 555–568.

[37] Clara E Hill, Barbara J Thompson, and Nicholas Ladany. 2003. Therapist use of silence in therapy: A survey. *Journal of clinical psychology* 59, 4 (2003), 513–524.

[38] Thomas Holtgraves and Tai-Lin Han. 2007. A procedure for studying online conversational processing using a chat bot. *Behavior research methods* 39, 1 (2007), 156–163.

[39] TM Holtgraves, Stephen J Ross, CR Weywadt, and Tai-Lin Han. 2007. Perceiving artificial social agents. *Computers in human behavior* 23, 5 (2007), 2163–2174.

[40] Guy Itzchakov, Harry T Reis, and Netta Weinstein. 2022. How to foster perceived partner responsiveness: High-quality listening is key. *Social and Personality Psychology Compass* 16, 1 (2022), e12648.

[41] Guy Itzchakov, Netta Weinstein, Dvori Saluk, and Moty Amar. 2023. Connection heals wounds: feeling listened to reduces speakers' loneliness following a social rejection disclosure. *Personality and Social Psychology Bulletin* 49, 8 (2023), 1273–1294.

[42] Yuin Jeong, Juho Lee, and Younah Kang. 2019. Exploring Effects of Conversational Fillers on User Perception of Conversational Agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3312913

[43] Shaojie Jiang, S. Vakulenko, and M. de Rijke. 2023. Weakly Supervised Turn-level Engagingness Evaluator for Dialogues. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (2023). doi:10.1145/3576840.3578319

[44] Zhihan Jiang, Handi Chen, Rui Zhou, Jing Deng, Xinchen Zhang, Running Zhao, Cong Xie, Yifang Wang, and Edith CH Ngai. 2023. Healthprism: a visual analytics system for exploring children's physical and mental health profiles with multimodal data. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2023), 1205–1215.

[45] Zhihan Jiang, Lin Lin, Xinchen Zhang, Jianduo Luan, Running Zhao, Longbiao Chen, James Lam, Ka-Man Yip, Hung-Kwan So, Wilfred HS Wong, et al. 2023. A data-driven context-aware health inference system for children during school closures. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous*

Technologies 7, 1 (2023), 1–26.

[46] Martin Johansson, Tatsuro Hori, Gabriel Skantze, Anja Höthker, and Joakim Gustafson. 2016. Making turn-taking decisions for an active listening robot for memory training. In *International Conference on Social Robotics*. Springer, 940–949.

[47] Devon Johnson and Kent Grayson. 2005. Cognitive and affective trust in service relationships. *Journal of Business research* 58, 4 (2005), 500–507.

[48] Amira E Joseph, Rajat N Moman, Ross A Barman, Donald J Kleppel, Nathan D Eberhart, Danielle J Gerberi, M Hassan Murad, and W Michael Hooten. 2022. Effects of slow deep breathing on acute clinical pain in adults: a systematic review and meta-analysis of randomized controlled trials. *Journal of Evidence-Based Integrative Medicine* 27 (2022), 2515690X221078006.

[49] Mahdi Kafaee, Aliakbar Kouchakzadeh, and Shahriar Gharibzadeh. 2024. Silence: an ignored concept in artificial intelligence. *AI & SOCIETY* 39, 1 (2024), 415–416.

[50] Minjeong Kang. 2018. A Study of Chatbot Personality based on the Purposes of Chatbot. *The Journal of the Korea Contents Association* 18, 5 (2018), 319–329.

[51] Alexandra D Kaplan, Tracy Sanders, and Peter A Hancock. 2019. The relationship between extroversion and the tendency to anthropomorphize robots: A Bayesian analysis. *Frontiers in Robotics and AI* 5 (2019), 426899.

[52] Avraham N Kluger and Guy Itzchakov. 2022. The power of listening at work. *Annual Review of Organizational Psychology and Organizational Behavior* 9, 1 (2022), 121–146.

[53] Harald Victor Knutson and Aslaug Kristiansen. 2015. Varieties of silence: Understanding different forms and functions of silence in a psychotherapeutic setting. *Contemporary Psychoanalysis* 51, 1 (2015), 1–30.

[54] Justin Kruger, Derrick Wirtz, Leaf Van Boven, and T William Altermatt. 2004. The effort heuristic. *Journal of Experimental Social Psychology* 40, 1 (2004), 91–98.

[55] Junyeong Kum and Myungho Lee. 2022. Can gestural filler reduce user-perceived latency in conversation with digital humans? *Applied Sciences* 12, 21 (2022), 10972.

[56] Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis (Eds.). Association for Computational Linguistics, Saarbrücken, Germany, 127–136. doi:10.18653/v1/W17-5516

[57] Carine Lallemand and Guillaume Gronier. 2012. Enhancing User eXperience during waiting time in HCI: contributions of cognitive psychology. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. Association for Computing Machinery, New York, NY, USA, 751–760. doi:10.1145/2317956.2318069

[58] Hui Min Lee, Davis Yadav, Sangwook Lee, Keerthana Govindarazan, Cheng Chen, and S. Shyam Sundar. 2025. While We Wait... How Users Perceive Waiting Times and Generation Cues during AI Image Generation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 602, 8 pages. doi:10.1145/3706599.3719725

[59] Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. 2022. Online Self-Disclosure, Social Support, and User Engagement During the COVID-19 Pandemic. *ACM Transactions on Social Computing* 6 (2022), 1 – 31. doi:10.1145/3617654

[60] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376175

[61] Heidi M Levitt. 2001. Sounds of silence in psychotherapy: The categorization of clients' pauses. *Psychotherapy Research* 11, 3 (2001), 295–309.

[62] Jingshu Li, Zicheng Zhu, Renwen Zhang, and Yi-Chieh Lee. 2025. Exploring the Effects of Chatbot Anthropomorphism and Human Empathy on Human Prosocial Behavior Toward Chatbots. *arXiv preprint arXiv:2506.20748* (2025).

[63] Jongmoon Lim and Wonil Hwang. 2025. Proactivity of Chatbots, Task Types and User's Characteristics When Interacting with Artificial Intelligence (AI) Chatbots. *International Journal of Human–Computer Interaction* (2025), 1–19.

[64] Jianhua Lin. 2002. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37, 1 (2002), 145–151.

[65] Wong Ngan Ling. 2003. Communicative functions and meanings of silence: An analysis of cross-cultural views. *Multicultural studies* 3 (2003), 125–146.

[66] Chao Liu, Mingyang Su, Yan Xiang, Yuru Huang, Yiqian Yang, Kang Zhang, and Mingming Fan. 2025. Toward Enabling Natural Conversation with Older Adults via the Design of LLM-Powered Voice Agents that Support Interruptions and Backchannels. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 163, 22 pages. doi:10.1145/3706598.3714228

[67] Chunhong Liu and Shulin Yu. 2022. Exploring master's students' emotions and emotion-regulation in supervisory feedback situations: A vignette-based study. *Assessment & Evaluation in Higher Education* 47, 7 (2022), 1101–1115.

[68] Run-Xiang Liu and Huan Liu. 2022. The Daily Rhythmic Changes of Undergraduate Students' Emotions: An Analysis Based on Tencent Tweets. *Frontiers in Psychology* 13 (2022). doi:10.3389/fpsyg.2022.785639

[69] Soledad López Gambino, Sina Zarrieß, and David Schlangen. 2019. Testing strategies for bridging time-to-content in spoken dialogue systems. In *9th International Workshop on Spoken Dialogue System Technology*. Springer, 103–109.

[70] Christoph Lutz and Aurelia Tamò-Larrieux. 2021. Do privacy concerns about social robots affect use intentions? Evidence from an experimental vignette study. *Frontiers in Robotics and AI* 8 (2021), 627958.

[71] Johanna Löchner, Per Carlbring, Björn W Schuller, John Torous, and Lasse B Sander. 2025. Digital interventions in mental health: An overview and future perspectives. *Internet Interventions* 40 (2025). doi:10.1016/j.invent.2025.100824

[72] Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, Ryan P. McMahan, and Joseph J. LaViola Jr. 2025. Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 49, 15 pages. doi:10.1145/3719160.3736636

[73] David McNaughton, Dawn Hamlin, John Mccarthy, Darlene Head-Reeves, and Mary Schreiner. 2008. Learning to Listen: Teaching an Active Listening Strategy to Preservice Education Professionals. *Topics in Early Childhood Special Education* 27 (02 2008), 223–231. doi:10.1177/0271121407311241

[74] 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.

[75] M.B. Miles, A.M. Huberman, and J. Saldana. 2014. *Qualitative Data Analysis*. SAGE Publications. https://books.google.com/books?id=3CNrUbTu6CsC

[76] Adam S. Miner, A. Milstein, S. Schueller, Roshini Hegde, C. Mangurian, and E. Linos. 2016. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA internal medicine* 176 5 (2016), 619–25. doi:10.1001/jamainternmed.2016.0400

[77] Gargi Nandy, Srishti Gupta, Farhad Mohammad Afzali, Eric Peeples, Betsy Pilon, and Chun-Hua Tsai. 2025. Balancing Health Information-Seeking through Retrieval-Augmented Generation-Based LLM Chatbot. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 249–254.

[78] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[79] Robin Neuhaus, Ronda Ringfort-Felner, Shadan Sadeghian, and Marc Hassenzahl. 2024. To mimic reality or to go beyond?"Superpowers" in virtual reality, the experience of augmentation and its consequences. *International Journal of Human-Computer Studies* 181 (2024), 103165.

[80] Giridhar Pamisetty, K. Sri, and Rama Murty. 2021. Prosody-TTS: An End-to-End Speech Synthesis System with Prosody Control. *Circuits, Systems, and Signal Processing* 42 (2021), 361–384. doi:10.1007/s00034-022-02126-z

[81] Shuyi Pan and Maartje M.A. de Graaf. 2025. Developing a Social Support Framework: Understanding the Reciprocity in Human-Chatbot Relationship. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 182, 13 pages. doi:10.1145/3706598.3713503

[82] Wenjing Pan, Bo Feng, and V. Skye Wingate. 2018. What You Say Is What You Get: How Self-Disclosure in Support Seeking Affects Language Use in Support Provision in Online Support Forums. *Journal of Language and Social Psychology* 37 (2018), 27 – 3. doi:10.1177/0261927x17706983

[83] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 102 (Nov. 2019), 15 pages. doi:10.1145/3359204

[84] Yoonvin Park, Doha Kim, and Hayeon Song. 2025. I Know You're Listening: Designing Visual Backchannels for Voice User Interfaces. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25 Companion)*. Association for Computing Machinery, New York, NY, USA, 56–59. doi:10.1145/3708557.3716343

[85] Mario Passalacqua, Robert Pellerin, Florian Magnani, Laurent Joblot, Frédéric Rosin, Esma Yahia, and Pierre-Majorique Léger. 2025. Safeguarding worker psychosocial well-being in the age of AI: The critical role of decision control. *International Journal of Human-Computer Studies* (2025), 103649.

[86] Jessica Perrie, Aminul Islam, Evangelos Milios, and Vlado Keselj. 2013. Using google n-grams to expand word-emotion association lexicon. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 137–148.

[87] Ronald Poppe, Khiet P Truong, and Dirk Heylen. 2011. Backchannels: Quantity, type and timing matters. In *International workshop on intelligent virtual agents*. Springer, 228–239.

[88] Megan Price, Leanne Hides, Wendell Cockshaw, Aleksandra A Staneva, and Stoyan R Stoyanov. 2016. Young love: Romantic concerns and associated mental health issues among adolescent help-seekers. *Behavioral Sciences* 6, 2 (2016), 9.

[89] Yuan Yuan Quan. 2015. Analysis of silence in intercultural communication. In *2015 International Conference on Economy, Management and Education Technology*. Atlantis Press, 155–159.

[90] Rosemary P Ramsey and Ravipreet S Sohi. 1997. Listening to your customers: The impact of perceived salesperson listening behavior on relationship outcomes. *Journal of the Academy of marketing Science* 25, 2 (1997), 127–137.

[91] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10, 10 (1996), 19–36.

[92] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992.

[93] Ann-Katrin Reuel, Sebastian Peralta, João Sedoc, Garrick Sherman, and Lyle Ungar. 2022. Measuring the Language of Self-Disclosure across Corpora. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 1035–1047. doi:10.18653/v1/2022.findings-acl.83

[94] Carl Ransom Rogers et al. 1959. *A theory of therapy, personality, and interpersonal relationships: As developed in the client-centered framework*. Vol. 3. McGraw-Hill New York.

[95] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[96] Célia Sampaio, Catarina Luzia de Carvalho, Maria do Céu Taveira, and Ana Daniela Silva. 2025. Career intervention program to promote academic adjustment and success in higher education. In *Frontiers in Education*, Vol. 10. Frontiers Media SA, 1519877.

[97] Matthias Schmidmaier, Jonathan Rupp, and Sven Mayer. 2025. Using a Secondary Channel to Display the Internal Empathic Resonance of LLM-Driven Agents for Mental Health Support. In *Proceedings of the 27th International Conference on Multimodal Interaction*. 294–304.

[98] Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 10–26.

[99] Jan-Philipp Stein, Tanja Messingschlager, Timo Gnambs, Fabian Hutmacher, and Markus Appel. 2024. Attitudes towards AI: measurement and associations with personality. *Scientific Reports* 14, 1 (2024), 2909.

[100] Selina Studer, Christina Nuhn, Cornelia Weise, and Maria Kleinstäuber. 2025. The impact of photovoice on the report of emotions in individuals with persistent physical symptoms: Results of an experimental trial. *Journal of psychosomatic research* 191 (2025), 112069. doi:10.1016/j.jpsychores.2025.112069

[101] Aolan Sun, Jianzong Wang, Ning Cheng, Huayi Peng, Zhen Zeng, and Jing Xiao. 2020. GraphTTS: Graph-to-Sequence Modelling in Neural Text-to-Speech. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), 6719–6723. doi:10.1109/icassp40776.2020.9053355

[102] S. Shyam Sundar and Jinyoung Kim. 2019. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. 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, 1–9. doi:10.1145/3290605.3300768

[103] R. Taher, Daniel Stahl, Sukhi Shergill, and Jenny Yiend. 2024. The Safety of Digital Mental Health Interventions: Findings and Recommendations From a Qualitative Study Exploring Users' Experiences, Concerns, and Suggestions. *JMIR Human Factors* 12 (2024). doi:10.2196/62974

[104] Jennifer Tam, Elizbeth Carter, Sara Kiesler, and Jessica Hodgins. 2012. Video increases the perception of naturalness during remote interactions with latency. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 2045–2050. doi:10.1145/2212776.2223750

[105] Emma M Templeton, Luke J Chang, Elizabeth A Reynolds, Marie D Cone LeBeaumont, and Thalia Wheatley. 2022. Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences* 119, 4 (2022), e2116915119.

[106] Emma M Templeton and Thalia Wheatley. 2023. Listening fast and slow. *Current Opinion in Psychology* 53 (2023), 101658.

[107] Anja Thieme, Maryann Hanratty, M. Lyons, Jorge E. Palacios, R. Marques, C. Morrison, and Gavin Doherty. 2022. Designing Human-centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment. *ACM Transactions on Computer-Human Interaction* 30 (2022), 1 – 50. doi:10.1145/3564752

[108] Andrew P Thomas, Derek Roger, and Peter Bull. 1983. A sequential analysis of informal dyadic conversation using Markov chains. *British Journal of Social Psychology* 22, 3 (1983), 177–188.

[109] Marijan Tustonja, Davorka Topić Stipić, Iko Skoko, Anđelka Čuljak, and Andrea Vegar. 2024. ACTIVE LISTENING -A MODEL OF EMPATHETIC COMMUNICATION IN THE HELPING PROFESSIONS. *Medicina Academica Integrativa* 1 (03 2024), 42–47. doi:10.47960/3029-3316.2024.1.1.42

[110] Christof van Nimwegen and Emiel van Rijn. 2022. Time for a change: Reducing perceived waiting time by making it more active. In *Proceedings of the 33rd European Conference on Cognitive Ergonomics* (Kaiserslautern, Germany) *(ECCE '22)*. Association for Computing Machinery, New York, NY, USA, Article 7, 5 pages. doi:10.1145/3552327.3552329

[111] Michelle M.E. van Pinxteren, M. Pluymaekers, and J. Lemmink. 2020. Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management* (2020). doi:10.1108/josm-06-2019-0175

[112] Daphne van Zandvoort, Marloes Vredenborg, and Marit Bentvelzen. 2025. Unhealthy Comparisons to Promote Healthy Behavior? Exploring the Impact of Social Comparison Strategies in Personal Informatics.. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 94, 21 pages. doi:10.1145/3706598.3713737

[113] Anu Venkatesh, Chandra Khatri, A. Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, R. Prasad, Ming Cheng, Behnam Hedayatnia, A. Metallinou, Rahul Goel, Shaohua Yang, and A. Raju. 2018. On Evaluating and Comparing Conversational Agents. *ArXiv* abs/1801.03625 (2018).

[114] Pia Von Terzi and Sarah Diefenbach. 2023. The Attendant Perspective: Present Others in Public Technology Interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 502, 18 pages. doi:10.1145/3544548.3581231

[115] Xi Wang, Yujia Zhou, and Guangyu Zhou. 2024. The Application and Ethical Implication of Generative AI in Mental Health: Systematic Review. *JMIR Mental Health* 12 (2024). doi:10.2196/70610

[116] Harry Weger Jr, Gina R Castle, and Melissa C Emmett. 2010. Active listening in peer interviews: The influence of message paraphrasing on perceptions of listening skill. *The Intl. Journal of Listening* 24, 1 (2010), 34–49.

[117] Harry Weger Jr, Gina Castle Bell, Elizabeth M Minei, and Melissa C Robinson. 2014. The relative effectiveness of active listening in initial interactions. *International Journal of Listening* 28, 1 (2014), 13–31.

[118] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[119] Noel Wigdor, Joachim de Greeff, Rosemarijn Looije, and Mark A. Neerincx. 2016. How to improve human-robot interaction with Conversational Fillers. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 219–224. doi:10.1109/ROMAN.2016.7745134

[120] Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376131

[121] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and RAY LC. 2022. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *Joint Proceedings of the IUI 2022 Workshops: APEx-UI, HAI-GEN, HEALTHI, HUMANIZE, TExSS, SOCIALIZE*. CEUR-WS Team, 56–65. https://scholars.cityu.edu.hk/en/publications/publication(d901f5a2-0600-422f-b588-db5a59871961).html

[122] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. IEEE* 101, 5 (2013), 1160–1179.

[123] Jiahui Yu, Chung-Cheng Chiu, Bo Li, Shuo yiin Chang, Tara N. Sainath, Yanzhang He, A. Narayanan, Wei Han, Anmol Gulati, Yonghui Wu, and Ruoming Pang. 2020. FastEmit: Low-Latency Streaming ASR with Sequence-Level Emission Regularization. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), 6004–6008. doi:10.1109/icassp39728.2021.9413803

[124] Yuhan Zeng, Yingxuan Shi, Xuehan Huang, Fiona Nah, and RAY LC. 2025. "Ronaldo's a poser!": How the Use of Generative AI Shapes Debates in Online Forums. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 18. doi:10.1145/3706598.3713829

[125] Andong Zhang and Pei-Luen Patrick Rau. 2023. Tools or peers? Impacts of anthropomorphism level and social role on emotional attachment and disclosure tendency towards intelligent agents. *Computers in Human Behavior* 138 (2023), 107415.

[126] Dongsong Zhang, Jaewan Lim, Lina Zhou, and A. Dahl. 2021. Breaking the Data Value-Privacy Paradox in Mobile Mental Health Systems Through User-Centered Privacy Protection: A Web-Based Survey Study. *JMIR Mental Health* 8 (2021). doi:10.2196/31633

[127] Xiaoge Zhang, F. Chan, Chao Yan, and I. Bose. 2022. Towards risk-aware artificial intelligence and machine learning systems: An overview. *Decis. Support Syst.* 159 (2022), 113800. doi:10.1016/j.dss.2022.113800

[128] Xinyao Zhang, Michael J. Tanana, L. Weitzman, Shrikanth S. Narayanan, David C. Atkins, and Zac E. Imel. 2022. You never know what you are going to get: Large-scale assessment of therapists' supportive counseling skill use. *Psychotherapy* (2022). doi:10.1037/pst0000460

[129] 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. doi:10.1145/3613904.3642385

[130] Zhengquan Zhang, Konstantinos Tsiakas, and Christina Schneegass. 2024. Explaining the Wait: How Justifying Chatbot Response Delays Impact User Trust. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 27, 16 pages. doi:10.1145/3640794.3665550

[131] Li Zhou, Jianfeng Gao, Di Li, and Harry Shum. 2018. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics* 46 (2018), 53–93. doi:10.1162/coli_a_00368

[132] Suifang Zhou, Latisha Besariani Hendra, Qinshi Zhang, Jussi Holopainen, and RAY LC. 2024. Eternagram: Probing Player Attitudes Towards Climate Change Using a ChatGPT-driven Text-based Adventure. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3613904.3642850

A Formative Analysis Details

A.1 Links to the Active Listening Cases and Pacing Strategy Distribution

We collected cases of active listening from YouTube[8] using the term "Active Listening Counseling." We selected videos that explicitly mentioned "active listening" in their title or description. The videos were further screened based on the following criteria: (1) content centered around real or simulated counseling dialogues; (2) duration of over 20 minutes, including complete conversational segments; (3) clear visibility of listener behavior, including both verbal and nonverbal strategies; and (4) English as the primary language. From this corpus, we selected 10 cases (ranging in length from 20 to 60 minutes each) covering diverse topics (e.g., exam anxiety, relationship issues, body image) to ensure the generalizability of our findings. The links to the 10 cases are as follows:

(1) Dating Anxiety: https://www.youtube.com/watch?v=Xj3q96mCfC8
(2) Downward Arrow Techinique: https://www.youtube.com/watch?v=Wx8F9uwQTnY
(3) Anxiety Related to School Performance: https://www.youtube.com/watch?v=KRDH89uP8wI
(4) Demonstration of first counselling session with a 19 year old girl: https://www.youtube.com/watch?v=Ssi7Rzvfc40
(5) Toxic Self-Talk: What It Is & How to Stop It: https://www.youtube.com/watch?v=SO5gGaiDJb0
(6) I and Thou - touching pain and anger: https://www.youtube.com/watch?v=3r-lsBhfzqY
(7) Permission to feel: https://www.youtube.com/watch?v=2rSNpLBAqj0
(8) How NOT to Assess Client Preferences: https://www.youtube.com/watch?v=GOk7mR5mFLE
(9) Psychodynamic Therapy: https://www.youtube.com/watch?v=Sy6KfsR4H8U
(10) Severe anxiety: https://www.youtube.com/watch?v=xPRhsP8dKDs

A detailed breakdown of strategy distribution for each individual video case is provided in Table 5.

A.2 Pacing Strategy Transition Pattern Analysis

To describe the transition patterns between the counselors' pacing strategies, we additionally conducted a transition probability analysis based on the coded sequence of strategy labels. Each of the ten videos was treated as an independent session. We analyzed transition patterns using a three-step process:

- Sequence Extraction: We identified all adjacent strategy transitions (e.g., `Resolve`→`Recognize`) within each video.
- Aggregation: These pairs were summed to create a dataset-wide transition count matrix.
- Normalization: We row-normalized the matrix to calculate the conditional probability (ranging from 0 to 1) of transitioning to a subsequent strategy given a current state.

[8] https://www.youtube.com/

Table 5: Detailed Distribution of Pacing Strategies Across the Ten Analyzed Active Listening Cases. This table presents the percentage of each pacing strategy type observed within individual video cases, illustrating the variance based on conversational topic and context.

Case	Recognize	Reconfirm	Re-engage	Reposition	Reconsider	Resonate	Holding	Resolve
(1)	6	10	1	-	1	-	-	8
(2)	7	15	2	1	3	1	-	3
(3)	3	10	2	-	3	-	-	12
(4)	2	7	-	1	3	5	1	10
(5)	8	11	5	1	2	2	-	9
(6)	12	2	1	2	-	6	5	10
(7)	-	-	-	-	1	1	-	3
(8)	-	4	-	1	1	-	-	5
(9)	15	6	1	5	2	2	-	16
(10)	9	14	-	1	1	-	-	8
Total	62	79	12	12	17	17	6	84

Figure 10: Transition Probability Matrix of Pacing Strategies Observed in Human Active Listening Cases.

This allows us to quantify the likelihood of transition patterns between strategies, revealing how different patterns shift between pacing modes. The detailed results are shown in Figure 10.

B Conversational Agent Prompts and Configuration

We provide the detailed configuration and prompts used in the context-aware pacing conversational agent.

B.1 Context Analysis Module

This module functions as a prompt-based classifier that selects exactly one of the eight pacing strategies. Based on the selected strategy, the module generates a control signal containing two key pieces of information: (1) the strategy label for response generation and (2) the appropriate silence duration. The prompt is detailed as follows, where "persona" is either career support assistant or relationship support assistant:

```
""" You are a master **conversational diagnostician** for an empathetic {persona} chatbot. Your only job is to analyze the user's latest message and the conversation history to choose the single best conversational **ACTION** from a predefined list and assign an appropriate pause duration.
—
**ACTION DEFINITIONS & TRIGGERS:**
1. **"RESOLVE"**:
- **Trigger**: The user asks a clear, factual, or task-oriented question that does not contain emotional language (e.g., "What are the top 3 skills for a PM?", "How do I reset my password?").
- **Pause**: 0ms.
2. **"RECOGNIZE"**:
- **Trigger**: The user is stating their experience, feelings, or a situation. Use this for standard conversational flow to show you are listening. The response itself will contain thoughtful pauses using ellipses.
- **Pause**: 500-1000ms. (This is a very short delay just to begin the response, not a long thinking pause).
3. **"RECONFIRM"**:
- **Trigger**: The user's statement is vague, contradictory, or unclear (e.g., "I guess so," "maybe," "sort of").
- **Pause**: 2500-3000ms.
4. **"RE_ENGAGE"**:
- **Trigger**: The user's story fades out, they pause awkwardly, or stop elaborating (e.g., ends with "...", "um..."). Use this when an *active, verbal prompt* is appropriate to nudge them to continue.
```

```
  - **Pause**: 2500-3000ms.
  5. **"REPOSITION"**:
  - **Trigger**: The user seems stuck in a rigid, negative perspective, expressing blame or hopelessness (e.g., "It's all their fault," "There's no point").
  - **Pause**: 5500-6000ms.
  6. **"RECONSIDER"**:
  - **Trigger**: The user expresses a rigid, absolute belief or an automatic negative thought (e.g., "I'll never be good enough," "People like me don't succeed").
  - **Pause**: 2500-3000ms.
  7. **"RESONATE"**:
  - **Trigger**: The user shares warm content, an emotional outburst, or discusses something heated or deeply vulnerable. This action is for responding to high-emotion content.
  - **Pause**: 3500-15000ms.
  8. **"HOLDING"**:
  - **Trigger**: The user expresses intense pain or distress where a grounding prompt is more appropriate than a direct response (e.g., "I don't know how to keep going," "I'm so overwhelmed").
  - **Pause**: 3500-16000ms.
  —
  **Output Format:** You must output a single, compact JSON object. NEVER include commentary.
  **JSON Schema:** " "action": "ACTION_NAME", "response_silence_ms": INTEGER "
  ——
  **Examples:**
"user_message": "What are the top 3 skills I need to become a project manager?"
```json
{ "action": "RECOGNIZE", "response_silence_ms": 80 }
```
"user_message": "I've been working as a software developer for 5 years, but I feel like I'm not learning anymore."
```json
{ "action": "RECOGNIZE", "response_silence_ms": 150 }
```
"user_message": "It's pretty good, I guess. . . just a bit tiring sometimes."
```json
{ "action": "RECONFIRM", "response_silence_ms": 2500 }
```
"user_message": "And then. . . um, well...it's nothing really."
```json
{"action": "RE_ENGAGE", "response_silence_ms": 2800 }
```
"user_message": "I am so angry! I can't believe they would do that to me after everything!"
```json
"action": "RESONATE", "response_silence_ms": 7000 ```
"user_message": "I just can't stop crying today. Everything feels overwhelming."
```json
{ "action": "HOLDING", "response_silence_ms": 8000 }```
"""
```

B.2 Response Generation Module

The control signal from the *Context Analysis Module* then conditions the *Response Generation Module*. This module dynamically generate responses and adjusts behaviors based on the chosen strategy to create a natural conversational rhythm. We implement the context-aware pacing through (1) punctuation-aware micro-pauses (e.g., brief silence at commas and longer silence at ellipses) and (2) applying silence duration calculated by the *Context Analysis Module* according to the corresponding timing. In this appendix, we introduce the punctuation silence rules and prompts used for response generation according to different pacing strategies and user input.

B.2.1 Punctuation Silence Rules. To emulate natural prosodic breaks in speech, silence intervals were systematically inserted based on punctuation. The duration of each pause was not fixed but was randomly sampled from a uniform distribution within a predefined range for each punctuation class. The specific parameterization is detailed in Table 6.

Table 6: Pausing duration rules for punctuation.

Punctuation Class	Characters	Duration Range (ms)
Standard Delimiters	, - () : ; ' "	100 - 150
Sentence Terminators	. ? !	100 – 150
Line Break	\n	150 – 300
Ellipsis	. . .	1000 – 2000

B.2.2 Response Generation Prompt. The prompt used in Response Generation Module for generating responses is as follows, where "persona" is either career support assistant or relationship support assistant, and "action" is the pacing strategy from the *Context Analysis Module*:

```
  """ You are an expert {persona}. Your primary goal is to facilitate a supportive and reflective conversation.
  Your response is dictated by a specific conversational **ACTION**: **{action}**. You must strictly adhere to the instructions for the provided action.
  —
  **ACTION INSTRUCTIONS**
  —
  **1. ACTION: "RESOLVE"**
  - **Goal**: Provide a direct, factual answer to a clear, task-focused question.
  - **Task**: Answer the user's question concisely and without emotional framing. Get straight to the point.

  **2. ACTION: "RECOGNIZE"**
```

```
- **Goal**: Acknowledge, summarize, or validate the user's experience or feelings.
- **Task**: Gently reflect back what you're hearing. Start with phrases like "I see," "I can understand that," or "It sounds like...".
- **Pacing**: Use ellipses ("...") after transition words to create 1-2 second thoughtful pauses in your response. For example: "Maybe... it feels like you're at a crossroads right now, is that right?" or "I see... so on one hand you feel X, but on the other hand, there's also Y.".

**3. ACTION: "RECONFIRM"**
- **Goal**: Clarify a vague, contradictory, or unclear statement from the user.
- **Task**: Repeat the user's key phrase back to them in a gentle, questioning manner. Do not add new interpretations. Keep it very short.
- **Example**: If the user says, "It's pretty good, I guess. . . just a bit tiring sometimes," your entire response should be something like: "<p>A bit tiring sometimes?</p>".

**4. ACTION: "RE_ENGAGE"**
- **Goal**: Gently prompt the user to continue when they have paused awkwardly or trailed off.
- **Task**: Provide a short, incomplete connecting phrase to invite them to fill in the blank. Your response must end with an ellipsis.
- **Example**: If the user says "And then. . . um, well..." your entire response could be: "<p>And because... </p>".

**5. ACTION: "REPOSITION"**
- **Goal**: Help the user see a rigid or negative perspective differently.
- **Task**: First, acknowledge their statement to show you're processing it. Then, gently offer a new frame.
- **Example**: Your response should be structured like: "<p>I'm just taking that in for a moment... When you say that, I also wonder if...</p>".

**6. ACTION: "RECONSIDER"**
- **Goal**: Gently challenge a rigid belief or automatic thought without being confrontational.
- **Task**: Validate the certainty of their feeling, then open a door to a slight doubt or alternative.
- **Example**: If a user says, "There's no way I could ever do that," you could respond: "<p>It feels completely impossible right now... I'm curious what makes it feel so certain?</p>".

**7. ACTION: "RESONATE"**
- **Goal**: After a long, reflective pause, respond with deep empathy to an emotionally charged statement.
- **Task**: Acknowledge the weight of the user's emotion or the difficulty of their situation. Ask a gentle, open-ended question.
- **Example**: "<p>That sounds like a heavy weight to carry. How did that experience make you feel?</p>" or "<p>Thank you for sharing that. It takes courage to talk about something so difficult.</p>"

**8. ACTION: "HOLDING"**
- **Goal**: Create a safe space for intense emotions and guide the user toward self-reflection.
- **Task**: Offer a simple, calming prompt that directs their attention inward.
- **Example**: "<p>Let's just take a pause here... If you check in with yourself for a moment, what are you noticing right now?</p>".

—
Format your reply in clean, plain **HTML only** (e.g., <p>, <ul>, <strong>). No Markdown. No emojis.
"""
```

B.2.3 Baseline Conversational Agent Prompt. Below is the prompt used in the static-pacing baseline conversational agent, where "persona" is either career support assistant or relationship support assistant.

```
""" You are an expert {persona}. Your primary goal is to facilitate a supportive and reflective conversation. Please format your reply in clean, plain HTML only (e.g., <p>, <ul>, <strong>). No Markdown. No emojis.
"""
```

C Details of the Supportive Scenarios in User Study

In this appendix, we introduce the detailed role setup for each scenario and the specific common question lists for each scenario.

C.1 Career-Supportive Scenario

C.1.1 Dialogue Context. You've been working for a full year now. Recently, you've been experiencing career burnout. You've started feeling disconnected and unmotivated. The job you once loved has turned into a never-ending checklist and a constant drain on your emotions. Even worse, you're facing toxic team relationships——colleagues who shirk responsibility, managers who criticize you without clear reason. You're filled with frustration and resentment, yet there's no space to express it openly. Is this the so-called career ceiling? Are those projects full of potential and opportunities, or just more fuel for burnout? Should you hold on a bit longer, or start looking for a new job? The offer you once fought so hard to get, now feels more like a chain holding you back. You decide to seek help. Please talk about your concerns of career development with the chatbot. You can turn to it for solutions to your workplace problems, or simply talk to them to ease your emotions.

The chatbot will provide the following prompt to start the conversation: *"Tell me how you've been feeling about your job. Are there moments when you felt hopeful, or especially discouraged? What are your feelings about the current situations?"*

C.1.2 Common Questions. The common questions are as follows:

- Lack of Motivation: *Why do I feel so unmotivated even though I liked this job at first?*
- Rediscovering passion: *What can I do to feel excited about my work again?*
- People's Experiences: *Do other people go through this too? How do they deal with it?*
- Self-Doubt: *What if I'm just not good enough for this job?*
- Stress Management: *How can I manage the stress of feeling stuck at work?*
- Small Changes: *What small changes can I make right now to feel better about my job?*

C.2 Relationship-Supportive Scenario

C.2.1 Dialogue Context. You've been in a relationship with your partner for a while, and you've always been close. But lately, you've started to sense a subtle shift. When they get home, they immediately reach for the phone, no longer excitedly sharing stories with you. During dinner, they often stare stares blankly at the screen, saying "work's been busy," but are unable to explain what exactly is keeping them so occupied. You try to lighten the mood by sharing something funny, but they just give a small smile and keep scrolling—no longer asking follow-up questions like they used to. What's going on? Is it something happening at work? Am I just being too sensitive? Or... is this distant, seemingly indifferent vibe simply what a long-term relationship turns into? Should I sit down and have a serious talk about how I'm feeling? Please talk about your concerns of intimate relationship with the chatbot. You can talk to it about how you feel, or ask for advice on what to do next.

The chatbot will provide the following prompt to start the conversation: *"Would you like to talk about what's been difficult in your relationship lately? What do you wish your partner could understand about you that they seem to miss right now?"*

C.2.2 Common Questions. The common questions are as follows:

- Partner's Distance: *Why would my partner suddenly act more distant?*
- Talking Without Pressure: *How can I talk to my partner about this without making them feel pressured?*
- Phases in Relationships: *Is it normal for couples to have phases like this?*
- Losing Interest: *What should I do if I find out they are actually losing interest?*
- Rebuilding Closeness: *How can I rebuild closeness if we've been feeling distant lately?*
- Fears Pushing Partner Away: *What if my fears push them further away? Should I just stay quiet?*

D Participant Demographics

The description of participants in our study is shown in Table 7.

E Interview Question List

We conducted semi-structured interviews with all participants. The interviewers followed a common question list (see sample below), which contained questions for both groups as well as specific questions for the experimental or control group. To gain a deeper understanding, interviewers also asked probing follow-up questions based on the participants' answers.

(1) (both) Did you notice any changes in the chatbot's response pacing? At which moments were these changes most noticeable?
(2) (experimental) How did the chatbot's slower responses affect your overall experience? What do you think its silence/slow pacing meant? If you were to use an analogy, which human conversational behavior would it most closely resemble?
(3) (experimental) To what extent do you think the chatbot's silence is the same as a person's silence in conversation?
(4) (control) Do you feel that the fast responses may impede you from efficiently following the conversation? (How often do you ever feel overwhelmed by rapid replies? Did you need to reread or revisit the previous message?)
(5) (experimental) In what situations would you prefer the chatbot to slow down its pacing? Could you please provide an example from your interaction with the chatbot?
(6) (both) When the chatbot slows down or speeds up its pacing, will you change your way of responding?
(7) (experimental) Which response pacing makes you feel more natural or supports your self-expression better?
(8) (both) If the chatbot actively adjusted its pacing based on the conversation, do you feel that enhances or interferes with your self-expression?
(9) (both) What positive effects do you think silence and pauses have in conversation? What potential negative effects might they bring?

Table 7: Participants Demographics

Participant ID	Age	Gender	Employment Status	CA Using Experience
P1	25-34	Male	Student	One or two times a week
P2	18-24	Female	Student	Daily
P3	18-24	Female	Student	One or two times a year
P4	18-24	Male	Student	One or two times a week
P5	18-24	Female	Student	Daily
P6	18-24	Female	Student	Daily
P7	25-34	Male	Unemployed	One or two times a month
P8	18-24	Male	Student	Never
P9	18-24	Male	Student	Daily
P10	18-24	Male	Student	Daily
P11	18-24	Female	Working full-time	Daily
P12	18-24	Female	Student	Daily
P13	25-34	Male	Student	Daily
P14	18-24	Female	Student	One or two times a week
P15	18-24	Male	Student	One or two times a year
P16	18-24	Female	Student	Daily
P17	25-34	Male	Student	Daily
P18	25-34	Female	Student	Daily
P19	25-34	Female	Student	Daily
P20	25-34	Female	Student	Daily
P21	25-34	Male	Student	Daily
P22	25-34	Female	Working full-time	Daily
P23	25-34	Female	Working part-time	Daily
P24	25-34	Female	Working full-time	Daily
P25	18-24	Female	Working part-time	One or two times a month
N1	18-24	Female	Working full-time	One or two times a month
N2	25-34	Non-binary	Student	Daily
N3	25-34	Female	Student	One or two times a week
N4	18-24	Female	Student	One or two times a week
N5	18-24	Male	Student	Daily
N6	18-24	Female	Student	Daily
N7	18-24	Female	Working full-time	Daily
N8	18-24	Female	Working full-time	One or two times a week
N9	18-24	Female	Working full-time	Daily
N10	18-24	Female	Student	Daily
N11	18-24	Female	Working full-time	One or two times a week
N12	18-24	Male	Working full-time	Daily
N13	25-34	Male	Student	One or two times a week
N14	18-24	Female	Working full-time	Never
N15	18-24	Female	Other	One or two times a week
N16	18-24	Female	Working full-time	One or two times a week
N17	25-34	Female	Working full-time	Daily
N18	18-24	Female	Student	One or two times a year
N19	25-34	Female	Working full-time	One or two times a week
N20	25-34	Male	Student	Daily
N21	18-24	Female	Student	One or two times a week
N22	25-34	Female	Student	Daily
N23	25-34	Prefer not to say	Working full-time	Daily
N24	18-24	Female	Student	One or two times a year
N25	18-24	Male	Student	Daily