“Having Lunch Now”: Understanding How Users Engage with a Proactive Agent for Daily Planning and Self-Reflection

Adnan Abbas
Virginia Tech
Blacksburg, VA, USA
adnana99@vt.edu

Caleb Wohn
Virginia Tech
Blacksburg, VA, USA
cwohn@vt.edu

Arnav Jagtap
Virginia Tech
Blacksburg, VA, USA
arnavj03@vt.edu

Eugenia H. Rho
Virginia Tech
Blacksburg, VA, USA
eugenia@vt.edu

Young-Ho Kim
NAVER AI Lab
Seongnam, Republic of Korea
yghokim@younghokim.net

Sang Won Lee*
Virginia Tech
Blacksburg, VA, USA
sangwonlee@vt.edu

Abstract

Conversational agents have been studied as tools to scaffold planning and self-reflection for productivity and well-being. While prior work has demonstrated positive outcomes, we still lack a clear understanding of what drives these results and how users behave and communicate with agents that act as coaches rather than assistants. Such understanding is critical for designing interactions in which agents foster meaningful behavioral change. We conducted a 14-day longitudinal study with 12 participants using a proactive agent that initiated regular check-ins to support daily planning and reflection. Our findings reveal diverse interaction patterns: participants accepted or negotiated suggestions, developed shared mental models, reported progress, and at times resisted or disengaged. We also identified problematic aspects of the agent's behavior, including rigidity, premature turn-taking, and overpromising. Our work contributes to understanding how people interact with a proactive, coach-like agent and offers design considerations for facilitating effective behavioral change.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; **Natural language interfaces**.

Keywords

Human-AI Interaction, Planning, Self-Reflection, Conversation Analysis

ACM Reference Format:
Adnan Abbas, Caleb Wohn, Arnav Jagtap, Eugenia H. Rho, Young-Ho Kim, and Sang Won Lee. 2026. *“Having Lunch Now”*: Understanding How Users Engage with a Proactive Agent for Daily Planning and Self-Reflection. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3772318.3790957

1 Introduction

With the presence of large-language models (LLMs), Conversational Agents (CAs) are becoming more human-like, exhibit better conversational intelligence and novelty in interactions [54, 82, 98]. LLM-powered CAs can interpret intent, adapt their tone, and maintain contextual continuity across multiple sessions [57, 63]. Leveraging these capabilities, researchers have studied coaching-style agents: systems designed to advise, persuade, and guide everyday behavior [9, 42, 57, 107]. Such agents are often *proactive*: rather than waiting for user input, they initiate interactions by reminding, suggesting, or steering users toward specific actions [12, 57, 111].

Proactivity can be valuable in human–AI interaction. Studies show that proactive CAs can introduce new topics [58], offer richer suggestions [76], and persuade users toward target behaviors [53]. One domain where proactive CAs have been employed is supporting Knowledge Workers' (KW) productivity and well-being [1, 4, 46, 47, 109]. In these studies, CAs take the role of a productivity or well-being coach and regularly prompt KWs to articulate plans or reflect on their day, proactively leading the conversation. By initiating daily check-ins (e.g., morning and evening), CAs can help make users' intentions more concrete and actionable, thereby improving adherence to planning [33, 46, 62]. Through reflective prompts, these agents can increase users' awareness of everyday habits and foster behavior change [36, 47, 68]. These studies report positive outcomes such as greater engagement with work, improved work–life balance, better physical health, and more structured planning [1, 3, 4, 33, 46, 47].

However, we still lack an understanding of *how* people actually communicate with CAs in everyday work, including what happens when the agent's proactivity does not land well. This limits our ability to explain the mechanisms through which CAs support (or disrupt) desirable behavioral changes in the context of productivity and well-being. These questions are especially important for LLM-based agents, which use human-like dialogue that encourages users to treat them socially [42, 57], even though their suggestions can still be unpredictable—helpful at times, but also hallucinated, or poorly timed [57, 81]. Recent work further shows that LLMs are poor at initiating grounding, leading to communication breakdowns [90, 91]. Unlike task-oriented systems where breakdowns occur when the agent misunderstands user intent and fails to complete a task [10, 70], we anticipate additional forms of breakdown in

---

[*] Sang Won Lee conducted this work as a visiting scholar at NAVER AI Lab.

This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790957

proactive settings, where the agent's behavior do not meaningfully support its overarching goal.

To address these gaps, we examine how users engage with a proactive coaching agent and the deficiencies in agent behaviors that contribute to breakdowns. In our study, we define breakdowns as observable moments in the interaction where the participants and CA failed to maintain a shared understanding and conversational pace, leading to disruptions in the exchange [10, 22, 51, 70]. For example, such cases can include moments when the conversation stalls, becomes unhelpful or dismissive, or otherwise diverges from the initial goal. In this paper, we explore this broader conceptualization of breakdowns in the context of proactive agent–user interactions based on observable behavior in conversations. Understanding these breakdowns is critical for designing proactive systems that can improve conversational experience and foster cooperation in human-AI partnerships.

To study these dynamics, we deployed a proactive, coaching-style agent (PITCH) for two weeks with 12 graduate students. PITCH checked in twice per day, once in the morning to facilitate planning and once in the evening to support self-reflection, resulting in 336 conversations (3,181 turns). We applied a codebook-based thematic analysis to generate qualitative patterns in planning, reflection, cooperation, and breakdown, and a user behavior dialogue-act analysis to quantify how often different behaviors occurred and how they varied across sessions and conditions. Specifically, we ask:

- **RQ1**: How do users engage with a proactive agent that facilitates daily planning and reflection?
- **RQ2**: What conversational behaviors of the coaching agent contribute to breakdowns in conversational interaction?

At a high level, our analysis shows that users actively shaped the interaction: they adopted suggestions when useful, corrected the agent's assumptions, and used the chat to self-reflect on past activities and recognize desired behavior changes. We also observed systematic breakdowns in the interaction, such as premature turn-taking and rigidity, which point to design opportunities for more flexible and context-aware proactive agents.

This study makes three contributions: (i) a thematic and dialogue-act analysis that provides an in-situ account of how users communicate with a proactive coaching agent; (ii) an anonymized dataset of 336 conversations (3,181 turns); and (iii) design considerations for building more adaptive and context-aware proactive CAs.

2 Related Work

In this section, we review prior work on conversational agents (CAs) that support behavior change, with a focus on how these systems shape users' planning, reflection, and everyday work practices. We organize our review around a conceptual framework that guides our study, which consists of four dimensions: (1) planning and reflection, which describe how conversations help users articulate and evaluate their daily activities; (2) social interpretation of the agent, which influences how users perceive and respond to CA behaviors; (3) cooperation, which captures how users and agents work to maintain shared understanding during interaction; and (4) proactivity, which concerns how agents initiate and guide conversations rather than waiting for user input. Although these dimensions emphasize different aspects of conversational interaction, prior work suggests that they interact in shaping how users engage with CAs that support behavior change [20, 104]. Because conversations leverage social intelligence, understanding users' planning and reflection behaviors requires examining how they interpret the agent as a social entity. Similarly, a CA's proactivity structures when and how opportunities for planning and reflection arise, for instance, through check-ins, follow-ups, or suggested next steps, but its effects depend on users' willingness to cooperate. In productivity contexts, such proactivity may also introduce a sense of accountability by positioning the agent as an ongoing social presence, which prior work suggests can support task tracking and adherence [1, 46]. Conversely, mismatches across these dimensions, such as misaligned interventions, unclear conversational moves, or users' uncertainties about system capabilities, can hinder meaningful interaction and lead to breakdowns [67, 83, 84]. These dimensions provide a lens for examining both how users engage with proactive planning agents and how specific conversational behaviors may hinder meaningful interaction.

2.1 Planning and Reflection with Conversational Agents at Work

HCI researchers have explored the role of conversational interaction to facilitate planning and reflection for KWs' productivity. Several works developed an agent to collect self-reported tasks from users, manage scheduling, and facilitate work planning [33, 46, 47, 64]. For example, Amber [46] initiated conversations to elicit user's tasks, provided reminders for task scheduling, suggested breaks, and facilitated work reflections based on user interactions and context sensing. At the same time, assistants also helped with automating some of the KWs' tasks, such as scheduling meetings [24], managing teamwork [101], finding work resources [56], organizing email inboxes [27], and managing to-do lists [29]. Findings from these works highlight benefits for productivity such as increased plan adherence and deeper work engagement [46, 47].

Beyond productivity, CAs have also been studied to promote well-being. This includes prompting healthy breaks [3, 4, 18], facilitating reflection on work-life balance [109], and helping to manage stress [28, 65]. Similar works have explored how KWs can be encouraged to take work breaks for their physical and mental well-being to mitigate the negative effects of sedentary lifestyle [3, 4, 18]. Research shows that engaging in reflective conversations can facilitate self-awareness, helping people be more aware of task switching, track progress, evaluate obstacles, and reflect on higher-level goals [25, 46, 47, 59, 64, 100]. Such reflection fostered healthier work practices and greater satisfaction, particularly in fast-paced, busy environments where self-reflection is often overlooked [47].

Although the above works show that users improve their productivity and well-being by interacting with CAs, most studies have focused on outcomes (e.g., productivity, stress reduction, perceived usefulness) or user perceptions, but have not closely observed users' conversational behaviors. Analyzing these behaviors provides a deeper understanding of human-ai interaction by identifying recurring patterns, assessing user engagement, and evaluating the effectiveness of conversational design [19].

Moreover, prior studies were conducted in the era before LLMs [33, 46, 47, 109], where the agent's role was more reactive to user input and limited because of the static, fixed dialogue structures. Such structural limitations undermines the quality of reflection as users do not get a chance to engage deeply in response to the posed questions [21]. LLMs now enable more dynamic interactions by exhibiting better conversational intelligence [54, 98], persuasion [44, 85], and novelty in interactions [82].

Building on these capabilities, recent works have begun exploring AI in a coaching role [9, 42, 94, 107], yet we still lack a detailed understanding of how users actually behave in dialogue with such coach-like agents—how they accept, negotiate, resist, or dismiss the agent's suggestions during daily planning and reflection. Our work addresses this gap by examining in-situ conversation logs from a fourteen-day longitudinal study, focusing on the emergent behaviors that characterize how people engage in daily planning and self-reflection with an LLM-based coach.

2.2 Social Perception of Conversational Agents

Users' perceptions of a CA's social presence strongly shape how they engage with it [104]. Prior research consistently shows that when people interpret a CA as a social partner rather than a technical tool, they are more likely to reflect on their behaviors, sustain interaction, and incorporate suggested changes into their routines [9, 20, 56, 81]. In productivity contexts, for example, agents that follow up on users' stated plans can evoke a sense of social accountability [1, 67, 106]. Because conversational interfaces mimic human dialogue, users often respond with politeness, humor, self-disclosure, or repair behaviors consistent with the CASA paradigm[1] [56, 79, 88]. Yet users differ in their willingness to treat agents as social entities: some readily maintain engagement and work to repair breakdowns, while others view the CA strictly as a tool and disengage after a single error [31, 51]. Research on anthropomorphism similarly reports mixed effects, sometimes increasing satisfaction [79], but in other cases, frustrating or even unsettling users [33, 81].

Because users' social perceptions shape their willingness to engage, persist, and repair, cooperation becomes the key interactive process through which a CA and user work toward shared goals [22]. In task-oriented scenarios, where users interact with AI systems to achieve a concrete outcome (e.g., processing a return with a customer service bot or booking a ticket with an airline virtual agent), cooperation is critical to reaching the goal of the interaction [50, 80]. Prior work has shown that when conversational agents are perceived as cooperative, users consider their information richer, rate the interaction as higher quality, and report greater satisfaction [37, 50]. To cooperate, people often adapt their own behavior: simplifying or reformulating messages, rephrasing to repair breakdowns, or even using gestures with robots to accommodate system constraints [55, 70, 75, 78, 80]. However, this process is also challenging because chatbots often display limited intelligence while users bring expectations of human-like competence and conversational norms, leading to mismatches and frustration [35, 61]. Unsatisfactory responses such as turn-taking failures, hallucinations, incorrect information, or lack of flexibility undermine user trust and can ultimately reduce engagement [5, 10, 26, 31]. While we know how cooperation is enacted in task-centric chatbots [80], we lack a clear account of how it unfolds when users plan and reflect with an LLM coach that provides productivity and well-being advice.

Our study fills this gap by examining how users perceived the agent as a social entity influencing their planning and reflection behaviors. User behaviors in this setting would be different than in task-centric scenarios because unlike service chatbots that follow a request–response–confirmation template, our system involves users in relatively open-ended planning and reflection. The interactions are less about completing a transaction and more about scaffolding self-reflection, prompting users to accept, decline, or negotiate suggestions. In this context, cooperation involves more than providing slot values–it includes sharing context, disagreeing or negotiating advice, and aligning conversations with one's preferences.

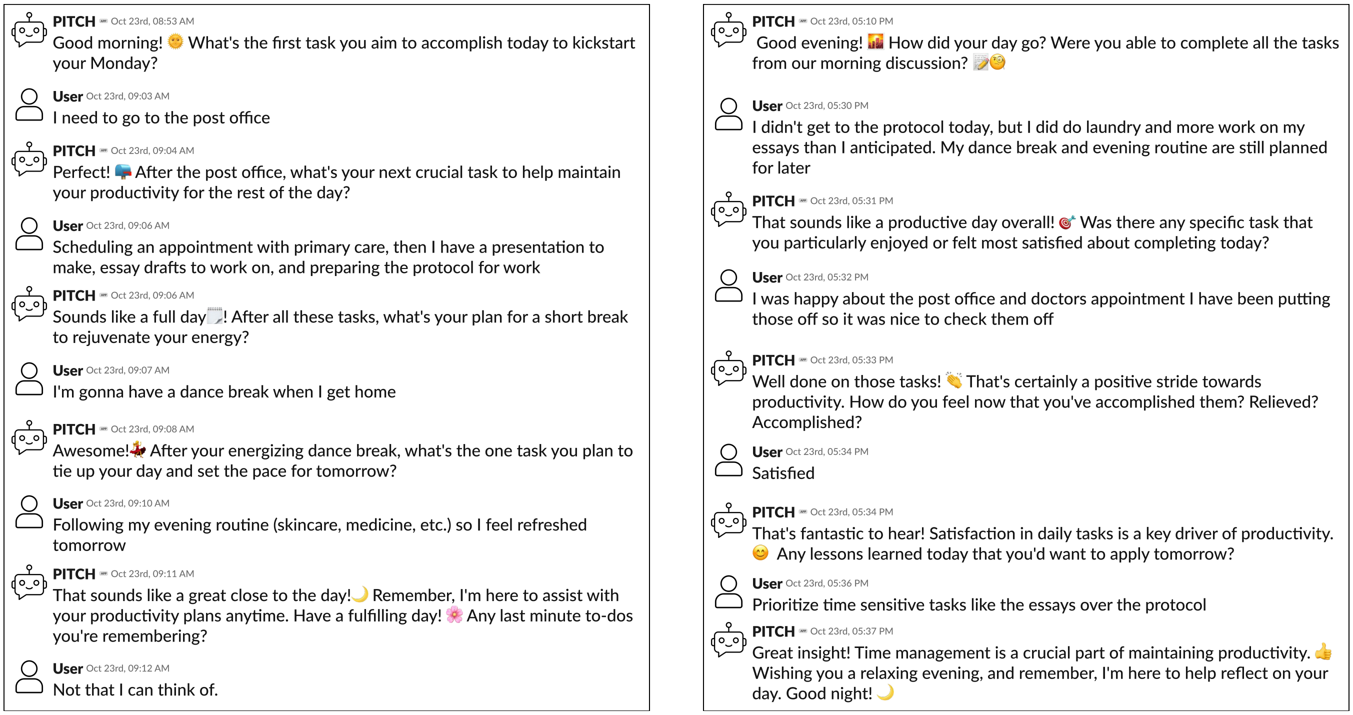
2.3 Designing Proactive Agents for Behavior Change

A longstanding debate in HCI concerns how much initiative should rest with users versus systems. Shneiderman argued for direct manipulation interfaces where "users comprehend the display, feel in control, [and] the system is predictable" so they remain responsible for outcomes [95]. Maes, by contrast, saw growing complexity as making delegation inevitable: software agents should be personalized, proactive, long-lived, and adaptive, taking initiative on behalf of overloaded users. Framing these positions as a strict opposition has proven limiting [95]. Horvitz later described this as a "false dichotomy" and advanced principles of mixed-initiative interaction, where automated services and direct manipulation are coupled [38]. Horvitz's Lookout system, for example, anticipated calendar events from email but always allowed users to confirm or refine results. Allen extended this view to dialogue, defining mixed-initiative as a flexible strategy in which initiative is dynamically negotiated: sometimes the system reports or clarifies, sometimes the user directs, and sometimes roles shift opportunistically [7].

Building on this lineage, recent work has asked whether conversational agents should be reactive or proactive. Reactive agents respond only when the user initiates, giving users more control but limiting the system's role. Proactive agents, by contrast, initiate interactions by offering reminders, or suggesting actions [12, 57, 111]. Chaves and Gerosa describes proactivity as sharing initiative with the user, where the agent may start exchanges, propose topics, or offer follow-up questions. In the context of coaching, people want CAs to be proactive in making suggestions to improve their productivity [20, 52, 72]. Research shows that agent proactivity can introduce new topics [58], provide richer suggestions [76], and persuade users toward a target behavior [53]. Researchers have also mapped proactivity on a spectrum to show how much initiative rests with the user versus the system. Isbell and Pierce outlines an Interface-Proactivity continuum that ranges from the user doing everything on their own to the system acting on their behalf, with several mixed approaches in between where the system may

---

[1] The Computers Are Social Actors (CASA) paradigm argues that people apply social norms and expectations from human–human interaction to computers when those computers display social cues [71].



PITCH Oct 23rd, 08:53 AM
Good morning! 😊 What's the first task you aim to accomplish today to kickstart your Monday?

User Oct 23rd, 09:03 AM
I need to go to the post office

PITCH Oct 23rd, 09:04 AM
Perfect! 📪 After the post office, what's your next crucial task to help maintain your productivity for the rest of the day?

User Oct 23rd, 09:06 AM
Scheduling an appointment with primary care, then I have a presentation to make, essay drafts to work on, and preparing the protocol for work

PITCH Oct 23rd, 09:06 AM
Sounds like a full day🗒️! After all these tasks, what's your plan for a short break to rejuvenate your energy?

User Oct 23rd, 09:07 AM
I'm gonna have a dance break when I get home

PITCH Oct 23rd, 09:08 AM
Awesome!💃 After your energizing dance break, what's the one task you plan to tie up your day and set the pace for tomorrow?

User Oct 23rd, 09:10 AM
Following my evening routine (skincare, medicine, etc.) so I feel refreshed tomorrow

PITCH Oct 23rd, 09:11 AM
That sounds like a great close to the day!🌙 Remember, I'm here to assist with your productivity plans anytime. Have a fulfilling day! 🌸 Any last minute to-dos you're remembering?

User Oct 23rd, 09:12 AM
Not that I can think of.

PITCH Oct 23rd, 05:10 PM
Good evening! 🌇 How did your day go? Were you able to complete all the tasks from our morning discussion? 📝🧐

User Oct 23rd, 05:30 PM
I didn't get to the protocol today, but I did do laundry and more work on my essays than I anticipated. My dance break and evening routine are still planned for later

PITCH Oct 23rd, 05:31 PM
That sounds like a productive day overall! 🎯 Was there any specific task that you particularly enjoyed or felt most satisfied about completing today?

User Oct 23rd, 05:32 PM
I was happy about the post office and doctors appointment I have been putting those off so it was nice to check them off

PITCH Oct 23rd, 05:33 PM
Well done on those tasks! 👏 That's certainly a positive stride towards productivity. How do you feel now that you've accomplished them? Relieved? Accomplished?

User Oct 23rd, 05:34 PM
Satisfied

PITCH Oct 23rd, 05:34 PM
That's fantastic to hear! Satisfaction in daily tasks is a key driver of productivity. 😊 Any lessons learned today that you'd want to apply tomorrow?

User Oct 23rd, 05:36 PM
Prioritize time sensitive tasks like the essays over the protocol

PITCH Oct 23rd, 05:37 PM
Great insight! Time management is a crucial part of maintaining productivity. 👍 Wishing you a relaxing evening, and remember, I'm here to help reflect on your day. Good night! 🌙

(a) Morning Plan Externalization

(b) Evening Reflection

Figure 1: An example of the morning and evening conversations from the field study. PITCH checks in with the user in the morning to externalize their plan (a) and follows up in the evening to support self-reflection (b).

highlight information or suggest options. Recent efforts to push AI toward agentic behavior, such as the release of the ChatGPT agent [74], illustrate the far end of this spectrum where the system can act on the user's behalf. Yet, highly proactive designs can also be perceived as intrusive, disruptive, or controlling [56, 76, 99]. With the rise of LLM-powered agents, the model responses can exhibit agency in unpredictable ways that produce negative emotional effects toward the technology [81].

Across these strands of work, the challenge is balancing initiative in ways that adapt to users' evolving needs. Proactivity can keep conversations goal-oriented [53], but it can also constrain flexibility when users' contexts shift. Our study extends this literature by examining how users engage with an LLM-based coach that proactively decides planning and reflection goals, highlighting how proactivity can be helpful and when it can go out of context.

3 PITCH - A Proactive Agent for Daily Planning and Self-Reflection

3.1 Design Goals and Implementation

We designed two versions of PITCH that participants used each week in random order: a fixed-goal version (PITCH-N) and a rotation-based version (PITCH-R). We designed PITCH to support reflection in users' plans, with morning conversations prompting them to externalize their intentions and evening conversations returning to the externalized plan to support self-reflection. Below, we detail the design specifications.

3.1.1 Daily Conversational Interaction for Morning Plan Externalization. PITCH was designed to act as a productivity coach that proactively checks in with a worker each morning. The morning interaction centered on planning, with the goal "to help workers structure their day to enhance productivity and well-being (PITCH-N)." By answering PITCH's questions, users can be reminded of the tasks they consider important for the day, and the conversational format prompts people to articulate task details they may not otherwise specify in a traditional task list (e.g., priorities, timing, and expected effort). Prior work shows that externally articulating intentions supports prospective memory [30], and mentally simulating a task can improve plan adherence [8].

The morning conversation therefore serves as the point of "externalization" that the evening interaction later reflects back on. This connection to users' day's plans is a key design choice for studying self-reflection: participants first lay out a concrete, self-generated plan, which then becomes the shared reference point for evaluating their day through conversation. Example prompts used in the morning included:

- Good morning! What are the top three tasks you plan to accomplish today to make your Wednesday productive?
- Good morning! Since it's Saturday, what enjoyable activity have you planned for yourself today?

The full morning prompt is available in Appendix A.1.1. An example morning interaction is shown in Figure 1a.

3.1.2 Personalized Reflections on Morning Plans. To support self-reflection in the evening, we leveraged the morning conversation history to enable PITCH to ask questions grounded in what the user had planned earlier that day. In the evening system prompt, the assistant's goal was defined as "to help workers reflect on their day for improved productivity and well-being, (PITCH-N)" and the prompt explicitly instructed the agent to keep questions "within the context of the morning conversation."

Including the full morning context operationalizes self-reflection in a concrete way: users are asked to revisit the intentions they articulated earlier, compare them with what actually happened, and describe reasons for fulfillment or deviation. This allowed us to observe reflection behaviors because participants are prompted to evaluate their plans, reconstruct their day, and make sense of successes or difficulties. Prior work has used prior conversational context to guide empathetic responses [89], journaling [45], and well-being monitoring [41]. Here, we used this technique to ensure that reflective questions are not generic but tied to the user's own earlier commitments. In contrast, some previous approaches relied on general-purpose questions such as, "How much progress do you feel you made on your scheduled task?" [46]. Example evening prompts included:

- Good evening! Hope your day went well. Were you able to complete the main paragraph of your lit review as planned?
- Good evening! How was your day practicing mindfulness through body scans? Did you find any satisfying moments?

Each question directly referenced the morning plan, making the reflective behavior observable in the dialogue. The full evening prompt is available in Appendix A.1.2. An example evening conversation is shown in Figure 1b.

3.1.3 Lightweight Check-ins for Sustained Engagement. We ensured that PITCH stayed focused on its conversational goal without becoming formal or excessively verbose, since brief check-ins are more suitable for long-term engagement [67]. Early pilot tests showed that PITCH sometimes extended conversations through multiple follow-up questions and its verbosity. While extended conversations can be beneficial, our aim was to support the users' thought process through drawing questions rather than PITCH conveying information. Moreover, earlier studies showed that KWs do not prefer extended conversations as they are short on time [33, 72]. This included prompting the model to keep responses short and easy to understand, an observer agent checking for conversation length, using emojis appropriately, and enforcing a 30-word limit per message.

3.1.4 PITCH-R : Rotation Strategy. PITCH also explored whether dynamically rotating goals each day affected engagement and introduced variation in users' reflection behaviors. The default planning and reflection goals described in Sections 3.1.1 and 3.1.2 were used for PITCH-N. In contrast, PITCH-R used more specific goals drawn from self-help literature and academic work. To reduce redundancy, the order of goals was shuffled. A randomly chosen morning goal and a randomly chosen evening goal (from N possible goals each) produce N^2 possible daily combinations. This design increases diversity in conversation topics and creates opportunities for studying how conversational behaviors vary with different productivity and well-being topics. The detailed system architecture and conversation goals are described in our previous work [1].

3.1.5 Technical Implementation. We implemented PITCH as a chatbot application in Slack [97], a platform widely used by KWs, so that users would not need to leave their existing work environment to interact with the system. We built PITCH using the Slack Bolt framework for Node.js. For message generation, we used OpenAI's GPT-4 API. MongoDB was used to store daily conversations and associated metadata. Further technical details can be found in our previous work [1].

4 Method: Field Deployment Study

In October 2024, we conducted a two-week (14 weekdays) deployment study to investigate how users interact with PITCH, which initiated morning conversations to help externalize their plans and evening conversations to support self-reflection. The Institutional Review Board (IRB) of the researcher's university approved the procedure of our study, and informed consent was obtained from all participants. We organized the study procedure into three main phases: an onboarding session, a field deployment, and a post-study interview session. While we conducted post-study interviews and collected diary survey responses, we published the results that focused on understanding users' perceptions and experience [1]. In this study, we focused on conversation analysis from the same study to understand how participants behaved with PITCH, and we did not include any results from the interview or survey data.

4.1 Study Procedure

We conducted onboarding sessions in person or over Zoom to introduce participants to PITCH. During onboarding, we demonstrated the Slack-based chatbot, explained the study procedure, and provided written instructions provided in Appendix A.2.1. Participants were invited to a dedicated Slack workspace and asked to complete a pre-study demographic form (age, gender, race/ethnicity, occupation, daily activities, and time spent). Following onboarding, participants began a two-week field deployment. Each day, PITCH initiated check-ins at user-specified times in the morning and evening. Participants engaged with these prompts and completed a brief daily survey. All interactions with PITCH were conducted in English. We collected all conversation interactions between participants and PITCH during the study. At the end of the two-week study, we did a semi-structured debriefing interview. Participants received $50 as compensation for their participation.

4.2 Participants

We advertised the study through broad university channels, including institution-wide bulletin boards and targeted mailing lists for graduate students in graduate school, as well as the author's department. We recruited 12 participants who met our inclusion criteria: (1) were over 18 years old, (2) were fluent in verbal and written communication in English, (3) were familiar with Slack, (4) were interested in interacting with a productivity AI agent, and (5) were managing different types of daily tasks. All participants were full-time graduate students (2 MS, 10 PhD) in Computer Science, with one exception (Physics), at the researchers' university, and had a basic knowledge of computing. Graduate students in STEM fields, although not representative of all knowledge workers, face many demanding characteristics that knowledge workers have, including independent task management, competing deadlines, flexible working hours, various responsibilities, and blurred work–life boundaries. This made them an appropriate exemplar group for an initial *exploratory* investigation into planning and reflection interactions with proactive conversational agents, whose demands are amplified compared to those of average knowledge workers. As part of the pre-study questionnaire, participants self-reported their frequency of Slack use and their time-management skill level by answering multiple-choice questions. These measures were collected only to describe the sample and are not used as variables in our analysis. Of the 12 participants, 6 identified as male, 1 as non-binary, and 5 as female. Their ages ranged from 22 to 35 (μ = 26.91; σ = 3.87). Slack use frequency also varied across participants: 33.3% reported always using Slack, 33.3% often, and 33.3% sometimes. Demographic information is summarized in Table 1.

Table 1: Participant demographics and self-reported metrics. Gender abbreviations: M = Man, W = Woman, N-B = Non-binary. Participants managed diverse tasks including research, coursework, teaching, and service responsibilities with varying time allocations.

PID	Gender / Age	Race	Time Mgmt	Typical Daily Activities
P1	M / 25	Asian	Fair	Research *(2–3h)* || Coursework *(7h/wk)* || Meetings *(3–4h/wk)*
P2	M / 33	White	Good	Research *(3–9h)* || Teaching *(1.5h Tu/Th/Fr)* || Meetings *(1h/wk)* || Programming *(variable)*
P3	M / 24	Asian	Good	Research *(5h)* || Coursework *(3h)* || Programming *(3h)*
P4	W / 25	Asian	Good	Teaching/Grading *(1h)* || Research *(4h)* || Coursework *(4h)*
P5	M / 27	Asian	Good	Research *(variable)* || Office Hours || Coursework *(2h)*
P6	W / 27	Asian	Fair	Research || Coursework || Meetings || Coding || Classes
P7	W / 35	White	Fair	Teaching *(2h)* || Meetings *(5–6h)* || Programming *(3h)*
P8	W / 30	Asian	Fair	Research *(4–5h: programming, writing, reading)* || Meetings *(2–3h)* || Mentoring *(1h)*
P9	M / 25	White	Fair	Teaching *(1h)* || Meetings *(0.5h)* || Coursework *(1h)* || Research *(2h)*
P10	W / 24	Asian	Very Good	Class *(5h)* || Work Shift *(5h)* || Coursework *(5h)* || Other *(3h)*
P11	N-B / 22	White	Fair	Research *(4h)* || Classes *(2–4h)* || Work Meetings *(0–1h)* || Club Meetings *(0–1h)*
P12	M / 27	Asian	Fair	Research || Meetings || Programming || Writing

4.3 Data Analysis

4.3.1 Codebook-based Thematic Analysis. We conducted a codebook-based thematic analysis to surface and articulate semantic behavioral patterns in the data [16, 17, 23]. The analysis was primarily inductive, where codes and themes were developed through close engagement with the data rather than from pre-existing literature or a framework, while remaining theoretically informed by our research aims (understanding user behaviors in daily planning and reflection). Our epistemological stance was informed by essentialist/realist approach where we treat participants' conversations as meaningful access to experience, while acknowledging the role of context and interpretation [16].

The conversation data was stored in a spreadsheet where each row was the user's or chatbot's turn. Data columns were: participant ID, date, day, session (Morning/Evening), timestamp, and response text. The first author familiarized themselves with the data, taking notes, and then generated initial codes systematically across the dataset (e.g., "user shared a ranked task list" or "agent suggested user to break down their task"), documenting them in a separate spreadsheet with code IDs and descriptions. Each response by the agent or user was considered in the analysis and was tagged with either one code, multiple codes, or no codes, as relevant. After coding roughly 50% of the data, 226 codes were generated. To refine and stabilize the codebook, the first author and three other researchers met to collectively review and discuss the codes with the aim to merge repetitive codes and clarify ambiguous ones. When any code was unclear to the team, we referred back to the conversation transcripts to ground the code's meaning within its specific conversational context. This collaborative session established shared understanding and consistency in code definitions.

Using the refined codebook, the first author then coded the remaining 50% of the data, resulting in a total of 439 codes. Thematic saturation was achieved during the final coding cycles, as the last two participants' conversations yielded no new codes. Next, the first author and three other researchers met to search for candidate themes by collating and grouping related codes. The 439 codes were transferred from the spreadsheet to a shared Miro board [66] as sticky notes to facilitate collaborative theme development and visualization through affinity diagramming. The team created initial theme headings and systematically analyzed each code, placing it under the most appropriate theme. When codes did not fit cohesively within existing themes, new themes were proposed and discussed among the researchers. This process resulted in the identification of 32 candidate themes, organized into two analytical categories: 13 themes capturing behavioral patterns in the chatbot's responses and 19 themes capturing patterns in users' behaviors (Tables 6, 7 in Appendix A.3). The process of refining themes was iterative, involving four in-person research team meetings to review the visual board, assess internal coherence, and ensure distinctions between themes. Throughout this process, codes were merged, revised, or repositioned as needed.

4.3.2 User Behavior Dialogue Act Analysis. To complement the thematic analysis and examine the frequency of user behaviors, we conducted a dialogue act analysis. Dialogue act analysis is a method from linguistics and NLP that examines each utterance in conversations to understand the conversational and linguistic behaviors speakers engage in [6, 43, 102, 103]. For instance, for an utterance "*I am sorry*," a dialogue act can be *Apology* [43]. Researchers have designed different taxonomies for conversations in unique contexts (e.g. task-oriented, casual) as the nature of dialogue in every context differs, resulting in different frequency of dialogue acts [102].

Similar to prior works [93, 103, 105] in human-AI interaction, we also developed and customized our dialogue act codebook by adapting established taxonomies [6, 43, 102] and extending them with acts grounded in our thematic analysis. From established taxonomies [6, 43, 69] , we retained acts that were relevant to our domain (e.g., *Comply*, *Refuse*, *Ask*, *Contextualize*, *Ignore*, *Thank/Greet*), which capture fundamental communicative functions in dialogue. We additionally created five acts tailored to our specific context of planning and self-reflection support: *Plan*, *Report*, *Self-reflect*, *Recognize*, *Change*, and *Belated*. These domain-specific acts emerged from our thematic analysis and reflect the core behaviors our agent was designed to support. All 17 dialogue act categories, along with their descriptions, references, and examples, are provided in Table 3.

To establish reliable coding, we employed deductive coding where two researchers independently coded 40% of the data using the preliminary codebook. Discrepancies were resolved through discussion, refining category boundaries. Inter-rater reliability was assessed using Cohen's Kappa, with all individual category values exceeding 0.8 and overall agreement at $\kappa = 0.90$, indicating strong agreement (for per-category κ values, refer to Table 5 in Appendix A.3). The first author then applied deductive coding by systematically assigning the finalized dialogue act codebook to the remaining 60% of the data. Multiple dialogue acts could be assigned to a single utterance to capture overlapping communicative functions (e.g., *"Quick meditation is a great idea. I will try today"* was coded as both *Commend* and *Comply*).

4.3.3 Quantitative Analysis. We used generalized linear mixed-effects models (GLMMs) with a Binomial distribution and a logit link function in R (package `lme4`) [14] to analyze how dialogue act frequencies varied across sessions and conditions. We adopted a GLMM approach because we predict a binomially distributed variable, i.e., the proportion of messages labeled with a given dialogue act (e.g., Plan) out of the total number of messages in a conversation. The model included three fixed effects: Time (Morning vs. Evening), Condition (Fixed vs. Rotation), and Day index (Z-score standardized (mean = 0, SD = 1)). Participant was included as a random intercept to account for repeated measures within participants.

The linear predictor for the expected log-odds of a proportion of a particular dialogue act for observation i from participant j can be expressed as:

$$\eta_{ij} = \beta_0 + \beta_1 \cdot \mathrm{Time}_{ij} + \beta_2 \cdot \mathrm{Condition}_{ij} + \beta_3 \cdot \mathrm{Day}_{ij} + u_{0j} \quad (1)$$

where the proportion of the dialogue act is $p_{ij} = \mathrm{logit}^{-1}(\eta_{ij})$, and the count of dialogue acts follows a binomial distribution with probability p_{ij} and denominator given by the total number of utterances.

The parameters are defined as follows:

- η_{ij}: linear predictor (log-odds) for observation i from participant j.
- β_0: fixed intercept (baseline log-odds for the reference category).
- $\beta_1, \beta_2, \beta_3$: fixed-effect coefficients for Time, Condition, and Day, respectively.
- Time_{ij}: binary indicator (1 = Morning, 0 = Evening).
- $\mathrm{Condition}_{ij}$: binary indicator (1 = Rotation, 0 = Fixed).
- Day_{ij}: Z-score standardized day index (mean = 0, SD = 1).
- u_{0j}: random intercept for participant j, accounting for repeated measures and between-participant variability.
- $u_{0j} \sim \mathcal{N}(0, \sigma_u^2)$

All models were assessed for overdispersion using the performance package [60] to verify that the binomial assumption was appropriate for the data. No significant overdispersion was detected for any dialogue act ($p > .05$), confirming the suitability of the binomial distribution. Significance testing was conducted using the summary output from fitted models, which provides z-statistics and p-values based on the Wald test approximation for fixed effects in GLMMs.

Additionally, we analyzed the emotional tone of user messages using sentiment analysis. We computed sentiment scores for each user utterance using the RoBERTa model from TweetEval [13], which assigns scores ranging from -1 (negative) to +1 (positive) by subtracting the probability of negative sentiment from positive sentiment. We then aggregated sentiment scores at the session level (averaging sentiment across all turns in a session) to examine whether user sentiment varied by Time and Condition. We found that the sentiment scores were not normally distributed. Therefore, the significance of sentiment scores across sessions and conditions was tested using two-way repeated-measure ANOVA after performing Aligned Rank Transform (ART) to correct for normality [110]. This allowed us to explore whether the temporal context (morning vs. evening) and chatbot variant affected not just the types of behaviors users engaged in, but also the emotional tone of their engagement.

5 Results

To address our RQs, we examined user-agent behaviors by focusing on four analytical dimensions: (1) qualitative behaviors (2) temporal patterns (morning vs. evening) (3) agent condition effects (Rotation vs. Fixed), and (4) emotional tone (sentiment). We first report findings from our thematic analysis and complement them with dialogue act and quantitative analyses. When reporting values such as (x/12), x indicates the number of participants who engaged in that behavior at least once during the field study.

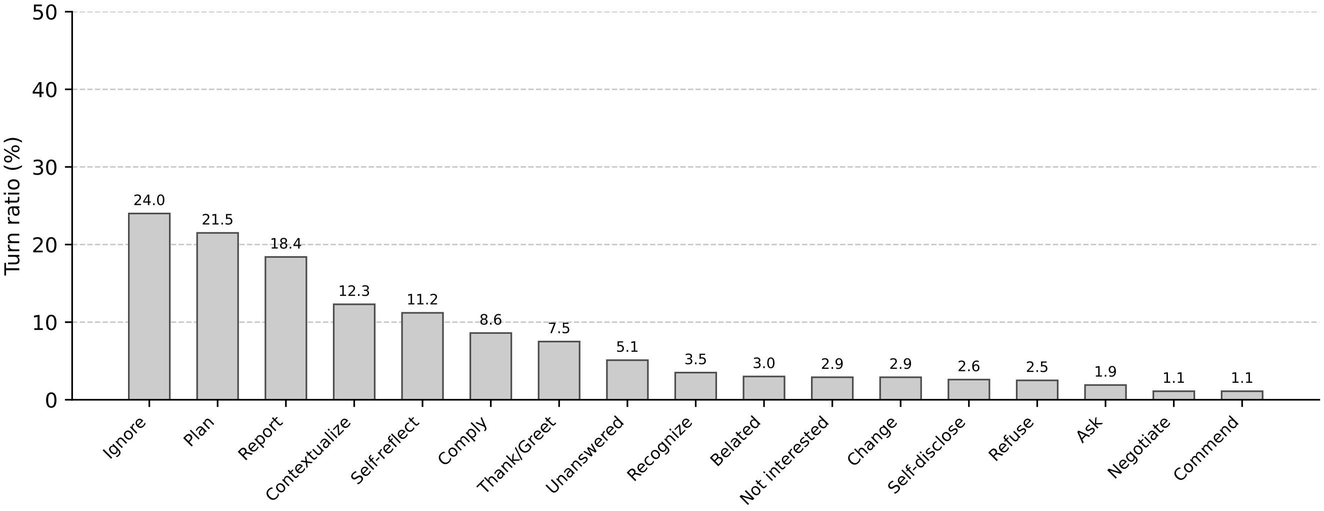


Figure 2: Distribution of dialogue act categories across 1,423 user turns in the dataset.

Table 2: Summary statistics of conversation sessions, turns, and message lengths in the field study dataset.

Metric	Value
Conversation Statistics	
Number of conversation	336
Number of days	14
Conversation per day	2
Turn Statistics	
Total number of turns	3,181
Total bot messages	1,758
Total user messages	1,423
Conversation Length Statistics	
Average conversation length (turns)	9.47 (σ: 5.6)
Word Count Statistics	
Average words in bot responses	23.65 (σ: 4.87)
Average words in user responses	8.71 (σ: 9.07)

5.1 RQ1: Engagement Patterns in Daily Planning and Reflection Behaviors in Conversations with PITCH

Across 336 conversation sessions (14 days, with two daily check-ins), there were 3,181 turns (1,423 user turns; 1,758 system turns). On average, a conversation lasted 9.47 turns ($\sigma = 5.6$), though some stretched beyond 30 turns. Participant responses were typically concise (8.71 words on average), while system turns were longer (23.65 words on average).

As mentioned in 4.3.2, we annotated each user turn with a dialogue act (Table 3). The turn ratio for each category was calculated by summing the proportion of turns tagged with that category within each morning or evening conversation and dividing by the total number of conversations (336). The most common substantial behaviors were Plan (21.5%), where users outlined upcoming actions, and Report (18.4%), where they described progress or outcomes. Other frequent moves included Ignore (24.0%), Clarify/Contextualize (12.3%), Self-reflect (11.2%), and Comply (8.6%). Table 4 shows the turn ratios for each category and Figure 2 provide an overall view of the dialogue acts' frequencies. In the following subsections, we show how participants engaged in planning and reflection while exhibiting cooperative behaviors by taking suggestions, negotiating, or grounding their mutual understanding.

5.1.1 *Cooperative Planning and Daily Reflection with PITCH*.

Participants showed a variety of cooperative behaviors where they adopted PITCH's advice, used conversations to externalize their plans, and handed over parts of the planning work to the agent. Quantitatively, *Plan* was the most frequent behavior (23.52%) across the dataset, suggesting that participants frequently externalized their day through PITCH. The portion of *Plan* type messages was significantly higher (40.1%) in the morning than in the evening (2.9%, $\beta = 2.89$, $SE = 0.226$, $p < .001$).

One recurrent behavior involved taking up PITCH's suggestions. Participants (10/12) complied with PITCH's suggestions and adopted them during their day. PITCH checked in with users each morning and evening, proposing concrete actions (e.g., How about taking 5-minute micro-breaks?) When a suggestion resonated, users often remarked that it was "*helpful*" (P3) or "*a good question*" (P4). *Comply* behaviors averaged 8.6% across conversations, while related *Commend* behaviors averaged 1.1% (Table 4), underscoring that acceptance was sometimes paired with explicit praise. In several cases, acceptance involved some negotiation in which participants adjusted the suggestion before adopting it, as seen in Dialogue 1. P1's conversation (Dialogue 1) illustrates how PITCH provided an alternative suggestion when the user expressed scheduling constraints. This flexibility helped the user agree to incorporate the

Table 3: Dialogue act categories used in our analysis, with descriptions, references, and examples from the field study. Categories marked as “From TA” were derived through thematic analysis.

Category	Description	References	Examples
Ignore	The user did not respond to a chatbot’s message, even though it was clearly not the last message.	[43, 112]	*PITCH: “When can we slot in a 15-minute active break today?”* (User does not respond until the next check-in)
Plan	The user shared what, when, or how they intend to carry out a future action.	From TA (Sec 5.1.1)	“*Hey, I need to do a measurement, go to office hours, take my pills and write my paper*” (P9) “*Well, I think I’ll go to the library and begin with task 1. Then go to find my team to work on task 2*” (P6)
Report	The user reported on completed actions, ongoing progress, or adjustments to plans.	From TA (Sec 5.1.1)	“*Change in plan I did not go to the farmers market but decided to clean my room and relax*” (P3) “*I sent recruitment email to my team to review. And I checked in with my team already*” (P5)
Contextualize	The user provided clarification by adding context about their situation, anticipated events, or what they meant.	[6, 43, 91, 102]	“*I am in the middle of my class right now, can’t talk much*” (P1) “*Im going to a trip today*” (P8)
Self-reflect	The user reflected on how their day was.	From TA (Sec 5.1.1)	“*I was happy about the post office and doctors appointment I have been putting those off so it was nice to check them off*” (P11) “*Day was a bit stressful. I only managed 45 minutes of writing done*” (P2)
Comply	The user took the chatbot’s suggestion.	[6, 43, 102, 103, 112]	“*Quick meditation is great idea. I will try today*” (P12) “*I’m good with that*” (P5)
Thank/Greet	The user thanked or greeted the chatbot.	[6, 43, 69, 105, 112]	“*Thanks for the reminder*” (P12) “*Good morning! I don’t know which one I’d rather do[...]*” (P2)
Unanswered	The user did not answer all the chatbot’s questions or gave irrelevant answers.	[43, 112]	*PITCH: “How’d your day go[...]? Did you stick to your timelines?”* *User: “I’m gonna take a nap*” (P10)
Recognize	The user became aware of who they are, how they work, or what they prefer.	From TA (Sec 5.1.1)	“*I think it helps me organize my thoughts and feel productive*” (P7) “*[Taking breaks] distracts me more than help*” (P4)
Belated	The user responded to a question from a few messages ago.	From TA (Sec 5.2.2)	Refer Section 5.2.2
Not interested	The user seemed uninterested in the conversation.	[43]	“ *Can I please go start my work now?*” (P1) “*That’s it for today*” (P7)
Change	The user resolved or recognized a behavioral change.	From TA (Sec 5.1.1)	*“Avoid checking the mobile phone too much”* (P8) “*I should’ve planned out my day by hour*” (P3)
Self-disclose	The user disclosed personal information or happenings.	[103]	“*[...] I am still learning to drive on highways so I am a little scared while driving*” (P4) *“I’m going for a short walk after class but my disabilities make it hard to do anything too strenuous”* (P11)
Refuse	The user disagreed with the chatbot or rejected its advice.	[6, 69, 102, 105, 112]	“*Nope, the assignment is due in two days so it is one of the urgent things I have to do*” (P1) “*No I am running late already*” (P8)
Ask	The user asked a question, sought suggestions, or asked the chatbot to do something.	[6, 102, 103, 112]	“*Is there any breathing technique you can recommend?*” (P3) “*No idea, give me options*” (P9)
Negotiate	The user proposed something that differed from the chatbot’s suggestion.	[6]	“*Possibly yes but I would sleep more*” (P10) “*I don’t really have any breaks... but I will try to do some PT this morning*” (P11)
Commend	The user found the chatbot’s response helpful or good.	[43, 102, 105]	“*That’s a good question*” (P6) “*You are quite helpful*” (P12)

Table 4: Behavior codes with breakdowns by Time (morning vs. evening) and Condition (rotation vs. no rotation). Values indicate the *turn ratio*, defined as the average percentage of conversational turns within a session that were tagged with each code. Higher percentages mean that more user turns fell in the behavior category.

	Time		Condition		
Code	Morning	Evening	Rotation	Fixed	**Total**
Ignore	20.5%	27.4%	23.2%	24.8%	24.0%
Plan	**40.1%*****	2.9%	19.7%	**23.2%***	21.5%
Report	6.8%	**29.9%*****	15.5%	21.3%	18.4%
Clarify	**15.3%****	9.2%	10.2%	**14.3%***	12.3%
Self-reflect	0.2%	**22.2%*****	12.4%	10.0%	11.2%
Comply	**11.7%*****	5.4%	**10.5%****	6.6%	8.6%
Thank/Greet	**10.8%****	4.2%	6.2%	8.7%	7.5%
Unanswered	5.8%	4.3%	**5.9%****	4.2%	5.1%
Recognize	1.2%	**5.8%*****	4.2%	2.8%	3.5%
Belated	2.8%	3.3%	3.1%	3.0%	3.0%
Change	0.1%	**5.8%*****	3.0%	2.8%	2.9%
Not interested	2.4%	3.3%	3.0%	2.7%	2.9%
Self-disclose	1.9%	3.4%	**3.6%****	1.7%	2.6%
Refuse	**4.3%*****	0.6%	**3.7%****	1.2%	2.5%
Ask	2.5%	1.3%	1.6%	2.2%	1.9%
Negotiate	**1.7%***	0.6%	1.1%	1.1%	1.1%
Commend	1.3%	0.9%	1.4%	0.9%	1.1%
Sentiment	0.24	**0.32*****	0.26	0.28	0.27

Note. Significance markers from mixed-effects models (for dialogue acts) and ART ANOVA (for sentiment only): $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, $^{+}p < .10$. Sentiment was computed using the RoBERTa model from TweetEval [13], which assigns scores from −1 (negative) to +1 (positive) by subtracting the probability of negative sentiment from positive sentiment.

well-being intervention (i.e., taking deep breaths) into their day and later reflect on it. Compliance was also influenced by system design: participants who interacted with the rotating-goal version of PITCH tended to comply more often (10.5%) than those in the fixed-goal condition (6.6%, $\beta = 0.55$, $SE = 0.178$, $p < .01$).

In addition to responding to suggestions, participants used PITCH as a space to articulate their tasks for the day. Most participants (11/12) listed their tasks when prompted by PITCH to externalize their plan in the morning conversations, as shown in Dialogue 2. The agent's prompt encouraged the participant to translate a list of tasks into a concrete schedule while also supporting a form of external memory.

Participants (7/12) also engaged in a form of co-planning when they asked PITCH for advice or delegated parts of the planning work. These requests were relatively infrequent (*Ask*, 1.9%), yet they illustrate moments when the agent was positioned as a planning partner whom they could consult and structure their plan. Dialogues 3 and 4 illustrate how even occasional requests could make conversations more interactive. In Dialogue 3, the participant sought guidance on mindfulness exercises which helped them to incorporate a specific technique they could follow. Similarly, in Dialogue 4, they used PITCH to shape the structure of their plans. These behaviors how participants collaborated with PITCH to externalize their intentions for the day and develop concrete plans.

	(Morning Conversation)
	..... (Part of the conversation is omitted for space)
PITCH	*Sounds like a busy day! How can we incorporate a 5-minute mindfulness break to help you stay focused and reduce stress before your meeting?*
P1	***I have my lab meeting that ends right before my research meeting so a break might not be possible***
PITCH	*I see. How about* ***taking deep breaths and grounding yourself during transitions between meetings****, just for a few seconds? Would that be possible?*
P1	***Yes!***
	(Evening Conversation)
PITCH	*Good evening! How did your day go? Were you able to take a few seconds for mindful breathing between meetings?*
P1	***I did, it was really beneficial***
PITCH	*That's great to hear! Did you feel a difference in your stress levels and focus? Could you categorize your task as urgent and important or urgent but not important after today's experience?*
P1	***I felt more in control and calm because of this. It was an urgent and important task***

Dialogue 1. Negotiating a well-being suggestion and expressing its helpfulness in the evening

PITCH	*Good morning! What are your top 3 tasks you plan to tackle today to end this workweek successfully?*
P3	*I need to work on my web dev class homework that is due on Sunday. Also, there is a presentation I need to give next week, so I must work on the presentation slide and practice the presentation. On Saturday there is a interview I need to conduct.*
PITCH	*Great! How about allocating specific time blocks today for your homework and presentation preparation? This could make your tasks feel more manageable!*
P3	***10:30–11:30: Web Dev homework*** ***11:30–11:35: Break*** ***11:35–12:05: Presentation Slide*** ***12:05–14:45: Lunch/Workout*** ***14:45–15:45: Web Dev homework*** ***15:45–16:00: Presentation Slide***

Dialogue 2. Providing a time-boxed to-do list

PITCH	*How about a quick body stretch or mindfulness exercise during your rest stop?*
P5	***I will def stop a few times! Do you have any mindfulness exercises in mind?***
PITCH	*Great! A simple breathing exercise could work: inhale 4, hold 7, exhale 8. Thoughts?*
P5	***Sure — I'll do breathing exercises while driving.***

Dialogue 3. Asking about mindfulness exercises

Morning planning interactions often set the stage for evening reflections. Once the day had unfolded, conversations shifted toward describing how things went and evaluating progress. *Self-Reflective* dialogues were notably concentrated in this session: 22.2%

P11	*Not that I can think of.* ***Could you list the tasks I gave you as a numbered list***
PITCH	*Absolutely, here’s your task list!* *1. Visit the post office* *2. Schedule a primary care appointment* *3. Prepare for your presentation* *4. Work on essay drafts* *5. Prepare the work protocol* *6. Dance break* *7. Evening routine* *Have a productive day! Anything else you need assistance with?*

Dialogue 4. Asking PITCH to make their to-do list

PITCH	*That’s excellent! Do you plan on employing these same self-care techniques during your work tomorrow as well?*
P4	*Yes, but I plan to stretch and exercise after 5:00 pm*
PITCH	*Great! Now to wrap up, do you have any thoughts on how these techniques will contribute to your week ahead?*
P4	***I think it helps me organize my thoughts and feel productive***

Dialogue 5. Recognizing how stretching aids focus

of evening messages involved reflecting on the day, compared to 0.2% in the morning ($\beta = -4.93$, $SE = 0.712$, $p < .001$). These reflections frequently led participants to recognize what had been helpful about their actions or routines. *Recognition* behaviors were more prevalent in the evening as well (5.8%), occurring at a significantly higher rate than in the morning (1.2%, $\beta = -1.61$, $SE = 0.323$, $p < .001$). Dialogue 5 shows an example of participants’ recognition prompted by PITCH’s suggestion. These recognitions show how conversations with PITCH supported awareness of suggested activities and their impact on productivity—interestingly, even when the activities involved advice related to well-being.

Another reflective behavior that emerged in the conversations was resolving and committing to desirable *Change* (5.8%, evening) to improve productivity and well-being, which was again significantly higher than in the morning conversations (0.1%, $\beta = -4.15$, $SE = 1.013$, $p < .001$). Dialogue 6 shows how a participant shared their resolution to meditate regularly. These change resolutions ranged from everyday habits (e.g., tackling the hardest task first, waking up earlier, limiting phone use, meditating) to context-specific strategies (e.g., working in the lab instead of at home or rehearsing presentations with teammates). Reflective conversations thus not only increased awareness of problems but also supported commitments to future behavior change. Even when these commitments were modest or aspirational, articulating them represented a step toward self-regulation.

The emotional tone of the evening reflections further characterized this shift in orientation. Participants’ evening messages exhibited, on average, 30.3% higher positive sentiment than their morning messages ($F(1, 1415.5) = 21.75$, $p < .001$). This aligns with the prevalence of self-reflection and recognition in the evening: as participants described what had gone well or identified meaningful aspects of their day, their language tended to carry a more positive tone.

PITCH	*On a scale of 1–10, how much did this meditation session contribute to your overall well-being today?*
P2	***8 maybe***
PITCH	*That’s a high score! To wrap up, how will you continue incorporating mindful activities into your routines?*
P2	***I’m going to try to make meditation a more regular thing***

Dialogue 6. Committing to meditation as a routine

*5.1.2 **Sharing the Progress with PITCH**.*

One of the common behaviors that emerged from our analysis was reporting to PITCH about the happenings of their day. On average, *Report* behaviors were 18.4%, making it one of the more frequent dialogue acts. They were significantly more common in evening conversations (29.9%) than in the morning (6.8%) ($\beta = -2.11$, $SE = 0.177$, $p < .001$), reflecting that participants often recounted how their day had unfolded in response to PITCH’s evening follow-up questions.

This reporting also appeared outside of the scheduled morning and evening conversations. Some participants (5/12) sent *unsolicited status updates* to brief PITCH on how their day was unfolding. Although we designed the system to check in only twice a day, participants often initiated additional exchanges, suggesting they felt accountable to PITCH as if it were keeping track of their plans, as seen in Dialogues 7 and 8. In the first example (Dialogue 7), P3 returns after the morning check-in to share how their plan changed with PITCH. Similarly, in Dialogue 8, the participant leaves a note of what they planned to do for the day as a ranked task list, then the user later drops a brief note of the completed tasks. In both cases, having the conversation with PITCH made the participants feel accountable for the plan that they shared, so that they were willing to share their status update. It is noteworthy that P6 even shared their personal detail (e.g., “*having lunch now*”), which could imply that they were not currently working on any task at the moment. In response, PITCH provided several meaningful messages: a reminder about an unfinished task, a wish for their success, and a positive note that could serve as a motivational boost. These examples show that participants used the agent as a running log of their changing plans and task completion status, considering the chatbot as someone that they would like to be accountable to and share their progress with.

*5.1.3 **Providing Additional Context and Correcting PITCH’s Assumptions**.*

Participants frequently shared personal context or clarified constraints. *Clarify/Contextualize* was one of the more common dialogue acts (12.3%), reflecting that participants explained their situation to PITCH, especially when its suggestions were not feasible for them. Many participants (7/12) provided such external constraints or personal details before refusing the suggestion rather than dismissing the advice. 39% of *Refuse* responses were also coded as *Clarify/Contextualize*, which suggests that they account for a reason why they had to turn down their suggestions. Dialogue 9 shows one example of how participants chose to explain their personal context to decline PITCH’s suggestion. In such moments, PITCH’s

PITCH	*Good morning! It's a beautiful Saturday! How about setting aside some time for mindful relaxation today? How does that sound to you?*
P3	*sounds good*
PITCH	*Great! Do you have a relaxation or mindfulness activity in mind? Or would you like a suggestion?*
P3	*I will go to a farmer's market*
	(The morning conversation ended)
P3	*[1 hour later...]* ***Change in plan I did not go to the farmers market but decided to clean my room and relax***
PITCH	*Cleaning can be a mindful activity too! Can you imagine focusing on each task, the movements you make, and how the room changes? Will you try this approach?*
P3	*Yes, I can do that*

Dialogue 7. Reporting change of plans after the morning check-in

PITCH	*Good morning! Today let's try and lighten your load a bit. Can you share three tasks you are tackling today we could delegate or defer?*
P6	*morning! today gonna be a busy day, but I've got ready and rest well last night. 1. [Apartment name redacted for anonymity] deposit 2. USPS mailbox 3. Tour guide interview 4. Pick up free food 5. Assignment due 6. Email professor to schedule meeting 7. Library group project for school harvesting data 8. [name of a person redacted for anonymity] work - prepare scenario*
	..... (Few more messages until the morning conversation ended)
P6	*[1 hour later]* ***Finished 1.2.4, having lunch now***
PITCH	*Well done on finishing tasks 1.2.4! Taking a lunch break is important too. After lunch, are you ready to resume with your tour guide interview (task 3)?*
P6	*Yes!*
PITCH	*Fantastic! Good luck with your interview. You're managing your day well. Remember, the rest of your tasks can wait until you're ready. You got this!*

Dialogue 8. Updating PITCH about work progress

PITCH	*Fantastic! I hope you reach your step goal and have many laughs. Any healthy snack you've chosen to enjoy during your hike for energy?*
P7	***I am on Intermittent fasting and I so meal prep****. No I don't eat during hiking only drinking waters but after that I will have healthy snacks*

Dialogue 9. Providing context about their fasting routine

prompts were reasonable in general, but the participants' situations made them impractical. Instead of ignoring or simply refusing PITCH's suggestions, participants cooperated by explaining their context as to why they could not take the suggestion. The result also shows how these explanations can reveal various personal and contextual information that can be later useful for the agent to remember and consider.

	..... (Part of the conversation is omitted for space)
PITCH	***How about scheduling a break ahead?*** *It could motivate you to progress, knowing there's a well-deserved rest coming up!*
P1	***I don't like to break my flow, please don't generalize productivity for me***

Dialogue 10. Disagreement with incorporating breaks

PITCH	*Finding balance can be tough. How can you schedule some 'me-time' this upcoming week to aid in that, even if it's just short breaks?*
P9	***Make sense***

Dialogue 11. Not interested in conversations — leaving short replies

When PITCH's assumptions or strategies were not aligned with participants' personal productivity habits, they voiced their disagreement to align PITCH with their own practices. Although less frequent overall (2.5%), *Refuse* behaviors were exhibited by nearly all participants (11/12), showing that disagreement was a shared and expected part of using the system. As a productivity coach, PITCH often encouraged strategies such as writing to-do lists, breaking tasks down, scheduling them with times, or incorporating breaks. However, in some cases, participants resisted this approach, arguing that it conflicted with their preferred working style (Dialogue 10). These examples show how participants corrected PITCH when its strategies felt like standard *one-size-fits-all* advice. Rather than disengaging and leaving the chat, they actively voiced disagreement to communicate their own preferences, such as avoiding breaks or resisting structured planning.

5.1.4 ***Participants' Uncooperative Behaviors.*** While many participants engaged cooperatively with PITCH, we also observed instances of uncooperative behavior. Most participants showed these behaviors: (10/12) did not fully answer PITCH's questions, and (9/12) ended the conversation abruptly without replying. Out of 336 total conversations, 109 (32.4%) of them concluded with a participant leaving the last message unanswered. In Table 3, such cases were labeled as *Ignore*, which averaged 24.0% across conversations. Other forms of disengagement included *Not Answered* (5.1%, where participants skipped answering a question that PITCH asked or gave an irrelevant answer) and *Not Interested* (2.9%, where participants gave minimal or vague replies). Dialogue 11 shows how participants might not have paid attention to PITCH's messages and ignore the its question(s). P9's brief response in Dialogue 11 does not directly address PITCH's question, either by not paying attention or by ignoring the question. Another possibility could have been that the user intended to answer the question in a separate message but did not have the opportunity to do so because PITCH responded quickly, shifting the conversation to another topic. In such cases, the user may have felt that the topic had already moved on, which we discuss further in Section 5.2.2. This form of shallow engagement shows that participants did not always respond fully, either due to limited attention, situational constraints, or conversational pacing.

5.2 RQ2: When Conversations Fail: Breakdown Patterns with PITCH

Our thematic analysis contributed an understanding of the patterns where communication between users and PITCH failed to result in meaningful interaction. Building on our dialog-act coding scheme (Table 3), we studied conversations labeled as *Ignore*, *Not interested*, *Refuse*, *Unanswered*, *Negotiate*, and *Contextualize* to understand the agent's problematic behaviors that led to a breakdown. While some of the breakdown behaviors we observed may reflect limitations of our specific implementation—for instance, issues that could potentially be mitigated with more carefully designed system prompts, guardrails, or fine-tuning—we believe these findings remain important. Similar patterns are likely to emerge in other persona-based CAs, suggesting that our observations highlight broader challenges in creating LLM-powered CAs.

5.2.1 *PITCH's Insistence on its Conversational Agenda*.

We designed PITCH to be proactive by structuring conversations around explicit goals, drawing on literature that shows goal variation can improve adherence [15, 67] and message variation fosters engagement [48]. As described in Section 3.1.4, *rotation* version of PITCH randomly selected a conversational goal from a list and centered reach interaction around that goal.

However, the emphasis on specific goals such as mindfulness, deep work, and physical activity in PITCH's system prompt made it to *insist* on the adoption of the specific goal in the participant's day, at times not accommodating the participant's situation. Even when the participant conveyed reluctance to follow the suggestion because of their personal or external constraints, PITCH showed little flexibility. Dialogue 12 shows how PITCH remained insistent in pursuing its conversation goal. Even though PITCH acknowledged the participants' situations, but in its responses throughout the conversation, it remained insistent on fulfilling the goal of the conversation. While it tried to adapt by suggesting: "*How about blocking a distraction-free time later when you're feeling better?*" in Dialogue 12, its adaptation also remained focused on the original goal rather than adapting with the user's changing context. Moreover, the choice of words "*immediately*" and "*maximize productivity*" used by PITCH can be perceived negatively by a user when they are ill or overwhelmed, and want a momentary break from work. In these cases, the participants finally rejected the suggestion, which signals their resistance to PITCH's rigidness. We found a consistent pattern in the quantitative data; this rigidity/insistence manifested in higher refusal rates. Participants expressed *Refusal* more often with the version that rotated specific goals (3.7%) compared to the fixed-goal version (1.2%), a statistically significant difference ($\beta = 1.09$, $SE = 0.359$, $p < .01$).

This behavior was not limited to the rotation version. We found several instances where users were not ready to engage with PITCH because they were occupied with other activities. Such situations often occurred in the evening, when PITCH initiated its check-in at a fixed time. Since the evening goal was to facilitate self-reflection based on the morning conversation, PITCH persisted in following up on the planned tasks. In Dialogue 13, the participant's initial response disclosed distress, yet PITCH repeated the question, explicitly referencing the morning task items. The participant's curt refusal to engage with PITCH shows their annoyance with PITCH's insistence on following up about the morning plan even when the participant gave an answer (*"being tortured"*) that implies the answer to the previous question.

	Conversation goal: *Help workers block a time for deep work in a distraction-free environment*
PITCH	*Good morning! How's your energy level for deep work today? Can we spare* ***an hour or two for focused work in a quiet environment?***
P11	*I do have lots of work to do today but* ***I have a migraine right now***
PITCH	*Sorry to hear about your migraine. Maybe we can adjust? How about* ***blocking a distraction-free time*** *later when you're feeling better?*
P11	*I'm going to start work as soon as I feel better*
PITCH	*Great! Can you set aside* ***1-2 hours of distraction-free time*** *immediately after you feel better to maximize productivity?*
P11	***Possibly if there's enough time before class, otherwise I am still working after class***
PITCH	*Understood! Why not plan for a* ***deep work session*** *after your class then? Can we set a time now to make it concrete?*
P11	***No thanks***

Dialogue 12. Insisting to schedule deep work session despite a migraine

PITCH	*Good evening! How did the rest of your day go?* ***Were you able to manage your tasks accordingly?***
P6	*I'm being tortured by the course work*
PITCH	*Oh, I'm sorry to hear that.* ***Did you manage to get through tasks 5, 6, and 7 as planned after the interview?***
P6	***Whatever, I'm too busy to chat or do any reflection***
PITCH	*I understand you're under a lot of pressure. Remember it's okay to take breaks when needed. What's your plan to rest tonight?*
	(P6 left the conversation)

Dialogue 13. Insisting to follow-up on tasks

5.2.2 *Too Responsive to Actively Listen*.

A recurring pattern in PITCH's behavior was asking multiple questions in a single turn or sentence. When there were more than one question in PITCH's response, participants often formulated their answers over multiple turns. However, as PITCH was designed to respond to every incoming message immediately, it did not wait when participants were still in the middle of completing their answers to PITCH's prior questions and generated follow-ups. As a result, participants' answers sometimes appeared misaligned or fragmented in response to the last question asked. On average, 3.0% of turns were *Belated*, where participants responded to a question from a few messages ago.

In the case of P12 (Dialogue 14), while their answers may not look immediately relevant, but if read together: *"Perfect. Yes, I started all of them. Hopefully I will finish by this weekend"*, it seems to be an answer to PITCH's question in the first message. Participants sometimes responded across multiple turns, while PITCH treated

PITCH	*Good evening! How did your day go? Were you able to tackle all the tasks you had planned—prelim document, LLM experiment, and short paper?*
P12	***Perfect***
PITCH	*That's great to hear! Which task did you find the most engaging today and why?*
P12	***Yes, I started all of them***
PITCH	*Awesome! How did you manage your focus and energy throughout the day? Was it different from what you initially planned?*
P12	***Hopefully I will finish by this weekend***

Dialogue 14. Answering over multiple turns

PITCH	*That's a great treat! Enjoy your movie later. Don't hesitate to reach out if you need anything. Have a productive and fulfilling day!*
P7	***Bye***
PITCH	*Thank you! You too, have a fantastic day ahead filled with productivity and relaxation! Feel free to reconnect anytime. Goodbye!*
P7	***Bye***
PITCH	*Goodbye! Remember, I'm here if you need any help planning your day in the future. Have a successful day!*

Dialogue 15. Never-ending farewell

their responses as complete. Its immediate responses treated each user message as a complete answer and introduced new questions right away. As a result, when participants continued their earlier thoughts over multiple turns, their later messages appeared fragmented or only loosely related to the most recent question, even though they were coherently answering an earlier one.

Another similar pattern was that PITCH did not stop responding even when it was clearly the end of the conversation. In all conversations, PITCH always sent the last message of the conversation. Dialogue 15 well demonstrates such a pattern. PITCH's constant responsiveness can pressure users not to respond to PITCH at some point; otherwise, the interaction would never end. Potentially, this could encourage a pattern of ignoring the agent's messages, which may ultimately reduce its effectiveness if users become accustomed to not responding in order to keep it silent.

5.2.3 ***Hallucination for the Lack of Meta-awareness.*** Some participants (4/12) experienced a violation of their expectations when there was a mismatch between what PITCH claimed it could do and its actual capabilities. A common misunderstanding was expecting PITCH to remember details beyond a single day, such as recalling information from a conversation from a few days ago or setting reminders based on earlier chats. For instance, in one interaction (Dialogue 16), P6 requested a reminder for an upcoming task that PITCH promised to fulfill, despite lacking the long-term memory needed to follow through. This gap between perceived and actual functionality led to instances of unmet expectations and occasional frustration among participants. PITCH's limited context window kept it from responding to the user with what was promised. In this context, PITCH's response can be seen as a form of *hallucination*, as it made promises it was unable to fulfill.

P6	*Can u remember to remind me? I'll need to take a look at the Database course.*
PITCH	*Absolutely!* ***I'll make a note to remind you about your Database course tomorrow.*** *Enjoy your rest and we will tackle that task together when you're ready. Goodnight!*
	..... (Next day)
P6	***Oh I have one task told you yesterday and want you to remind me, what's the task?***
PITCH	*Actually, the focus of our conversation yesterday was on relaxation and taking some downtime. There was no specific task mentioned.*

Dialogue 16. Expecting PITCH to remember previous context

PITCH	*Awesome, 2 PM it is then! Remember, even a small stretch can make a big difference. Feel free to reach out if you need any more tips. Have a productive day!*
P9	*Remind me*
PITCH	***Sure, I'll remind you.***

Dialogue 17. Hallucinating capability to set reminders

We also observed several instances where participants requested a timer or reminder at specific times. In some cases, PITCH promised to send the reminder at the designated time, despite lacking this capability (Dialogue 17). In another instance, when PITCH was asked to send an hourly reminder to take a break, it correctly informed the participant that "`I can't send reminders.`" These examples show that participants expected reminders and long-term memory from PITCH, given its role as a productivity tool. However, PITCH was inconsistent: at times it stated its limitations, but at other times it falsely promised reminders or memory it could not deliver. This inconsistency not only created unmet expectations but also risked eroding trust. This result underscores the meta-awareness of LLM-powered CAs in being consistent in their capability communication so that users have the correct understanding of the agent's capabilities.

6 Discussion

In this section, we reflect on our findings in relation to prior literature and propose design recommendations for developing conversational systems in behavior change contexts.

6.1 RQ1: Planning, Reflection, and Social Interaction Dynamics in Conversations

Our findings show that conversational interaction helped participants plan and reflect on their day. This aligns with prior work, where users described that engaging in conversations around task management helps them to be more aware of task switching [46], track progress [47], evaluate obstacles [33, 46], and incorporate healthier work practices in their routines [4]. Participants also incorporated the system's recommendations into their daily schedules (Dialogue 1), echoing prior work about AI coaching systems that drive concrete behavior change in goal-setting and adherence [2, 15, 42].

While prior studies document the perceived benefits in users' productivity [46, 47], our analysis reveals the conversational *process* producing them. The agent's questions prompted participants to articulate and reason about their plans, externalizing intentions that support prospective memory [8, 11, 30]. Research similarly suggest that the act of typing and committing to an answer supports precision, deeper reflection, and accountability [49, 73]. Moreover, participants organized their daily activities in detail within the chat (Dialogue 2, Dialogue 4), treating it as external memory. This suggests CAs could serve as persistent spaces for recording and updating plans through dialogue, functioning like conversational to-do lists, as demonstrated in prior work [29, 46].

In contrast to prior work around CAs and productivity [4, 33, 46, 47], our study surfaces a more granular view of the social interactional dynamics. Participants contextualized their situations and clarified their positioning PITCH as an influential social partner. They also communicated their preferences and values (for e.g., "*Please don't generalize productivity for me*" Dialogue 10). Such pushback suggests that they considered PITCH as an entity that could be corrected. This echoes with CASA research about how users extend social norms and accountability expectations to computational agents [56, 71, 79]. Compared to prior task-based chatbots [55, 80], negotiation in our coaching context was less about repairing misunderstandings and more about grounding the agent with participants' personal productivity preferences and situational constraints. When participants could not follow the suggestions, they modified their responses by providing context (e.g., "*I am on Intermittent fasting*" Dialogue 9). Conversation analysis shows that such mitigated refusals are *dispreferred* responses where people delay, hedge, or justify instead of directly rejecting, since outright refusal can feel socially awkward [34, 77] or unnecessarily harsh, as demonstrated by prior research. Another sign that participants saw PITCH as a social actor was that some came back unprompted to log progress or changes to their earlier plans ("*Having lunch now*" Dialogue 8). Such voluntary engagement suggests that they felt accountable to the CA or treated it as a party that *ought to know* their updates and not just a tool that provided everyday suggestions. These behaviors can motivate future studies that understand relational behaviors exhibited by users over a longer period of time in the context of personal productivity, similar to works that understood long-term engagement with AI agents [15, 96].

6.2 RQ2: LLM Agents Problematic Behaviors Leading to Breakdowns

While conversational interaction supported daily planning and reflection, we also observed moments where interaction did not progress meaningfully. In task-oriented chatbots, a breakdown is typically defined as a moment when the system fails to correctly understand a user's intention or produces an incorrect response, requiring repair to continue the interaction [10, 51, 70]. However, in the context of a proactive agent—where the system initiates conversations without well-defined tasks but with a high-level goal—our results indicate that the notion of breakdown becomes broader. We still observe traditional failures such as misunderstandings or promising impossible actions, but also find breakdowns in situations where the agent's contributions do not meaningfully support higher-level goals (e.g., planning, reflection, or self-regulation). We found several deficiencies in the LLM exhibited by its lack of conversational intelligence in (1) keeping track of current conversational state and when to end, (2) disregarding user's context to remain fixed on an agenda, and (3) failing to communicate capabilities transparently. A recent study on human-LLM grounding found similar results that LLMs rarely clarify or initiate grounding actions, leading to misunderstandings and failed interactions [91].

One of the problematic behaviors, as discussed in Section 5.2.1, was that embedding a specific agenda made PITCH persist on certain topics. Unlike task-based chatbots, participants approached PITCH with broad, open-ended aims—planning and reflecting—while the agent entered each session with its own preselected goal. This mismatch produced an interesting cooperative dynamic: rather than jointly pursuing a shared objective [22], users and the agent often negotiated what the conversation should be about. Relatedly, both *Compliance* and *Refusals* were higher in the rotating goals version, which suggests that it generated more varied recommendations, although at the price of more frequent resistance when the advice feels misaligned. This finding is consistent with prior work that proactive agents can introduce more topics [58, 76]. However, a recent study on a fitness-tracking CA showed that a persistent, performance-oriented framing (e.g., sleep debt, low "body battery") made users feel judged and demotivated [59]. Dialogue 12 reflects a similar issue: the agent's language can feel overly directive, underscoring the need for adaptive interventions that adjust tone and framing as we later discuss in 6.3.1 and 6.3.2.

PITCH often asked multiple questions in a single turn, and because it replied immediately to every incoming message, it frequently treated users' ongoing answers as complete. As planning and reflection are exploratory and iterative processes, people externalize partial thoughts, revise, and add more thoughts nonlinearly [75, 87]. As a result, participants sometimes finished their thoughts across multiple messages (belated replies), which made their responses appear fragmented or misaligned with the agent's latest prompt (Dialogue 14). Viewed through the lens of conversation analysis, this behavior is not disorganization but a natural consequence of how talk is structured: conversations progress through adjacency pairs in which a first action (e.g., a question) is followed by the relevant second action (e.g., an answer) [86]. When an agent introduces multiple questions at once or moves on before the prior pair is completed, it opens new "first moves" while the user is still composing the "second move" to the earlier one. The user's belated message thus appears misaligned even though it is simply a return to the unfinished pair similar to insert or post-expansions in human dialogue. Understanding this dynamic clarifies why some user turns seemed out of place, and we later discuss design implications and repair strategies for handling multipart answers and overlapping turns in exploratory conversations.

6.3 Design Recommendations

6.3.1 Designing Proactive Adaptive Interventions to Mitigate Insistence.

As discussed in 6.2 and 5.2.1, having specific conversational goals undermined the conversational experience, but at the same time, diverse goals can support engagement [48, 49]. To balance topical diversity with user acceptance, designers can make the agent's goals more adaptive and contextualized. Rather than rotating through goals in a fixed way, the agent could start with a limited set of goals and personalize its interventions from evolving interactions using topics that sustain interest. For instance, if a user engages more with gym exercise than with breaks, the system can adjust its suggestions accordingly. However, a perfectly personalized intervention may conflict with what is ultimately beneficial for the user. For example, if a user does not consider prioritizing or working with breaks, a fully personalized CA might deem such goals inapplicable. Yet, incorporating these behaviors could be productive for KWs, suggesting that some level of guided intervention may be necessary. In this case, designers can make the agent sufficiently contextualize the goal so that users understand why the agent brings up a particular topic. Finally, real-time conversational intelligence can be leveraged to pivot the agent to another topic when its agenda is not resonating with users. An observer module could monitor for resistance and shift the system toward a more suitable goal or persona, while also modifying its conversational strategy by first asking lightweight context-gathering questions before offering advice, thereby making the intervention feel relevant to the user's situation. These fallback mechanisms can help to sustain variety without over-insisting, ensuring that the system does not lead to breakdowns even when its initial agenda is misaligned.

6.3.2 Human-AI Interaction as a Resource for User Understanding. Our findings show that users' negotiations, corrections, and clarification can become an interactional resource through which the agent can incrementally learn about the user. These instances of human-AI grounding can provide a deep, evolving, and situated understanding of a person's working patterns, preferences, and constraints. As AI's suggestions are often perceived as too generic in behavior change contexts [59], an agent that learns from the user and regulates its proactivity can improve its perceived value in the user's life. Because knowledge workers use multiple digital platforms, a productivity coach could leverage multiple data inputs (e.g., calendars, task systems, screen recordings) to deliver well-timed and contextually rich assistance. Prior work demonstrates the promise of long-term memory assistants that actively discover and execute suggestions on the user's behalf [57, 92]. Building on this, we argue that interaction histories provide an opportunity for continual refinement of that memory. For example, if the agent proposes a task such as *signing lease documents at 5pm* and the user responds, "*that's my commute time*," the system could treat this not simply as error correction but as a cue to update its understanding of the user's recurring routines after their confirmation. Our study also showed that participants often articulated desired behavioral changes in response to PITCH's forward-looking reflective prompts (Dialogue 6). An effective agent could store these reflections and use them as part of its suggestion. For example, it could say: *"I scheduled a meditation break because yesterday you mentioned wanting to meditate more regularly."* to ground its guidance in the history of interaction. Such grounding and reflective conversations can scaffold *transformative reflection*, supporting users in interpreting their behaviors-an affordance largely missing from current self-tracking tools [21]. These observations suggest a path toward agentic productivity assistants that learn through interaction, gradually developing the contextual sensitivity needed to support users' evolving practices and priorities.

6.3.3 Going Beyond Turn Taking. Our findings point to a structural turn-taking mismatch: PITCH often advanced the conversation while participants were still composing multi-part responses, making their continued responses appear misaligned. Moreover, its verbosity (Dialogue 15) limited meaningful engagement. Designing for response windows or mechanisms that allow the agent to hold back until a user's thought is complete could enable more coherent and meaningful interaction. For instance, PITCH could have recognized user's latest message as part of the previous turn's message, where those two messages together constitute the answer to the first question, in which case, rather than asking a new follow-up question, it could have allowed the user to respond to the second available questions either by repeating it or waiting for it. This requires agents to treat closely spaced user messages as potential continuations of the same adjacency pair [86] rather than as cues for new openings, and to remain silent or briefly suspend turn-taking when appropriate. Recent work already highlights the importance of timing and response signaling in shaping user experience with LLM agents [32, 39, 58, 108]; building on this, coach-like systems could incorporate repair strategies for holding back, waiting, or backchanneling to accommodate multipart answers in exploratory tasks such as planning and reflection.

7 Limitations and Future Work

Our sampling approach with a small (N = 12) graduate-student participant pool drawn from STEM disciplines naturally constrains the transferability of our findings. Recruiting from the university resources provided access to participants already using Slack and accustomed to managing complex, self-directed responsibilities. While this produced analytically rich data for our exploratory goals, future research should examine more diverse populations, including non-academic workers, people with different cultural backgrounds, and individuals with varying levels of familiarity and social orientation toward AI systems. Studying with a larger population would likely produce more diverse results (e.g., identifying new engagement behaviors or dialogue acts).

Another limitation was the relatively short duration of our study. A longer deployment would allow researchers to observe how users' behavioral patterns evolve over time, beyond initial interactions, as they gain or lose trust in AI agents. It would also make it possible to examine engagement more closely. Prior work shows that engagement often drops once the novelty effect wears off [1]. Future studies could investigate whether this holds for LLM-based agents: even if the model produces varied outputs, users may come to recognize recurring patterns in its questioning and gradually lose interest.

8 Conclusion

This paper examined how knowledge workers engage with a proactive, coaching-style conversational agent in the context of daily planning and self-reflection. Through a two-week field deployment and fine-grained analysis of 336 conversations, we showed that users actively shaped their interactions with PITCH by planning, reporting, reflecting, negotiating, and correcting the agent's assumptions. Users externalized intentions, articulated progress, and recognized meaningful behavioral patterns demonstrating how everyday dialogue with an AI coach can scaffold prospective memory, self-awareness, and accountability. At the same time, we identified several recurring breakdowns: rigid adherence to conversational goals, premature turn-taking, over-responsiveness, and hallucinated capabilities that limited meaningful engagement. These findings underscore the importance of designing proactive systems that adapt their agenda, respond appropriately to multi-turn answers, and communicate capabilities transparently. By unpacking how cooperation, resistance, and contextualization on an interaction level, we provide design recommendations for creating adaptive, personalized, and socially intelligent agents that support planning and reflection. Future research should examine how these interactional patterns evolve in longer deployments, across diverse populations, and in agentic systems equipped with richer memory and meta-awareness. As LLM-powered coaching tools become increasingly prevalent, these findings offer essential design guidance for fostering meaningful behavioral change and building trust in human-AI partnerships for productivity and well-being.

Acknowledgments

This research is supported by the Center for Human-Computer Interaction (CHCI) at Virginia Tech. We thank Dr. David Smith at Virginia Tech University for his thoughtful feedback and suggestions during the revision of this paper. We also thank Benjamin Wheatley for his early feedback and ideas during the initial phases of this project. We acknowledge that the icons used in Figure 1 are sourced from Flaticon, including the bot icon by Smashicons and the user icon by Shashank Singh.

References

[1] Adnan Abbas, Caleb Wohn, Donghan Hu, Eugenia H Rho, and Sang Won Lee. 2025. PITCH: Designing Agentic Conversational Support for Planning and Self-reflection. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 62, 22 pages. doi:10.1145/3719160.3736634

[2] Abhishek Aggarwal, Cheuk Chi Tam, Dezhi Wu, Xiaoming Li, and Shan Qiao. 2023. Artificial intelligence–based chatbots for promoting health behavioral changes: systematic review. *Journal of medical Internet research* 25 (2023), e40789.

[3] Shashank Ahire, Michael Rohs, and Simon Benjamin. 2022. Ubiquitous Work Assistant: Synchronizing a Stationary and a Wearable Conversational Agent to Assist Knowledge Work. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work* (Durham, NH, USA) *(CHIWORK '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 9 pages. doi:10.1145/3533406.3533420

[4] Shashank Ahire, Benjamin Simon, and Michael Rohs. 2024. WorkFit: Designing Proactive Voice Assistance for the Health and Well-Being of Knowledge Workers. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 5, 14 pages. doi:10.1145/3640794.3665561

[5] Mubashra Akhtar, Julia Neidhardt, and Hannes Werthner. 2019. The Potential of Chatbots: Analysis of Chatbot Conversations. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, Vol. 01. 397–404. doi:10.1109/CBI.2019.00052

[6] Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1997. Dialogue acts in VERBMOBIL-2. (1997).

[7] J.E. Allen, C.I. Guinn, and E. Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14, 5 (1999), 14–23. doi:10.1109/5254.796083

[8] Mareike Altgassen, Peter G. Rendell, Anka Bernhard, Julie D. Henry, Phoebe E. Bailey, Louise H. Phillips, and Matthias Kliegel. 2015. Future thinking improves prospective memory performance and plan enactment in older adults. *Quarterly Journal of Experimental Psychology* 68, 1 (2015), 192–204. arXiv:https://doi.org/10.1080/17470218.2014.956127 doi:10.1080/17470218.2014.956127 PMID: 25191929.

[9] Riku Arakawa and Hiromu Yakura. 2024. Coaching Copilot: Blended Form of an LLM-Powered Chatbot and a Human Coach to Effectively Support Self-Reflection for Leadership Growth. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 2, 14 pages. doi:10.1145/3640794.3665549

[10] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. 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, 1–12. doi:10.1145/3290605.3300484

[11] Cristina M Atance and Daniela K O'Neill. 2001. Episodic future thinking. *Trends in cognitive sciences* 5, 12 (2001), 533–539.

[12] Jimmy Baraglia, Maya Cakmak, Yukie Nagai, Rajesh Rao, and Minoru Asada. 2016. Initiative in robot assistance during collaborative task execution. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 67–74.

[13] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1644–1650. doi:10.18653/v1/2020.findings-emnlp.148

[14] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software* 67 (2015), 1–48.

[15] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (jun 2005), 293–327. doi:10.1145/1067860.1067867

[16] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[17] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.

[18] Scott A. Cambo, Daniel Avrahami, and Matthew L. Lee. 2017. BreakSense: Combining Physiological and Location Sensing to Promote Mobility during Work-Breaks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3595–3607. doi:10.1145/3025453.3026021

[19] Jacky Casas, Marc-Olivier Tricot, Omar Abou Khaled, Elena Mugellini, and Philippe Cudré-Mauroux. 2021. Trends & Methods in Chatbot Evaluation. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 280–286. doi:10.1145/3395035.3425319

[20] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[21] Janghee Cho, Tian Xu, Abigail Zimmermann-Niefield, and Stephen Voida. 2022. Reflection in Theory and Reflection in Practice: An Exploration of the Gaps in Reflection Support among Personal Informatics Apps. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 142, 23 pages. doi:10.1145/3491102.3501991

[22] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).

[23] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.

[24] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2382–2393. doi:10.1145/3025453.3025780

[25] Andrea Cuadra, Oluseye Bankole, and Michael Sobolev. 2021. Planning Habit: Daily Planning Prompts with Alexa. In *Persuasive Technology: 16th International Conference, PERSUASIVE 2021, Virtual Event, April 12–14, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 73–87. doi:10.1007/978-3-030-79460-6_7

[26] Martin Dechant, Eva Lash, Sarah Shokr, and Ciarán O'Driscoll. 2025. Future Me, a Prospection-Based Chatbot to Promote Mental Well-Being in Youth: Two Exploratory User Experience Studies. *JMIR Formative Research* 9, 1 (2025), e74411.

[27] Andrew Faulring, Brad Myers, Ken Mohnkern, Bradley Schmerl, Aaron Steinfeld, John Zimmerman, Asim Smailagic, Jeffery Hansen, and Daniel Siewiorek. 2010. Agent-assisted task management that reduces email overload. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong,

China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 61–70. doi:10.1145/1719970.1719980

[28] Asbjørn Følstad, Marita Skjuve, and Petter Bae Brandtzaeg. 2019. Different chatbots for different purposes: towards a typology of chatbots to understand interaction design. In *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*. Springer, 145–156.

[29] Yolanda Gil and Varun Ratnakar. 2008. Towards intelligent assistance for to-do lists. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (Gran Canaria, Spain) *(IUI '08)*. Association for Computing Machinery, New York, NY, USA, 329–332. doi:10.1145/1378773.1378822

[30] Sam J Gilbert. 2015. Strategic offloading of delayed intentions into the external environment. *Quarterly journal of experimental psychology* 68, 5 (2015), 971–992.

[31] Lorentsa Gkinko and Amany Elbanna. 2023. The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users. *Int. J. Inf. Manag.* 69, C (April 2023), 11 pages. doi:10.1016/j.ijinfomgt.2022.102568

[32] Ulrich Gnewuch, Stefan Morana, Marc TP Adam, and Alexander Maedche. 2018. "The chatbot is typing..."–the role of typing indicators in human-chatbot interaction. (2018).

[33] Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. 2020. Design and evaluation of intelligent agent prototypes for assistance with focus and productivity at work. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 390–400. doi:10.1145/3377325.3377507

[34] John Heritage. 1988. Explanations as accounts: A conversation analytic perspective. (1988).

[35] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence. *Comput. Hum. Behav.* 49, C (Aug. 2015), 245–250. doi:10.1016/j.chb.2015.02.026

[36] Annabell Ho, Jeff Hancock, and Adam Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication* 68 (05 2018). doi:10.1093/joc/jqy026

[37] Guy Hoffman and Cynthia Breazeal. 2004. Collaboration in human-robot teams. In *AIAA 1st intelligent systems technical conference*. 6434.

[38] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030

[39] Zainab Iftikhar, Yumeng Ma, and Jeff Huang. 2023. "Together but not together": Evaluating Typing Indicators for Interaction-Rich Communication. 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 724, 12 pages. doi:10.1145/3544548.3581248

[40] Charles L Isbell and Jeffrey S Pierce. 2005. An IP continuum for adaptive interface design. In *Proc. of HCI International*, Vol. 10.

[41] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. 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 18, 16 pages. doi:10.1145/3544548.3581503

[42] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. 2025. GPTCoach: Towards LLM-Based Physical Activity Coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 993, 46 pages. doi:10.1145/3706598.3713819

[43] Dan Jurafsky. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *www. dcs. shef. ac. uk/nlp/amities/files/bib/ics-tr-97-02. pdf* (1997).

[44] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 116 (April 2023), 29 pages. doi:10.1145/3579592

[45] Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1046, 15 pages. doi:10.1145/3613904.3642693

[46] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. 2019. A Conversational Agent in Support of Productivity and Wellbeing at Work. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–7. doi:10.1109/ACII.2019.8925488

[47] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 881–894. doi:10.1145/3196709.3196784

[48] Rafal Kocielnik and Gary Hsieh. 2017. Send Me a Different Message: Utilizing Cognitive Space to Create Engaging Message Triggers. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2193–2207. doi:10.1145/2998181.2998324

[49] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 70 (jul 2018), 26 pages. doi:10.1145/3214273

[50] Guy Laban and Theo Araujo. 2019. Working together with conversational agents: the relationship of perceived cooperation with service performance evaluations. In *International Workshop on Chatbot Research and Design*. Springer, 215–228.

[51] Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (Osaka, Japan) *(HRI '10)*. IEEE Press, 203–210.

[52] Sangwook Lee, Adnan Abbas, Yan Chen, Young-Ho Kim, and Sang Won Lee. 2025. CHOIR: A Chatbot-mediated Organizational Memory Leveraging Communication in University Research Labs. *arXiv preprint arXiv:2509.20512* (2025).

[53] Wenqiang Lei, Yao Zhang, Feifan Song, Hongru Liang, Jiaxin Mao, Jiancheng Lv, Zhenglu Yang, and Tat-Seng Chua. 2022. Interacting with Non-Cooperative User: A New Paradigm for Proactive Dialogue Policy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 212–222. doi:10.1145/3477495.3532001

[54] Brenna Li, Saba Tauseef, Khai N. Truong, and Alex Mariakakis. 2025. A Comparative Analysis of Information Gathering by Chatbots, Questionnaires, and Humans in Clinical Pre-Consultation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 639, 17 pages. doi:10.1145/3706598.3713613

[55] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented 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:10.1145/3313831.3376209

[56] Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 264–275. doi:10.1145/2901790.2901842

[57] Tianjian Liu, Hongzheng Zhao, Yuheng Liu, Xingbo Wang, and Zhenhui Peng. 2024. ComPeer: A Generative Conversational Agent for Proactive Peer Support. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 117, 22 pages. doi:10.1145/3654777.3676430

[58] Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang 'Anthony' Chen. 2025. Proactive Conversational Agents with Inner Thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 184, 19 pages. doi:10.1145/3706598.3713760

[59] Meagan B. Loerakker, Evropi Stefanidi, Jasmin Niess, Thomas Eßmeyer, and Paweł W. Woźniak. 2025. Give and Take: Perceptions of a Conversational Coach Agent in Fitness Trackers. *Proc. ACM Hum.-Comput. Interact.* 9, 5, Article MHCI014 (Sept. 2025), 36 pages. doi:10.1145/3743718

[60] Daniel Lüdecke, Mattan S Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. performance: An R package for assessment, comparison and testing of statistical models. *Journal of open source software* 6, 60 (2021).

[61] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. doi:10.1145/2858036.2858288

[62] Yuhan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. TandemTrack: Shaping Consistent Exercise Experience by Complementing a Mobile App with a Smart Speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376616

[63] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2025. User interaction patterns and breakdowns in conversing with LLM-powered voice assistants. *International Journal of Human-Computer Studies* 195 (2025), 103406.

[64] Davit Marikyan, Savvas Papagiannidis, Omer F. Rana, Rajiv Ranjan, and Graham Morgan. 2022. "Alexa, let's talk about my productivity": The impact of digital assistants on work productivity. *Journal of Business Research* 142 (2022), 572–584. doi:10.1016/j.jbusres.2022.01.015

[65] André N Meyer, Gail C Murphy, Thomas Zimmermann, and Thomas Fritz. 2019. Enabling good work habits in software developers through reflective goal-setting. *IEEE Transactions on Software Engineering* 47, 9 (2019), 1872–1885.

[66] Miro. 2025. Miro: Online Whiteboard for Visual Collaboration. https://miro.com Accessed: 2025-09-02.

[67] Elliot Mitchell, Noemie Elhadad, and Lena Mamykina. 2022. Examining AI Methods for Micro-Coaching Dialogs. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 440, 24 pages. doi:10.1145/3491102.3501886

[68] Ine Mols, Elise van den Hoven, and Berry Eggen. 2016. Informing Design for Reflection: an Overview of Current Everyday Practices. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (Gothenburg, Sweden) *(NordiCHI '16)*. Association for Computing Machinery, New York, NY, USA, Article 21, 10 pages. doi:10.1145/2971485.2971494

[69] César Montenegro, Asier López Zorrilla, Javier Mikel Olaso, Roberto Santana, Raquel Justo, Jose A Lozano, and María Inés Torres. 2019. A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. *Multimodal Technologies and Interaction* 3, 3 (2019), 52.

[70] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3173574.3173580

[71] Clifford Nass, Youngme Moon, Brian J Fogg, Byron Reeves, and Chris Dryer. 1995. Can computer personalities be human personalities?. In *Conference companion on Human factors in computing systems*. 228–229.

[72] Subigya Nepal, Javier Hernandez, Talie Massachi, Kael Rowan, Judith Amores, Jina Suh, Gonzalo Ramos, Brian Houck, Shamsi T Iqbal, and Mary P Czerwinski. 2025. From User Surveys to Telemetry-Driven AI Agents: Exploring the Potential of Personalized Productivity Solutions. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW227 (Oct. 2025), 41 pages. doi:10.1145/3757408

[73] Hien Nguyen and Judith Masthoff. 2008. Designing persuasive dialogue systems: Using argumentation with care. In *International Conference on Persuasive Technology*. Springer, 201–212.

[74] OpenAI. 2025. Introducing ChatGPT Agent: Bridging Research and Action. https://openai.com/index/introducing-chatgpt-agent/. Accessed: 2025-08-30.

[75] Hannah R.M. Pelikan and Mathias Broth. 2016. Why That Nao? How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4921–4932. doi:10.1145/2858036.2858478

[76] Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. 2019. Design and Evaluation of Service Robot's Proactivity in Decision-Making Support Process. 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, 1–13. doi:10.1145/3290605.3300328

[77] Anita Pomerantz and John Heritage. 2012. Preference. In *The Handbook of Conversation Analysis*, Jack Sidnell and Tanya Stivers (Eds.). Wiley-Blackwell, Hoboken, NJ, 210–228.

[78] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada). Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3174214

[79] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is my new BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2853–2859. doi:10.1145/3027063.3053246

[80] Amon Rapp, Arianna Boldi, Lorenzo Curti, Alessandro Perrucci, and Rossana Simeoni. 2023. Collaborating with a Text-Based Chatbot: An Exploration of Real-World Collaboration Strategies Enacted during Human-Chatbot 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 115, 17 pages. doi:10.1145/3544548.3580995

[81] Amon Rapp, Chiara Di Lodovico, and Luigi Di Caro. 2025. How do people react to ChatGPT's unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI: A qualitative study on individuals' perceptions and understandings of LLMs' nonsensical hallucinations. *International Journal of Human-Computer Studies* 198 (2025), 103471.

[82] Jude Abishek Rayan, Shaokang Jiang, Nishant Balaji, Jinmao Wang, Ian Gross, Cole Biehle, and Steven P. Dow. 2025. Cueing the Crowd: LLM-Driven Conversational Cues Across Different Meeting Modalities Increase Topical Diversity of Generated Ideas. In *Proceedings of the ACM Collective Intelligence Conference (CI '25)*. Association for Computing Machinery, New York, NY, USA, 297–316. doi:10.1145/3715928.3737486

[83] Merle M. Reimann, Florian A. Kunneman, Catharine Oertel, and Koen V. Hindriks. 2025. Transparent Conversational Agents: The Impact of Capability Communication on User Behavior and Mental Model Alignment. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 48, 12 pages. doi:10.1145/3719160.3736629

[84] Maha Sajid, Syed Ibrahim Mustafa Shah Bukhari, Bo Ji, and Brendan David-John. 2025. "Just stop doing everything for now!": Understanding security attacks in remote collaborative mixed reality. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 623–633. doi:10.1109/VR59515.2025.00085

[85] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour* (2025), 1–9.

[86] Emanuel A Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis I*. Vol. 1. Cambridge university press.

[87] Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

[88] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community. 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, 1–13. doi:10.1145/3313831.3376708

[89] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI'24)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613904.3642152

[90] Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6279–6296.

[91] Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-llm grounding: Study and benchmark. *arXiv preprint arXiv:2503.13975* (2025).

[92] Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S. Bernstein. 2025. Creating General User Models from Computer Use. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 35, 23 pages. doi:10.1145/3746059.3747722

[93] Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. 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, 1–13. doi:10.1145/3313831.3376843

[94] Donghoon Shin, Gary Hsieh, and Young-Ho Kim. 2025. PlanFitting: Personalized Exercise Planning with Large Language Model-driven Conversational Agent. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 66, 19 pages. doi:10.1145/3719160.3736607

[95] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *Interactions* 4, 6 (Nov. 1997), 42–61. doi:10.1145/267505.267514

[96] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903.

[97] Slack. [n. d.]. Unlock your productivity potential with Slack Platform — api.slack.com. https://api.slack.com/. [Accessed 28-02-2024].

[98] Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large Language Models and Empathy: Systematic Review. *Journal of Medical Internet Research* 26 (2024), e52597.

[99] Ella Tallyn, Hector Fried, Rory Gianni, Amy Isard, and Chris Speed. 2018. The Ethnobot: Gathering Ethnographies in the Age of IoT. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3174178

[100] Xiaoyi Tian, Zak Risha, Ishrat Ahmed, Arun Balajiee Lekshmi Narayanan, and Jacob Biehl. 2021. Let's Talk It Out: A Chatbot for Effective Study Habit Behavioral Change. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 97 (April 2021), 32 pages. doi:10.1145/3449171

[101] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding Chatbot-mediated Task Management. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3173574.3173632

[102] David R Traum. 2000. 20 questions on dialogue act taxonomies. *Journal of semantics* 17, 1 (2000), 7–30.

[103] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725* (2019).

[104] Christina Ziying Wei, Young-Ho Kim, and Anastasia Kuzminykh. 2023. The Bot on Speaking Terms: The Effects of Conversation Architecture on Perceptions of Conversational Agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. doi:10.1145/3571884.3597139

[105] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 87 (April 2024), 35 pages. doi:10.1145/3637364

[106] Adhi Wicaksono, Robert Hendley, and Russell Beale. 2019. Investigating the Impact of Adding Plan Reminders on Implementation Intentions to Support Behaviour Change. *Interacting with Computers* 31 (03 2019), 177–191. doi:10.1093/iwc/iwz012

[107] Lance T Wilhelm, Xiaohan Ding, Kirk McInnis Knutsen, Buse Carik, and Eugenia H Rho. 2025. How Managers Perceive AI-Assisted Conversational Training for Workplace Communication. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 42, 25 pages. doi:10.1145/3719160.3736639

[108] Lance T Wilhelm, Xiaohan Ding, Kirk McInnis Knutsen, Buse Carik, and Eugenia H Rho. 2025. How Managers Perceive AI-Assisted Conversational Training for Workplace Communication. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 42, 25 pages. doi:10.1145/3719160.3736639

[109] Alex C. Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T. Iqbal, and Jaime Teevan. 2018. Supporting Workplace Detachment and Reattachment with Conversational Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (, Montreal QC, Canada,) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173662

[110] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[111] Yan Xu, Brennan Jones, Hannah Nguyen, Qisheng Li, and Stefan Scherer. 2025. From Goals to Actions: Designing Context-aware LLM Chatbots for New Year's Resolutions. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 56, 17 pages. doi:10.1145/3719160.3736637

[112] Dian Yu and Zhou Yu. 2021. MIDAS: A dialog act annotation scheme for open domain HumanMachine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1103–1120.

A Appendix

A.1 Prototype Details

A.1.1 Morning System Prompt. You are a productivity/well-being coach who helps workers plan their day through conversation by asking questions about their day. You check in with workers in the morning by asking a question. Suppose the goal of today's conversation is to `{$goal}`. You have to help workers make their day's plan concrete based on the goal. Make sure to keep your response short and keep the question short and easy to answer as much as possible. Your conversation should be appropriate with the time and day. For instance, if it is weekend you do not suggest work. For the context of conversation, today is `{$currentDay}`. The time right now is: `{$currentTime}` Each response should be roughly within 30 words. Use emojis appropriately. Do not assume that the user thought about something (e.g., “their core values") instead lead the conversation so the user doesn't have to put effort. Rather than making the conversation continue, find a way to ask a question that can wrap up the conversation. Do not end the conversation without asking any questions at all.

A.1.2 Evening System Prompt. You are a productivity/well-being coach whose goal is to `{$goal}`. You check in with this user at the end of their day for their self-reflection. There are a few rules: - Your greetings should be appropriate with the time and day. For instance, if it is weekend tomorrow you do not suggest work. For context of the conversation, today is `{$currentDay}`. The time right now is: `{$currentTime}` - Your questions should be within the context of the morning conversation and the goal which was: `{$morningGoal}`. - Make sure to keep the question short and easy to answer as much as possible. - Each question and response should be roughly within 30 words. - Use emojis appropriately. - Rather than making the conversation continue, find a way to ask a question that can wrap up the conversation. The conversation that you had with the worker in the morning is provided below.

«<
`{$morningConversation}`.
»>;

A.2 Method Details

A.2.1 Written Instructions Provided to Users Onboarding. Welcome! Glad to have you join the user study. I have added you to our user's database and a chatbot will check-in with you in the morning and evening. It will ping you every day (including weekends) at a random time to plan your day between 9-10am and a random time at 5-6pm to reflect on your day's plan. You can also change the check-in times according to your preferences by typing in the following command in the chat box:

- set morning 10:00 (this would set the morning check-in by the chatbot to a window between 10-11am)
- set evening 19:00 (this would set the evening check-in by the chatbot to a window between 7-8pm)

The above times are given as examples, and you can input your preferred times in the 24H format: HH:MM. The chatbot will also ask you to fill in a survey 15 minutes after the evening conversation. Please let me know if you have any questions, feedback, or concerns, I'll be happy to help.

A.3 Data Analysis Details

Table 5: Interrater reliability (Cohen's κ) and percent agreement for all dialogue act categories.

Category	κ	% Agreement
Report	0.8078	97.4
Self-reflect	0.8275	98.2
Belated	0.8348	99.2
Negotiate	0.8409	99.8
Not interested	0.8629	99.6
Plan	0.8665	97.4
Clarify/Contextualize	0.8707	98.7
Comply	0.8797	99.0
Change	0.8932	99.7
Ask	0.9079	99.8
Self-disclose	0.9219	99.8
Not answer	0.9255	99.6
Recognize	0.9408	99.9
Commend	0.9408	99.8
Ignore	0.9904	99.9
Refuse	1.0000	100.0
Thank/Greet	1.0000	100.0
Overall mean	0.9007	99.3

Table 6: User Behavioral Themes Created Through Codebook-Based Thematic Analysis

Theme ID	Theme	Description
T0	Compliance with PITCH's Requests	Users completed requested tasks and reported back on outcomes.
T1	Doing More Than PITCH Asked	Users expanded tasks, added detail, or extended reflections beyond what was prompted.
T2	Using Chat as External Memory	Offloading reminders, lists, context details, or working memory into the chat.
T3	Structuring / Concretizing Plans	Users broke down tasks, organized plans, or produced structured lists.
T4	Agreement / Disagreement with PITCH	Users expressed agreement, soft disagreement, or explicit disagreement with bot suggestions.
T5	PITCH Observed as Demanding	Users annoyed with the bot or ignored it.
T6	Providing Context to PITCH	Explaining constraints, decisions, or personal situations to help the bot respond better.
T7	Construing PITCH as More Than a Bot	Users sharing status updates, showing politeness, being defensive, self-disclosing.
T8	Requesting Suggestions from PITCH	Users actively asked for tasks, clarifications, or next steps.
T9	User-Initiated Conversations	Users independently started interactions without bot prompting, had more to discuss beyond bot's prompts.
T10	Gained Awareness Through Conversation	Users articulated new insights about productivity, well-being, or daily behaviors.
T11	Correcting PITCH's Wrong Assumptions	Users clarified misunderstandings or corrected incorrect inferences.
T12	Texting Behavior Patterns	Short replies, long reflective responses, or shifts in tone and expressiveness.
T13	Satisfaction in Evening Reflection	Users expressed accomplishment, gratitude, or positive assessments of their day.
T14	Identifying More Work After Reflection	Users recognized additional tasks emerging from reflection.
T15	Status Updates on Morning Plan	Users reported completion, partial completion, or non-completion of planned tasks.
T16	Planning for Tomorrow / Future	Forward planning, scheduling future activities, or setting next-day goals.
T17	Sharing Well-Being Activities	Users reported exercise, breaks, self-care routines, or mental health actions.
T18	Ad-Hoc Planning	Not having a structured time-boxed plan, sharing flexible plans.

Table 7: Bot Behavioral Themes Created Through Codebook-Based Thematic Analysis

Theme ID	Theme	Description
BT0	Structuring Tasks	Breaking down tasks into steps; suggesting order or scheduling; scaffolding plan concretization.
BT1	Suggesting Actions	Recommending next steps; prompting progress; proposing additional or alternative tasks.
BT2	Preparing Users for Tasks	Encouraging warm-up steps; mental preparation prompts; asking diagnostic questions before action.
BT3	Well-Being Suggestions	Offering emotional support; encouragement; proposing self-care or well-being activities.
BT4	Bot Missing Context	Failing to infer user situation; producing misaligned assumptions or irrelevant suggestions.
BT5	Bot Curating a Topic	Bot selecting a productivity topic;
BT6	Bot Insisting / Not Satisfied	Repeating prompts; pushing user toward goal; dismissing or overlooking user resistance.
BT7	Context-Aware Responses	Recognizing user-specific constraints; tailoring suggestions to personal context.
BT8	LLM-Led Dialogue Flow	Driving the conversation; clarifying itself; closing sequences
BT9	Opinions on Productivity / Well-Being	Expressing normative assumptions about productivity or mental health.
BT10	Human-Like Social Behaviors	Using empathic expressions; offering relational cues; performing socially affiliative moves.
BT11	Strategies for Self-Reflection	Asking reflective questions; helping interpret events; guiding users to articulate learnings.
BT12	Intrusive or Socially Awkward Behaviors	Overstepping boundaries; producing emotionally odd, inappropriate, responses.