Exploring The Impact of Proactive Generative AI Agent Roles In Time-Sensitive Collaborative Problem-Solving Tasks

ANONYMOUS AUTHOR(S)

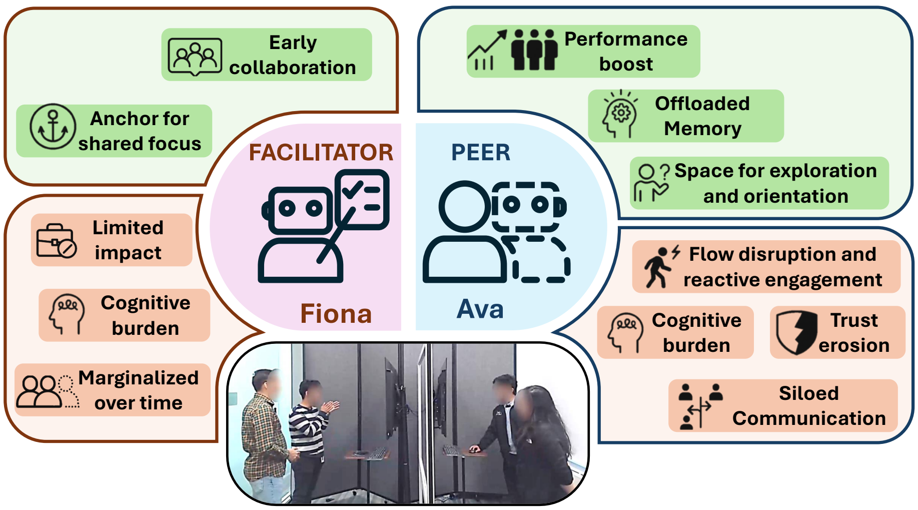


Fig. 1. Perceived influence of proactive AI agents on group performance and processes. The top panel provides an overview of sub-themes from a reflexive thematic analysis comparing a Facilitator agent ("Fiona") and a Peer agent ("Ava"). The benefits are highlighted in green, and the risks are highlighted in orange. The bottom image shows the study context, where a team is collaboratively solving a time-bounded digital escape-room task distributed across two screens.

Collaborative problem-solving under time pressure is common but difficult, as teams must generate ideas quickly, coordinate actions, and track progress. Generative AI offers new opportunities to assist, but we know little about how proactive agents affect the dynamics of real-time, co-located teamwork. We studied two forms of proactive support in digital escape rooms: a facilitator agent that offered summaries and progress cues, and a peer agent that proposed ideas and answered queries. In a within-subjects study with 24 participants, we compared group performance and processes across three conditions: no AI, peer, and facilitator. Results show that the peer agent occasionally enhanced problem-solving by offering timely hints and memory support, though it also disrupted flow and created over-reliance. In comparison, the facilitator agent provided light scaffolding but had a limited impact on outcomes. We provide design considerations for proactive generative AI agents based on our findings.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Co-located Collaboration, Generative AI, Proactive Agents, Escape Room, Group Processes

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

ACM Reference Format:
Anonymous Author(s). 2018. Exploring The Impact of Proactive Generative AI Agent Roles In Time-Sensitive Collaborative Problem-Solving Tasks. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 30 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Complex and time-sensitive problem-solving in the real world is rarely an individual task. From disaster response teams coordinating under time pressure to cybersecurity analysts mitigating an attack, group collaboration is the norm. Co-located collaboration, where team members work together in the same location and at the same time, often enhances productivity in tasks that rely on frequent communication and joint efforts, such as brainstorming, knowledge building, and planning [3, 31, 63, 102]. Prior research has shown that groups often outperform individuals because they can integrate diverse perspectives, share cognitive load, and adapt dynamically [94, 96]. Team effectiveness depends not just on member ability but on the processes of coordination, communication, and shared attention [37]. However, despite decades of CSCW and HCI research, developing technology that supports group processes and outcomes remains difficult.

Two recent research trends offer new possibilities for augmenting teamwork: generative AI and proactive systems. With advances in conversational and reasoning capabilities [24, 75, 78], generative AI has become a powerful collaborator [11, 35, 57, 77, 89]. Li et al. found that teams augmented with generative AI outperformed human-only groups across multiple performance measures [50]. Groups can also help regulate appropriate reliance, as members have opportunities to challenge, refine, and set boundaries around AI contributions [26, 39, 98]. At the same time, these systems are prone to persuasive but flawed outputs [100], leaving groups vulnerable to over-reliance, anchoring on AI suggestions, or deferring to them to avoid social conflict [12].

In parallel, research has explored how AI and intelligent systems can act proactively [32, 39, 73]. Proactive behaviors have been shown to improve trust, situational awareness, and engagement across diverse contexts, from creativity to decision support [101, 104]. Hwang et al. found that people produced a higher number and quality of ideas when their AI partner behaved as an autonomous teammate [36]. Teams often prefer AI that behaves as an active teammate, finding initiative-taking systems more supportive and peer-like. This line of work positions proactivity not just as a technical capability but as a design paradigm for how autonomous systems participate in collaboration.

Together, these trends converge in the growing CSCW and HCI framing of human–AI teams (HATs), where AI agents are understood not only as tools but as team members [9, 12, 41, 53, 91]. O'Neill et al. define HATs as "interdependence in activity and outcomes involving one or more humans and one or more autonomous agents, wherein each human and autonomous agent is recognized as a unique team member occupying a distinct role on the team, and in which the members strive to achieve a common goal as a collective." [65]. This emphasizes the importance of the roles of AI agents; these could be task-focused, such as generating ideas or solving problems, or process-focused, such as facilitating communication or maintaining shared attention [83, 91]. Generative AI has the potential to support both; its reasoning ability contributes task-specific guidance, while its conversational and summarization skills help facilitate coordination [35, 57, 101]. However, its limitations raise questions about how it should adopt such roles in practice.

While both generative AI and proactivity have shown promise independently, little is known about how they intersect in collaborative problem-solving. Should a proactive AI act as a *Peer*, contributing ideas as an imperfect teammate, or as a *Facilitator*, shaping group coordination and communication? How do these proactive roles influence not just team performance but also group processes? Addressing these questions is critical for designing AI that can effectively support collaboration in time-sensitive contexts. To explore this further, we investigate the following research questions in the context of co-located collaboration in time-sensitive problem-solving tasks:

RQ1: How do different AI agent roles (Peer vs. Facilitator) influence group performance?
RQ2: How do these AI agent roles shape group processes such as workload, communication, and coordination?

We explored these questions through the context of co-located teamwork in digital escape rooms. Escape rooms serve as a rich, high-pressure testbed for collaboration, requiring groups to solve interdependent puzzles under time constraints [45, 64]. We then developed two functional technology probes [34] in the form of generative AI agents: a facilitator, which provided discussion summaries and proposed group structures, and a peer, which contributed ideas as an imperfect teammate. The facilitator functionalities were based on previous work on such agents in group brainstorming and discussion contexts [19, 49, 102]. For the peer agent, we ran a formative study with 6 participants to understand the preferences and develop the design features. Through a within-subjects qualitative study with 24 participants (6 groups with 4 participants each), we investigate how generative AI in the role of a peer versus a facilitator impacts both group processes and performance.

Our findings highlight both the promise and pitfalls of proactive generative AI agents in group collaboration. The facilitator agent's summaries and collaboration cues initially captured attention, but its contributions were often sidelined when they became repetitive, lengthy, or poorly timed. The peer agent's thoughts and memory support sometimes boosted problem-solving, but they also risked over-reliance and disrupted flow. Importantly, teams followed varied trajectories: some relied on its thoughts before shifting to more reflective use; others' early enthusiasm led to dependence and later disillusionment; and in some, brief curiosity quickly turned into frustration and disengagement. Building on these insights, we provide design implications for tailoring facilitator and peer agents, and for supporting different trajectories of AI use in collaborative problem-solving.

In summary, we make the following three contributions:

(1) Functional technology probes of generative AI agents in facilitator and peer roles for group problem-solving tasks.
(2) An exploratory study of how proactive roles shape group processes and performance in time-sensitive collaboration.
(3) Design considerations for integrating proactive AI agents into group collaboration.

2 Background and Related Work

In this section, we review co-located collaboration, with a focus on escape rooms as a research setting, which is used to study how groups coordinate and solve problems under time pressure (Section 2.1). Next, we examine generative AI in collaborative contexts, highlighting recent shifts from reactive to proactive support (Section 2.2). Finally, we consider research on the roles of AI agents, carving out the roles of facilitator and peer (Section 2.3). By situating our work at this intersection, we show how embedding role-based proactive AI agents in escape-room settings offer new insights into both performance and group dynamics.

2.1 Co-located Collaborative Problem-Solving

Co-located synchronous collaboration describes situations where people work together in the same physical space and at the same time [20]. This mode of collaboration allows group members to share artifacts in a common workspace and benefit from subtle but important interactional cues [63]. For example, physical proximity enables coworkers to pick up on gestures, facial expressions, body posture, or shifts in attention that are often missed in remote settings. It also facilitates rapid feedback and turn-taking, helping groups quickly address misunderstandings or refine ideas as they arise. In addition, participants can easily co-reference objects in the shared environment, using gaze or pointing gestures to disambiguate expressions like "this" or "that." At the same time, these benefits come with challenges. Social evaluation pressures can make people hesitant to share ideas, and production blocking may occur when turn-taking delays individuals from voicing contributions before they forget or overthink them [60].

Escape rooms in particular offer a unique environment for studying co-located problem-solving, involving tightly-coupled interactions [84] and high synchronicity [51]. They combine the control of a laboratory study with the ecological validity of a naturalistic group task [14]. Groups must search for clues, solve puzzles, and coordinate under strict time pressure, creating a setting that naturally elicits intense interaction. They have been used for training and education purposes [23], and as platforms to study human behavior [29]. Escape rooms are well-suited for examining how groups work with AI under demanding conditions, also found in real-world settings such as disaster response, emergency medical teams, cybersecurity incident management, and air traffic control. In this paper, we extend existing literature by examining collaboration in a novel escape room setting, where groups interact with generative AI in a fast-paced, synchronous environment that amplifies both the opportunities and the challenges of teamwork with AI.

2.2 Generative AI Agents and Proactivity in Collaborative Contexts

Generative AI has already demonstrated potential to enhance creativity and problem-solving across diverse stages, such as ideation, prototyping, deliberation, and decision-making [12, 15, 30, 33, 76]. Most of this research, however, has focused on one-to-one interaction between a single user and a generative AI system. HCI studies have shown that individual improvements in AI performance do not guarantee better team outcomes, highlighting the importance of designing for team-level dynamics [74]. Only recently have studies begun to explore generative AI in multi-user, collaborative contexts. For example, researchers have investigated how groups use generative AI while co-designing [16, 46], conducting qualitative analysis [25], or engaging in cooperative play [81]. Others have examined how multiple participants jointly interact with tools like ChatGPT during group ideation [30, 76, 79], planning [102], cybersecurity vulnerability assessments [58], or creative activities such as music composition [85].

Early explorations of generative AI in teamwork have often positioned these systems as enhanced "AI-infused supertools" rather than as genuine collaborators [80]. This framing emphasizes the centrality of human expertise, where AI's role is primarily to provide support upon request. However, findings from several studies suggest that users desire AI agents that can act more like teammates, with greater initiative and autonomy [30, 46]. For example, when participants perceived an AI system as an autonomous partner, they generated more ideas and rated them as higher in quality [36]. Other work has shown that groups often prefer AI agents with stronger decision-making roles and peer-like participation, especially in creative or cooperative tasks [72, 101, 104]. Proactive communication from AI has also been found to enhance trust and situational awareness, underscoring the potential benefits of treating generative AI agents as active team members [101].

Despite this promise, most current applications of generative AI remain fundamentally reactive [27, 89, 99]. They depend on users to issue prompts or instructions before producing output. While this model lowers barriers to use, it also creates friction in collaborative contexts. Non-experts may struggle to articulate their intentions effectively, and even skilled users must divide attention between crafting prompts and participating in group discussions [27]. This additional cognitive load can disrupt conversational flow, reduce efficiency, and diminish engagement in the shared workspace [89, 90]. These limitations highlight the need for agents that can contribute more autonomously—responding to unfolding interactions without constant human direction.

Generative AI itself provides a foundation for building such proactive collaborators. Large language models (LLMs) in particular have demonstrated the capacity to simulate aspects of human cognition and social interaction, enabling them to participate more naturally in group exchanges [57, 66, 67]. Examples include LLM agents designed to play a devil's advocate role in deliberation [12], systems that surface relevant materials on shared displays to support discussion [38, 102], and AI teammates integrated into digital chat platforms [73]. Evaluations of these systems show promise but also raise concerns. Collaborative agents may unintentionally bias groups, disrupt social dynamics, or reinforce existing power imbalances [75]. Findings show that people often treat AI as a secondary partner in group discussions, particularly when the system lacks the capacity to participate fully in social dynamics [22]. Other work has highlighted a tendency for groups to over-rely on AI recommendations compared to individuals, raising concerns about dependence and reduced critical engagement [11, 100]. Early explorations of co-located settings, including prototypes of generative AI teammates in mixed reality [39] or shared displays [102], reveal both enthusiasm for their potential value and skepticism about their ability to navigate complex social interactions. Kraus et al. described proactivity as a "double-edged sword," valuable when it aligns with user needs but problematic when it intrudes or misfires [48].

However, much of what we know about proactive AI in collaboration comes from speculative design or wizard-of-oz studies that do not fully capture the complexities of functional deployment. In this paper, we address this gap by developing and testing functional probes of proactive generative AI agents. We empirically examine how teams respond to and collaborate with proactive generative AI agents in complex, time-sensitive problem-solving scenarios.

2.3 Roles of AI Agents in Collaborative Contexts

Roles are a critical lens for understanding how generative AI agents fit into collaborative contexts [77, 82, 95, 103]. In human teamwork, roles provide clarity, distribute responsibilities, and balance task-related and social demands. In the same way, when agents are introduced into groups, roles shape not only what it contributes but also how human members perceive and interact with them. Early studies have shown that people already apply social expectations to computers, treating them as legitimate teammates when interdependence exists between their actions [71, 83]. This highlights the need to carefully design and study the roles AI agents assume in group work.

Prior work has examined perceptions of AI in different social and functional roles. Kim et al. explored how people evaluated social versus functional AI, concluding that users tended to prefer functional systems, with usefulness acting as a key mediating factor [43]. However, their study was based on video demonstrations, leaving open questions about how people might respond when collaborating with functional and social agents in real tasks. Houde et al. argued that role specification could give users greater control and predictability in group brainstorming with AI, proposing roles such as responsive contributor, active reviewer, or conversation starter [32]. Liu et al. studied peer roles in children's collaborative learning and showed that framing an AI as a teammate or moderator changes conversational dynamics [52].

Bittner et al. provide a taxonomy of conversational assistant roles in collaborative work, grouping them into three categories: facilitator, peer, and expert [6]. Facilitators guide groups through structured processes, often using proactive or directive behavior grounded in scripts or models of the collaboration. They are common in contexts such as teaching, tutoring, or structured group interaction, where maintaining process flow is essential [18, 86]. Peers, in contrast, blend into the group as equals, contributing socio-emotionally while offering knowledge that is "enough but not too much." A well-designed peer agent avoids dominating discussions, encourages human contributions, and stays approachable

[17, 68]. Finally, expert roles emphasize domain-specific skill but remain largely reactive, providing help when prompted [97]. Wang et al. further distinguish task capabilities (e.g., executing, planning, evaluating) from social capabilities (e.g., coordinating, resolving conflicts, building shared understanding), clarifying what proactive agents can target [91].

Understanding the role of AI agents in collaborative problem solving also requires attention to group processes, not only outcomes [65, 101]. Group processes are the interdependent acts that transform individual inputs into collective results [37]. Communication, coordination, and workload distribution can be less visible than final performance, yet they are central to explaining team effectiveness, especially in time-sensitive contexts such as escape rooms.

In this work, we draw on these insights to focus on two roles, facilitator and peer, that occupy distinct parts of the design space. The facilitator lets us examine how an agent can structure and guide collaborative problem solving. The peer lets us explore what happens when the agent positions itself as an equal sparring partner. Unlike prior studies that place AI outside the group, assume perfect knowledge, or consider them as tools, we embed agents directly within co-located activity. By instantiating these roles as functional probes, we study how design features influence both group outcomes and the processes that mediate them.

3 Task Environment

Our study was designed around escape-room–style puzzles that serve as co-located collaborative problem-solving tasks [14, 45]. In designing the environment, we had three motivations: (1) tasks needed to require active communication and coordination among multiple group members, not passive or individual effort; (2) each puzzle had to accommodate four co-located participants; and (3) puzzles had to be adequately difficult to be engaging for at least 20 minutes for the group while remaining unsolvable by state-of-the-art multimodal generative AI models when viewed in isolation.

Puzzle Design. We created three puzzles (Puzzles 1–3), each consisting of three interconnected sub-puzzles. Sub-puzzles were distributed across two screens, such that solving them required integrating information from both displays. While individuals could walk between screens, this was slower and more effortful than communicating with teammates stationed at the other screen, thus encouraging interdependence.

There were nine unique sub-puzzles across all conditions. This avoided learning effects while maintaining a consistent "two-screen" theme. Sub-puzzles could be solved in parallel and each relied on exclusive puzzle elements, opening up possibilities for division of labor. We aimed for puzzles to last 15–25 minutes, though exact difficulty varied because solutions often depended on sudden "Aha!" moments when participants connected multiple pieces of information. Sub-puzzle designs were inspired by cooperative online puzzle games such as Acorn Cottage [1] and Alone Together [2].

We tested the puzzles against state-of-the-art multimodal and reasoning generative AI models (e.g., GPT-4.1, o3, o4) to evaluate model performance. While these models generated partial ideas by linking elements across screens, they consistently failed to produce full solutions. This reinforced our goal of designing tasks that were challenging for AI alone but could benefit from AI as a teammate, sharing partial reasoning with human collaborators.

Implementation. The puzzles were implemented using HTML, CSS, and JavaScript, and hosted on a Django server. Each puzzle's two screens corresponded to separate webpages. The countdown timer was synchronized through a backend SQLite database that stored the puzzle's status (START/STOP). Puzzle elements included both static images and interactive components. Here's an example of a sub-puzzle: in Puzzle 1 (Figures 2 and 3), Screen 2 contained green and yellow buttons. When pressed in the order shown by the Color Strip on Screen 1, these buttons triggered an "@"

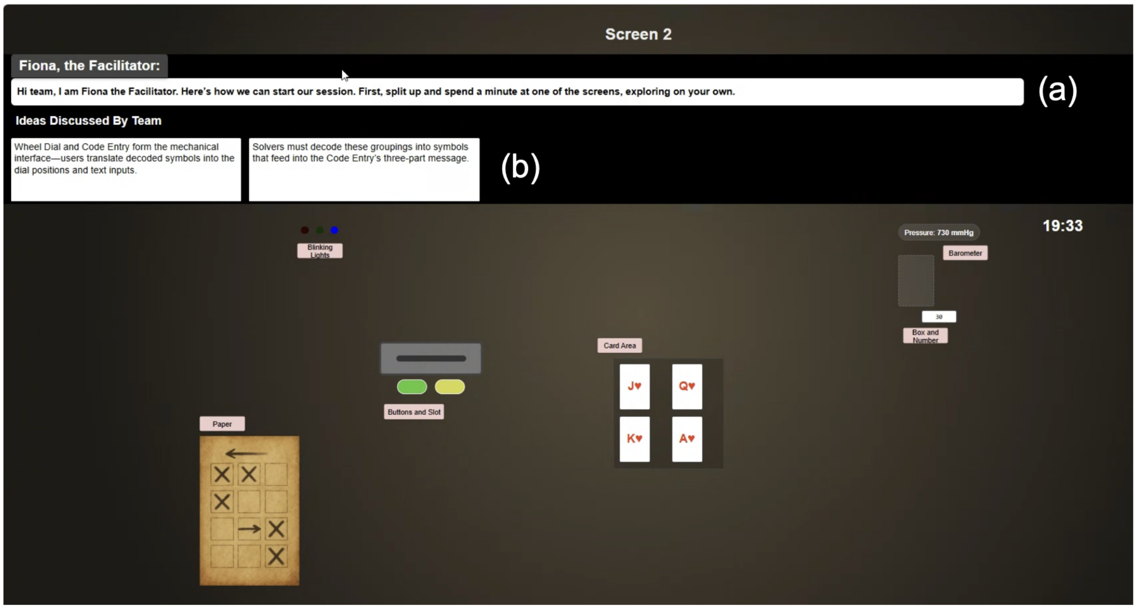


Fig. 2. Screenshot of Screen 2 of Puzzle 1 with the facilitator agent condition. There are two main features of the facilitator agent: (a) Fiona suggested structured collaboration strategies like the 1-2-4-All liberating structure [54], provided time reminders, and asked groups to divide up unsolved parts of the puzzle; (b) Fiona added the summary of ideas discussed by the group every 3 minutes.

symbol to appear in the slot above. This symbol then became a clue for the next sub-puzzle, which required linking the Symbol and Buttons element on Screen 1 with the Paper element on Screen 2.

4 Design of Generative AI Agents

We approached the agents as functional technology probes [34], designed to explore qualities of proactivity and interdependence that are central to human-AI teams [65]. Given the vast design space of proactive agents and the rapid evolution of generative AI, our probes are not intended as final or optimal solutions. Instead, they serve as design instances that help us investigate how different role configurations shape group dynamics and problem-solving processes.

To address our research questions, we set out three design goals for the agents: (1) they should work with multiple participants and take on active roles within the group, rather than acting from the outside; (2) they should not rely on perfect or pre-defined solutions, but collaborate with humans to construct answers in real time; and (3) they should act proactively, stepping in without waiting for explicit prompts. These goals align with what's currently possible with generative AI, including summarizing complex dialogue, contextualizing responses within group discussions, and supporting image-based puzzle solving. They also highlight the limitations of generative AI, including its inaccuracy and lack of social and cultural awareness.

Based on these goals, we designed two probes: *Fiona*, a process-focused facilitator; and *Ava*, a task-focused peer. Together, they represent two distinct regions of the design space for collaborative AI agents.

To ensure usability and minimize disruption, we conducted a pilot study with four participants outside of the research team. Across three sessions, the group solved two-screened puzzles under different conditions: without AI support,

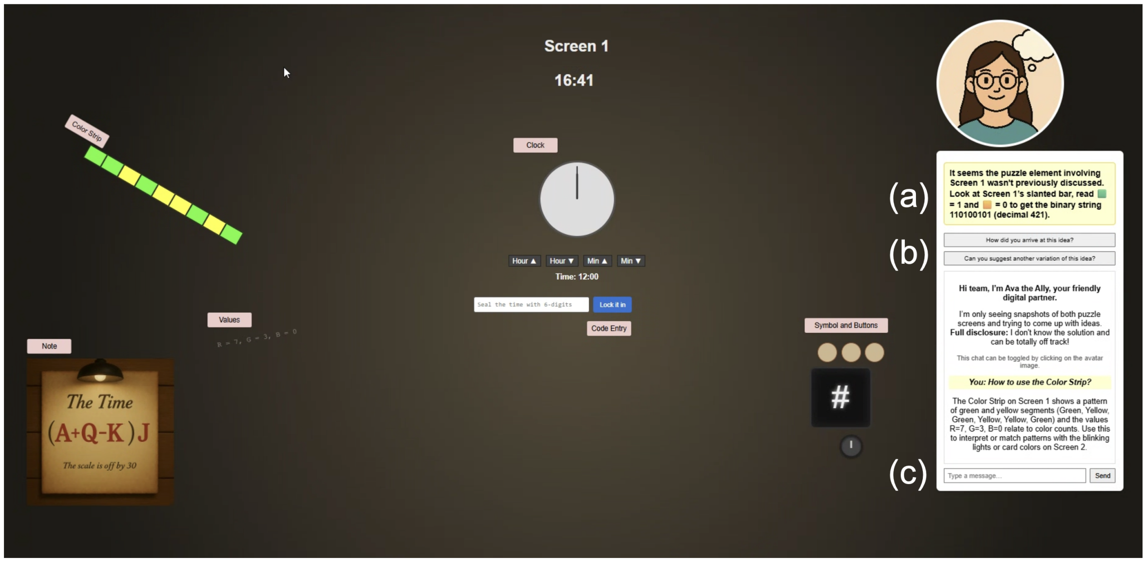


Fig. 3. Screenshot of Screen 1 of Puzzle 1 with the peer agent condition. There are three main features of the peer agent: (a) Ava proactively shared brainstorming thoughts every 3 minutes, based on puzzle screenshots and contextualized by group conversations; (b) Groups could follow up by asking Ava to explain or vary its ideas; (c) Ava was available as a chat-based partner on each puzzle screen, responding to questions.

with a facilitator agent duplicated across both screens, and with a peer agent duplicated across both screens. Each session lasted 20 minutes and was followed by a focus group interview to gather feedback on puzzle difficulty, room setup, and experiences with the agents, especially around usability. We iterated on the designs based on this feedback. In the following sections, we describe the agents' features and implementation details.

4.1 Fiona: The Facilitator Agent

Meta-cognition, or "cognition about cognition," enables groups to reflect on and regulate how they process information, approach problems, and coordinate efforts [87]. Prior work shows that groups with strong metacognitive skills are better able to monitor progress, adapt strategies, and leverage diverse perspectives, resulting in improved outcomes [61]. Two processes in particular can benefit groups: task monitoring, where groups regularly evaluate progress toward goals and adjust their approach [44], and metacognitive prompting, where questions or reminders encourage reflection on decision-making and collaboration [93]. Previous HCI research has supported these functions in contexts of group discussion and brainstorming [19, 49, 102].

Building on these insights, we designed the facilitator agent (Fiona) to scaffold groups' meta-cognitive processes during co-located problem-solving. Rather than replacing human judgment, we designed the facilitator agent to scaffold the group's own reflective capabilities, allowing them to remain aligned and adaptive to emerging ideas. The user interface of the facilitator is presented in Figure 2 along with the features. Fiona was implemented with two core features:

(1) *Group collaboration strategy prompts, time reminders, and coordination support:* The facilitator provided periodic reminders to regroup and prompted groups to divide up tasks if necessary. For example, Fiona started the session

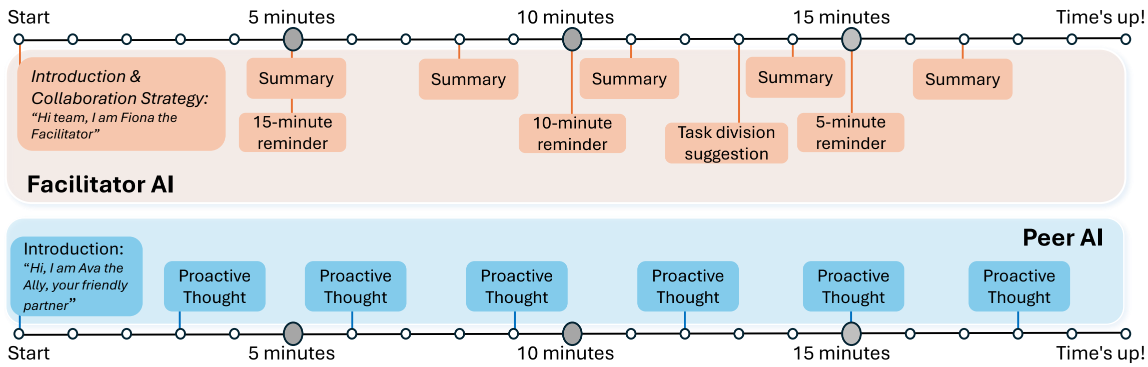


Fig. 4. Timeline of proactive interventions from the facilitator (top row) and peer (bottom) agents during the 20-minute session

by suggesting the 1-2-4-All liberating structure, a well-established facilitation technique where individuals first reflect independently, then pair up, and finally synthesize as a group [54]. This approach is relevant for small groups working across two shared displays, as it balances individual contributions with collective integration. Fiona also sent time reminders along with encouragement to keep communicating.

(2) *Periodic discussion summaries:* Every few minutes, Fiona generated concise summaries of the last segment of discussion and displayed them as idea cards. These summaries were designed to help groups step back and evaluate what had been covered, reinforcing task monitoring and ensuring shared awareness of progress.

4.1.1 Iteration based on Pilot. We found that frequent interruptions and long speech output from the facilitator disrupted the puzzle-solving process. The initial implementation used screen overlays to display up to four summaries grouped by puzzle elements, which blocked puzzle elements. It also relied on rigid countdowns to suggest collaboration structures that constrained the group's natural pacing.

In response, we removed strict timing enforcement to allow groups more flexible pacing during initial stages, eliminated screen overlays so task elements remained visible, and added a short demo/tutorial to set expectations. Fiona's language was made more concise, and the frequency of summaries was reduced to every three minutes. Instead of reading out summaries, Fiona briefly mentioned that new ideas had been added.

4.1.2 Implementation. To support time management, Fiona issued three reminders at five-minute intervals beginning at the 5-minute mark. An additional reminder at the 13-minute mark prompted groups to divide the remaining puzzle elements among members, work on them individually for one minute, and then share their ideas. Summaries were generated every three minutes, starting 5 minutes and 15 seconds after the first reminder. We present the timeline for these interventions in Figure 4. Each summary appeared as two cards highlighting the ideas discussed in the preceding three minutes, synchronized across both screens (Figure 2 (b)). Group collaboration strategy prompts, as well as reminders about time and task division, were delivered as static text in a top text box and read aloud in full using a female voice from the Edge browser (Figure 2 (a)). Summaries were generated with the multimodal GPT-4.1-mini model that used screenshots of the puzzle to ground the discussion. The ongoing dialogue between the group members was transcribed in real-time with the WhisperX model [4] and periodically stored in the database. For each summary, the model was prompted with the puzzle screenshots and the preceding three minutes of transcript. The full prompt is provided in Appendix B.1.

4.2 Ava: The Peer Agent

We designed Ava to act as a teammate who could contribute ideas without directly knowing the solution. The goal was to spark new directions and mimic peer-like collaboration rather than function as a facilitator or external advisor. This design draws on a growing body of work showing the value of AI as a creative collaborator [33, 53, 85]. For instance, prior studies have found that brainstorming with an AI partner can increase both the number and diversity of ideas compared to human-only groups [92]. Similarly, Muller et al. showed that "hybrid ideas"—those generated collaboratively between humans and AI—were more likely to be rated as the best ideas by the group [59]. At the same time, Shaer et al. cautioned that AI can sometimes overwhelm groups with excessive contributions, disrupting the collaborative flow [76].

While prior work demonstrates the potential of AI peers in brainstorming and idea generation, the escape-room context poses unique challenges. Unlike verbal brainstorming tasks, escape rooms require groups to integrate distributed visual clues, test out ideas quickly, and coordinate physical navigation between screens and members. However, no prior studies have examined how an AI peer might participate in such co-located, visual problem-solving tasks.

To better understand how Ava should interact in this setting, we conducted a formative study. Our aim was to surface user preferences for how an AI peer should contribute, when it should intervene, and how proactive its behaviors should be.

4.2.1 Formative Study. We recruited six participants for the formative study, all of whom were regular users of generative AI. At this time, we conducted individual testing to observe and gather insights on each participant's interactions and experiences. We used a ChatGPT instance with the GPT-4.1-mini model selected as the LLM to assist participants in solving the puzzle, as it's a multimodal model capable of reasoning directly over puzzle screenshots and producing fast responses. We used Puzzle 1 as the main task, which was split across two monitor screens, with ChatGPT enabled on a third monitor. To provide context, we prepared a starter prompt containing screenshots of both screens and an explanation that the puzzle elements were connected across them. Participants could then build on this prompt in their follow-up interactions with ChatGPT.

Each participant was individually tasked to solve the puzzle with the use of AI within a 20-minute time limit. Afterwards, a short semi-structured interview was conducted to gather their experiences. Based on the taxonomy for designing proactive AI agents [32], we first asked participants to talk about the AI's helpfulness and its relevance in solving the puzzle. Then we asked them to speculate on when a peer agent should contribute, including communication styles and modality, and where on the puzzle interface it should make its contributions. A summary of the common sentiments across the participants during the formative study is shown in Figure 5.

4.2.2 Design Features. Building on insights from the formative study, we designed Ava to function as a voice-enabled peer-like collaborator that sparks new directions, answers queries, and remains embedded in the group's ongoing puzzle-solving process. Ava was designed to be proactive but brief, so that it could help in time-sensitive tasks without overwhelming participants. Figure 3 shows the agent interface within Puzzle 1's Screen 2. The key features include:

(1) *Proactive Idea Contributions:* Every three minutes, Ava surfaced a new "thought" grounded in the puzzle elements visible on the two screens and contextualized by the group's recent dialogue (Fig. 3a). Ava announced, "I have a thought!" followed by a notification sound to get the attention of the group before reading out the thoughts. These contributions were intentionally framed as tentative peer ideas, rather than authoritative solutions, to maintain Ava's role as a collaborator rather than a solver. This design choice was motivated by

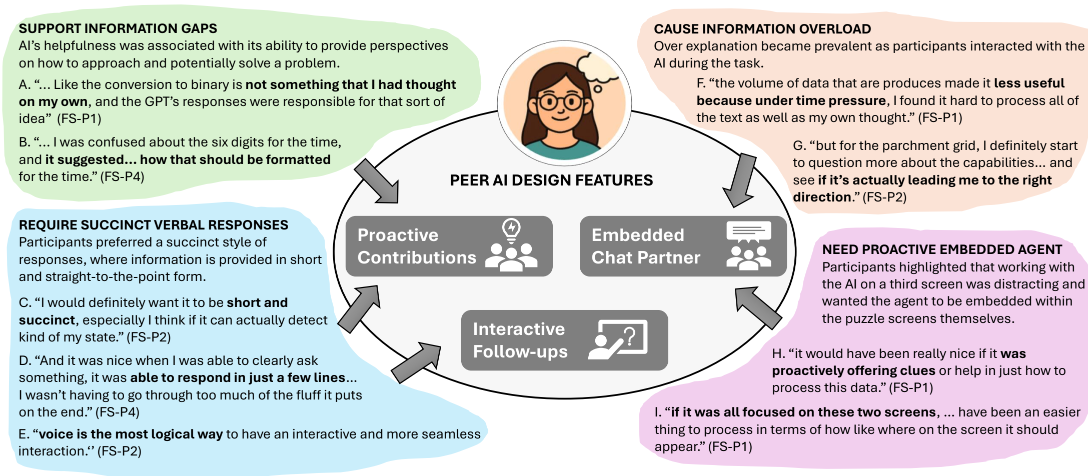


Fig. 5. Peer agent design considerations formalized from participant experiences and speculations during the formative study

formative findings that participants valued AI-generated perspectives but grew frustrated with prompting and lengthy responses. Ava's short and focused suggestions sought to trigger human reasoning without taking over the problem-solving process.

(2) *Interactive Follow-ups:* Each of Ava's ideas (Fig. 3b) included lightweight follow-up options such as "How did you arrive at this idea?" or "Can you suggest another variation?" This interactivity allowed group members to probe deeper only when they found an idea promising. This design directly responded to participants' requests for succinctness, with optional elaboration available on demand.

(3) *Embedded Chat Partner:* Ava was also available as a chat-based partner anchored within the puzzle interface (Fig. 3c). Positioning the agent directly on-screen minimized context switching between task work and AI interaction, which was a concern raised in the formative study. Ava's chat persona was framed as a "friendly digital partner" who only had access to puzzle snapshots, openly disclosing its limitations. This transparency helped set expectations and reinforced Ava's role as a peer rather than an omniscient solver. The chat feature opened up a two-way communication channel to get on-demand support.

4.2.3 Implementation. Ava presented six proactive "thoughts" to both screens during each puzzle session, delivered at three-minute intervals (shown in Figure 4). We chose this pacing to provide consistent nudges without overwhelming the group's own dialogue. The proactive thoughts were pre-generated using OpenAI's o3 model, which offered strong reasoning performance but required over a minute to generate a response. By preparing them in advance, we ensured that ideas could be delivered instantly during the session and that all groups experienced the same set of proactive contributions for comparability. The prompt for generating these thoughts is provided in Appendix B.2.

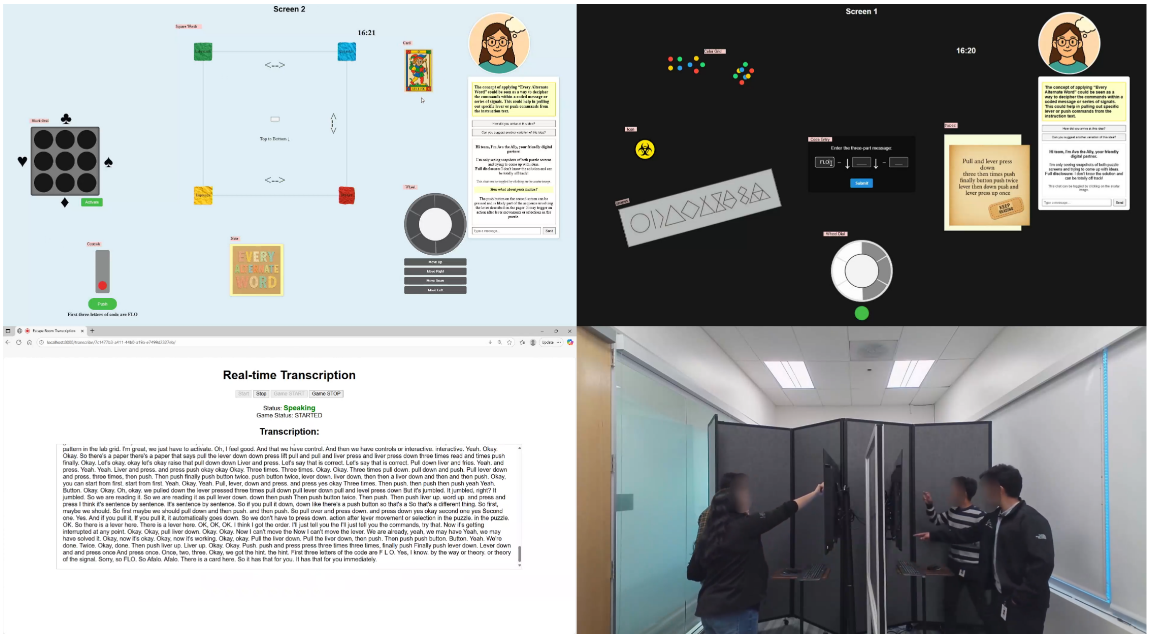


Fig. 6. Recording of Group 6, Session 2 (Puzzle 2 with the peer Agent). (1) The top panels show the two puzzle screens, each displayed on separate TVs connected to laptops. A divider between the screens increased the effort required to view the other screen, encouraging participants to communicate across displays. (2) Each participant wore a lavalier microphone, and their audio was merged into a single channel and transcribed in real time using WhisperX [4] (bottom left). These transcripts were used by the peer agent to contextualize responses and by the facilitator to generate summaries. (3) Finally, a room camera captured group interactions and overall activity throughout the session (bottom right).

We found that none of these ideas generated from this prompt gave away the solution, as the AI ideas used some of the distractions in the puzzle and couldn't guess the exact connection between the elements. For example, the first thought for Puzzle 1 was "Look at Screen 1's slanted bar, read green box = 1 and yellow box = 0 to get the binary string 110100101 (decimal 421)." Now, the binary conversion is not part of the solution, but reading the Color Strip on Screen 1 (Figure 3) can spark ideas for pressing the green and yellow buttons on Screen 2 (Figure 2) in that order. Similarly, half of the thoughts in all three puzzle conditions connected the right elements across the two screens. Therefore, the peer acted as an imperfect teammate, which was clarified before and during the start of the session. During the session, each proactive thought was contextualized in real time to match the group's ongoing discussion. This was accomplished using gpt-4o-mini, a faster model well-suited for tasks such as summarization and contextualization. Ava linked each pre-generated idea to the recent transcript summary using the prompt described in Appendix B.3.

In addition to proactive ideas, Ava also supported on-demand chat interactions. Chat interactions were powered by gpt-4.1-mini, a multimodal model capable of reasoning directly over puzzle screenshots and producing fast responses (less than 2 seconds). The prompt guiding this chat interaction is provided in Appendix B.4.

Table 1. The puzzle and AI conditions for the six study sessions

Group	Session 1	Session 2	Session 3
T1	Puzzle 2 + no AI	Puzzle 3 + facilitator AI	Puzzle 1 + peer AI
T2	Puzzle 1 + peer AI	Puzzle 3 + facilitator AI	Puzzle 2 + no AI
T3	Puzzle 3 + peer AI	Puzzle 1 + no AI	Puzzle 2 + facilitator AI
T4	Puzzle 2 + facilitator AI	Puzzle 1 + no AI	Puzzle 3 + peer AI
T5	Puzzle 1 + facilitator AI	Puzzle 2 + peer AI	Puzzle 3 + no AI
T6	Puzzle 3 + no AI	Puzzle 2 + peer AI	Puzzle 1 + facilitator AI

5 Methods

5.1 Study Design

We conducted a within-subjects user study with six groups of four participants each. Every group completed three sessions, with each session featuring a different puzzle and one of three AI conditions: no AI, peer Agent, or facilitator agent. Each session was capped at 20 minutes, providing sufficient time for groups to collaborate, interact with the AI agent (when present), and attempt to solve the puzzle.

To control for order effects such as learning, fatigue, or puzzle familiarity, we counterbalanced the condition order across groups using a six-sequence Latin square design (Table 1). This ensured that each condition appeared equally often in each position across the study and that each puzzle could be experienced with every AI condition twice. The study took place in a room where participants could move around and work on puzzles distributed across two screens, as shown in Figure 6.

5.2 Measures

We first administered a pre-study survey that included basic demographic questionnaires. In the main study, we administered surveys based on the conditions presented. When participants completed a session with an AI condition (peer or facilitator), we measured group coordination using the Perceived Coordination scale [70], subjective workload using NASA TLX [28], and the AI's impact using the AI Perception scale [5]. In the No AI condition, all surveys were administered except for the AI perception scale. We provide details of the survey scales in the Appendix D.

The performance of groups was measured by providing a score to each session based on their progress. Each puzzle had 3 sub-puzzles and was worth 5 points, with no partial points. So, there was a total of 15 points and no extra points for escaping early. We defined success in these puzzles as finding the elements that were connected across the two screens, interacting with them in a specific manner, and getting to the final solution by following through with the idea. We did not want to incentivize how fast participants got to the solution.

5.3 Participants

The user study comprised 6 groups, with 4 participants each. We recruited a total of 24 participants through voluntary convenience sampling from an internal company institution. The participant age range from 24–51 years old (M = 32.16, SD = 8.14) with 19 males and 5 females. In our participant pool, the majority of participants had little to no experience with escape room-style games, with 87% answering *Rarely* or *Never*, while 75% often used AI in their daily life and work. There was no direct compensation for participating in the study as directed by the review board in the institutional setting. The study was approved by ANONYMIZED (approval number: ANONYMIZED).

5.4 Procedure

Prior to the start of the study, we obtained informed consent from participants and subsequently administered the pre-study survey to collect basic demographic information. We then provided an overview of the study context and the overall tasks. Based on the session condition, we presented a brief overview of the AI agent that the participants will interact with before starting the puzzles. For the peer AI condition, we emphasized that the peer agent does not have the solution and can only share thoughts related to the puzzle. Similarly, we emphasized that the facilitator agent cannot provide ideas to solve the puzzle. We did not require or ask participants to use the agents explicitly. Lastly, we discussed that any external devices or internet access is not allowed during the session, and the experimenters will not be able to provide hints to solve the puzzles.

The main experiment consisted of three sessions, each with three different puzzles with a 20-minute time limit. Groupsinteracted with the AI agents based on the study design (see Table 1). After each session, the participants were provided with the main study survey questionnaires. After all the sessions were completed, we conducted a 25-30 minute focus group interview to gather the group's overall experience solving the puzzles as well as their perceptions and interaction with the different AI agents. During the focus group, we went through each of the three sessions and asked participants to describe their performance and teamwork. We followed up on how the AI agents impacted their performance and collaboration for the two AI conditions. The interview guide is provided in Appendix C. Each session lasted from 100-110 minutes, depending on the time taken to escape the puzzle rooms and the focus group durations.

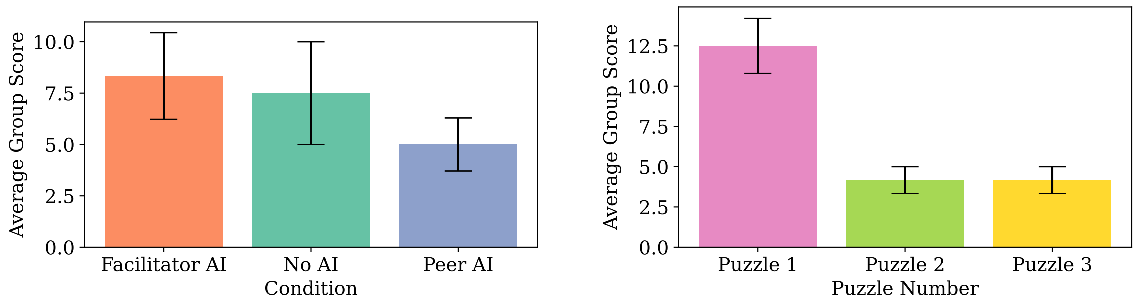
5.5 Data Collection and Analysis

We collected both quantitative and qualitative data from the six study sessions where groups worked together with and without generative AI agents. The different data sources were: (1) observation notes during the sessions and from their recordings; (2) surveys filled out by participants before the session and after experiencing each puzzle room; (3) focus-group interviews with each group.

The group interviews were conducted via Microsoft Teams using the recording and transcription features. Data quality checks on the automated transcriptions were conducted to ensure accurate translation, and any incoherent transcriptions were manually revised by the authors using the available recordings.

Our quantitative analysis drew on data from surveys after each puzzle-solving session during the study. Given the exploratory nature of the study, we report descriptive statistics. For Likert-scale responses, we present median (Md) values as measures of central tendency and interquartile range (IQR) to capture variability. For continuous variables, we report mean (M) and standard deviation (SD). Statistical inference is beyond the scope of this work, to avoid over-interpretation of the effects observed within the limited study population.

Our qualitative analysis followed the guidelines from Braun and Clarke's [7, 8] reflexive thematic analysis. We used an inductive and deductive approach to allow the codes and themes to be constructed from participants' experiences, yet still being guided by our research questions. Following the reflexive and interpretive nature of thematic analysis, we approached the analysis with the goal of building a thematic narrative that illustrates the potential influences of AI roles in collaborative group tasks. Therefore, we did not pursue inter-rater reliability since we considered the coding to be an interpretive and reflexive process rather than a fixed and stable outcome of the analysis. [55]. Following the reflexive approach from Vakeva's work [88], we also acknowledge that our interpretations were shaped by prior experiences with AI agents, which also influenced the design of the experimental conditions. Rather than treating these as biases, we view them as valuable perspectives that inform our critical interpretation of participants' accounts.



(a) Averaged by AI condition.

(b) Averaged by puzzle.

Fig. 7. Average group scores across puzzles and conditions. Error bar shows the standard error.

The thematic analysis was initiated by the first author, reading through the focus group interview transcripts and reviewing observation notes to be immersed in the context of the data. After becoming familiar with the data, the first author began extracting quotes relevant to the described experiences of the participants with the two AI roles, inductively generating initial codes to describe the extracted data. Some broad topic domains like ‘description of AI feature use’, ‘positive experiences of feature use’, ‘negative experiences of feature use’, and ‘suggested improvements to features’ were conceptualized. Afterwards, the initial codes were compared and arranged into relative topic domains that further described the variations in participant experiences with the facilitator and peer roles. Following the development of topic domains, the themes were generated in collaboration with the second author through iterative refinement and development.

6 Quantitative Findings

The facilitator condition produced the highest average score for each puzzle across groups (M = 8.33, SD = 5.16, range = 5–15), followed by the No AI condition (M = 7.50, SD = 6.12, range = 0–15). The peer condition showed the lowest overall performance (M = 5.00, SD = 3.16, range = 0–10), as shown in Figure 7a. Puzzle-level analyses show clear differences in difficulty (Figure 7b). Puzzle 1 was the easiest, with an average score of 12.5 (SD = 4.18), and four out of six groups solved it successfully. Puzzles 2 and 3 were substantially harder, with average scores of 4.17 (SD = 2.04 each). Notably, the two groups that failed Puzzle 1 encountered it under the peer condition.

We found high perceived group coordination scores throughout the sessions (Figure 8). Across conditions, participants consistently reported that teammates helped each other and coordinated smoothly, with median ratings of 4.0–5.0 on a 5-point scale. In terms of workload, average scores from the NASA-TLX survey revealed higher workload for the peer condition (M = 12.19, SD = 1.96) compared to both the facilitator (M = 10.60, SD = 2.39) and no AI conditions (M = 10.72, SD = 2.68) (Figure 9).

Based on the AI perception survey results, we found a lower rating for the facilitator compared to the peer for improving overall group score (Md = 2.0, IQR = 1.0–3.0 vs. Md = 2.5, IQR = 2.0–4.0; Figure 10). The facilitator also had a lower rating for improving group coordination (Md = 2.0, IQR = 1.75–2.25) compared to the peer (Md = 3.0, IQR = 2.0–3.25). For the peer, groups expressed widely varying perceptions, with ratings spanning from strong disagreement to agreement on whether it improved overall group score. On other dimensions, such as trustworthiness, the differences

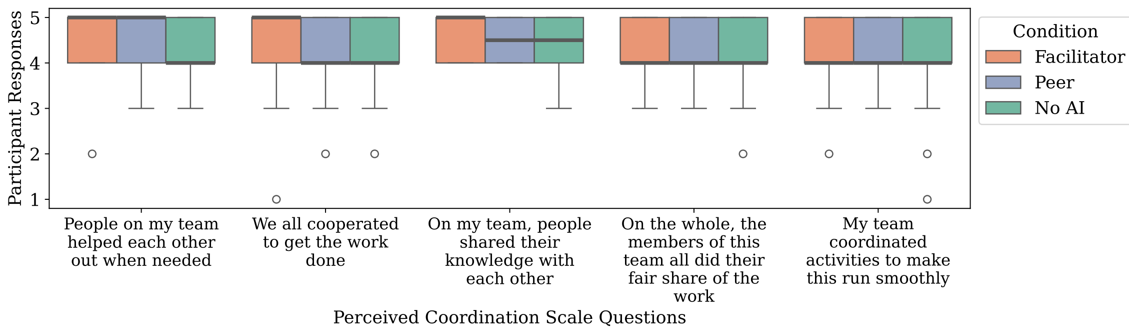


Fig. 8. Comparison of Team Coordination Survey Responses across Conditions

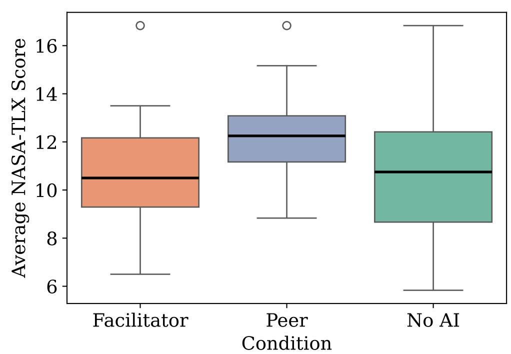


Fig. 9. Distribution of Average NASA-TLX Scores Across Conditions

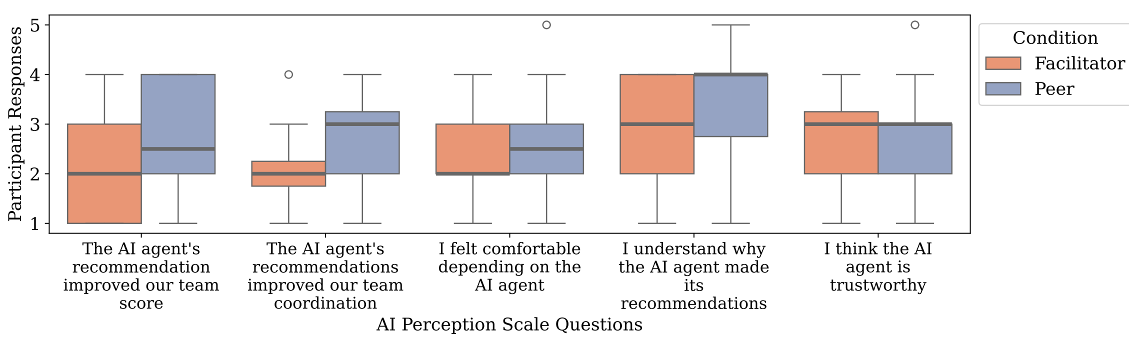


Fig. 10. Comparison of AI Perception Survey Responses between facilitator and peer conditions

between facilitator (Md = 3.0, IQR = 2.0–3.25) and peer (Md = 3.0, IQR = 2.0–3.0) were smaller, suggesting that perceptions diverged more strongly around impact on outcomes and coordination.

7 Qualitative Findings

In the following sections we present our findings from our group interviews and observations that describe their experiences with the two generative AI agents during collaborative problem-solving tasks. We conceptualize our results into themes that illustrate the influence of each agent role on their group performance and group processes. An overview of the sub-themes is presented in Figure 1.

7.1 Facilitator Agent

7.1.1 Theme 1: Provided Early Guidance and Subtle Anchors.

Structuring Early Collaboration. At the beginning of collaborative tasks, participants approached the facilitator agent with curiosity. Some groups followed its early suggestions, such as "look at other screen", even when these reiterated behaviors they were already engaged in. When the facilitator AI was introduced in the very first session (e.g., groups 4 and 5), its guidance strongly shaped how groups organized themselves. Participants treated its interventions almost like "rules," adopting practices such as rotating screens after short intervals and dividing puzzle elements among members. While these structures were not actively reinforced by the AI in later sessions, they often persisted with that workflow. Even participants in teams that did not have the facilitator in their first session mentioned that such guidance could have helped them orient more quickly. As P4 explained, "In session one, we totally had no idea about what is the setting…if it could provide something like 'split the group and focus on screen one,' that would help."

Anchors for Shared Focus. Once groups had begun to establish their own rhythm, the facilitator shifted into a quieter role: not directing the group, but anchoring its focus and maintaining structure when needed. For participants with prior experience in escape room puzzles, the facilitator's suggestions on work division and team structure were not novel, but were valued as reminders. The agent functioned as a grounding presence as P18 described, "It was more like a grounding element, like a teacher in the room or like a moderator in an exam room. Like you do your thing, but I'm there." This sense of quiet oversight was appreciated, with P18 further noting, "It gives us clear, you know, what we did, what we talked about, instead of giving us something that can lead us to somewhere else, I think this is the right amount of AI."

A typical facilitator summary looked like this: "The team discussed converting given values into a time format (hour, minute, second) and trying different button-press patterns according to arrow directions." (Group 5) Sometimes such summaries surfaced useful details that drew collective attention. P14 recalled, "I remember that moment when the AI cue with the offset 30 popped out and then everyone focused on that and we were able to solve." Subtle prompts to divide tasks also offered helpful nudges. As P22 reflected, "One time in the middle it kind of reminded us to split the task...that actually helped because we were going back and forth together."

7.1.2 Theme 2: Misaligned Support Led to Declining Use and Disengagement Over Time.

Limited Impact. Across groups, participants felt that the agent had little effect on puzzle performance. Summaries were frequently described as redundant, overly lengthy, or poorly timed—factors that limited their usefulness under time pressure. As P9 put it, "It was just saying what we already said… So I don't really think the summary helped at all." Similarly, P17 explained, "When I saw the summarization, probably that was like couple minutes ago… by then I already knew which parts I should work on." Participants often emphasized that their existing communication made the AI's inputs unnecessary.

Many groups relied more on human teammates than on the agent to coordinate work, with some comparing its role to an unneeded manager. P1 explained, “I only want to take hints from the AI or use it to remember what I’ve been saying. But…from a team process level, my teammates are more reliable.” Once communication patterns solidified, the AI’s contributions became increasingly redundant. As P15 noted, “We already had a good enough communication and collaboration from the first puzzle. So we kind of already solved that facilitation issue.”

Groups also described their interactions as primarily brainstorming-driven, which clashed with the agent’s retrospective style of summarization. P23 summarized this disconnect: “Our communication was brainstorming, not strategic. The AI summaries didn’t spark anything, they just restated what was already there.” While some acknowledged that short, targeted summaries might enhance coordination, the implementation was misaligned with the group’s fast-paced, improvisational style.

Added Cognitive Burden. Building on participants’ critiques of redundancy, some described the agent as not just unhelpful but disruptive. In more experienced groups, interventions sometimes interfered with momentum rather than supporting it. As P2 reflected, “Maybe if we were all standing around and doing nothing, then that might have helped. But…I think that actually disrupted our flow.”

Rather than lightening the workload, the agent introduced a subtle distraction when Fiona spoke up to introduce summaries. The two line summaries required extra attention at moments when participants already felt pressed for time. As P11 explained, “The time pressure gets me going, not wanting to read the AI… I would rather interact with the person who said that.”

Marginalized Over Time. We observed that initial curiosity about Fiona’s summaries quickly faded as the task progressed. While some participants glanced at outputs early on, most reported forgetting or ignoring the agent as time pressure mounted. As P7 described it, “At some point, I just forget about its existence. I kind of just ignored it.” By the latter half of puzzles, the agent was largely sidelined, with participants only occasionally glancing at the output if it appeared concise or relevant. This decline was amplified by prior negative experiences with the peer agent, which reduced trust: “My trust in the AI system was reduced because of the previous round…so I discouraged people from even looking at it.” (P22)

7.2 Peer Agent

7.2.1 Theme 1: Enhanced Problem Solving by Offering Timely Hints, Cognitive Offloading, and Exploratory Support.

Timely Hints Boosted Group Performance. Unlike the facilitator agent, Ava was remembered for its ability to steer groups toward solutions, especially at moments of impasse. Participants in Group 2 attributed much of their puzzle progress directly to its thoughts that were perceived as hints. As P5 explained, “Most of the tasks we solved were hints given by the agent… it basically directed us towards correct answers.”

The usefulness of the peer agent was closely tied to the timing of its contributions. Participants valued nudges most when momentum had stalled. As P10 summarized, “We were pursuing an idea for some time and maybe we weren’t going anywhere, and that’s exactly when it popped up… so it was useful.” Some participants also wished that the peer agent had been available in all sessions, ideally in a form that could be invoked on demand. As P22 explained, “I would have preferred an invokable AI agent…because we were stuck on a bunch of things and it was like maybe just, ‘Do you think that photo frame having three sides is relevant?”’

Offloaded Memory and Calculation Tasks. Beyond offering hints, the peer agent was valued as a dependable support for lower-level cognitive tasks, helping groups offload memory and calculation work. Participants used it to handle number-to-letter conversions, decode Morse code, recall prior details, and combine information across screens. As P2 explained, "A lot of those tasks have a long context… we had to rely on her to remember those things, but humans you probably make mistakes." Similarly, P4 emphasized its reliability in arithmetic: "I was trying to decode...doing the math, but I do it wrong… she do a really good job on remembering."

Participants also described how the peer agent supported problem-solving by confirming puzzle elements and providing targeted clarifications. For instance, P12 recalled, "It did have us identify that it is Morse code and what Morse code was saying." Others noted its usefulness in keeping the group oriented, as P11 explained: "One thing that helps is it actually reminds you some details… guides you back to what you want to focus." The system's responses also preserved parallel streams of thought by recording questions and ideas that might otherwise be forgotten. As P4 reflected, "Maybe everyone has different ideas and it's hard to memorize everyone's idea… with AI probably… that sort of records the idea [when you type out a question]."

Chat Provided Space for Orientation and Exploration. Groups engaged with Ava through text chat in varied ways, using it to orient themselves and probe puzzle elements. Some groups began by asking high-level questions—Groups 1 and 3, for instance, opened with "What are we trying to solve?" In fact, half of the groups interacted with Ava through chat even before its first proactive thought appeared around the three-minute mark.

Follow-up buttons for the proactive thoughts were used only once or twice per session. Instead, most queries focused on specific puzzle elements and their functions, often phrased with reference to their labels. Participants asked questions like, "What does the symbol between veg and triad mean?", "What are the blinking lights for?", or "Can you count the colors for us?" Groups occasionally debated an open question before typing it into the chat, sometimes duplicating the same query across both screens.

Chat was also a channel for exploration, where participants sought additional ideas after uncovering new puzzle information. For example, after manipulating grid elements in Puzzle 2 to reveal a heart symbol, P21 asked, "What do I do with a heart symbol?"

7.2.2 Theme 2: Misaligned Interactions Disrupted Flow, Added Effort, and Fragmented Communication.

Disrupted Flow and Reactive Engagement. Although the peer agent sometimes provided useful hints, participants also described how its outputs disrupted the natural flow of collaboration. Suggestions were often vague, mistimed, or irrelevant to the group's immediate focus. P14 noted, "Sometimes I see the answers from AI kind of confuse me… it points out which might be related to which, but not in a very specific way." Some wished for clearer, more directive cues, with P15 reflecting, "Whatever hints it gave, we need to interpret… probably it's better if AI just tells you like 'focus on this part."' In some cases, verbose input even made tasks feel harder than they were: "It made it look harder than what it actually was. We fell in the loop… I kind of was disappointed." (P14)

The timing of interventions often compounded this disruption, with unsolicited input breaking group momentum. P18 recalled, "Ava would come out of nowhere and be like, so if you look at this and this means this… and we're like, not right now." In several groups (2, 3, and 4), this dynamic pushed groups into a reactive mode, working collectively in response to AI suggestions rather than dividing tasks or generating independent hypotheses. As a result, collaboration became less self-directed and more tethered to interpreting Ava's outputs.

Cognitive Burden. Participants emphasized that interacting with the peer agent often introduced more effort than it saved. Typing prompts and parsing lengthy responses added friction in a fast-paced setting. As P4 explained, “Thinking in my head is faster than consolidating that and putting it in the prompt.” Others echoed concerns about the format: “The outputs right now were very long to process in a time-constrained setup.” (P8)

Participants also struggled to develop a clear mental model of the peer agent, citing unclear roles and capabilities. P9 mentioned, “Sometimes we don’t know how to ask a good question. We don’t know this AI, we don’t understand its capability.” Several noted that their limited experience with the system constrained its usefulness. P6 reflected, “If we had the AI maybe on the third one, we could interact in a better way, asking more correct questions that could be more efficient.”

Trust Erosion Through Over-Reliance and Unmet Expectations. Fiona’s interventions sometimes undermined trust within groups. Participants described how its confident outputs encouraged reliance without offering reasoning, which in turn reduced group-led problem solving. As P14 admitted, “It gave us information… but the issue is there was no reasoning behind it. It just gave us the final output, and we didn’t know if it was right. I just trusted her completely—I thought we were dumb to understand.” When such outputs proved unhelpful, the result was disappointment. P23 reflected, “I expected it to come up with something that we have not thought about... but it didn’t.”

The gap between expectations of an “intelligent” peer and the reality of inconsistent support eroded participants’ confidence in the system. As P22 summarized, “I think since most people didn’t trust it a lot, so it didn’t really add to collaboration.”

Siloed Communication and Fragmented Awareness. Compounding issues of cognitive burden and trust, the peer agent sometimes reshaped communication in ways that fragmented teamwork. When chat-based exchanges were not voiced aloud, the result was confusion and redundancy. As P1 described, “There was one time where… people had individually been talking to Ava… but without like talking out loud… then we were like, ‘oh wait, what did Ava say?”’ These private interactions created parallel conversations that left some participants out of the loop.

Several participants admitted that engaging with the AI reduced their group involvement. P16 reflected, “For me personally, I think this was the one that I was less engaged with the others… I spoke very less during this puzzle than during any of the other.” Others noted that typing to Ava pulled them away from shared context. This dynamic was often described as akin to interacting with a “fifth teammate,” but one that fragmented rather than enriched group collaboration. P3 explained, “It definitely changes the dynamic… you have this fifth person that you interact with kind of alone.”

7.2.3 Theme 3: Varied Trajectories of Agent Use Showing Contrasting Patterns of Reliance, Enthusiasm, and Disengagement.

Reliance on Hints Giving Way to Positive Reflections. In the first trajectory, we observed that groups 2 and 3 began with strong reliance on Ava’s hints. They treated its suggestions as decisive, turning to its “thoughts” when stuck. Ava often directed them toward answers, and the group followed its lead. They continued to engage with the peer throughout the session. In the focus group, these groups reflected that more deliberate engagement might have helped them collaborate effectively. As P6 recalled, "…we could have used the agent more, and the conclusion after the first experiment was that we should use the agent more. But the problem was in the next two [sessions] we didn’t end up having agent."

From Early Enthusiasm to Dependence and Disillusionment. In the second trajectory, groups 1 and 4 eagerly embraced the AI as a helpful partner, treating its early suggestions as breakthroughs. But this excitement soon shifted into dependence, with participants waiting for AI thoughts and query responses and interacting less with one another. When the AI's reasoning later fell short, the over-reliance turned into frustration and disappointment. For example, P3 described how, when Ava's first proactive thought surfaced, it seemed strikingly clever compared to the group's reasoning, prompting him to shift his attention toward the AI. But when a teammate solved the same sub-puzzle with a simpler approach, he lost trust in Ava altogether.

From Initial Curiosity to Frustration and Eventual Disengagement. The third trajectory reflected a sharper arc from curiosity to disengagement. At first, participants from groups 5 and 6 framed the AI as a potential teammate and actively explored its capabilities. Their expectations were high, and they experimented with asking questions to probe its usefulness. However, unfamiliarity with its limits soon led to confusion, and poorly timed or repetitive outputs disrupted the flow of collaboration. For example, in Group 6, P23 initially paid close attention to Ava's proactive thoughts, but when several suggestions repeated information the group had already used to solve an earlier sub-puzzle, she gradually stopped engaging with them. As such frustrations mounted, these groups redirected their attention to one another, gradually sidelining the AI and disengaging from it altogether.

8 Discussion

8.1 RQ1: How did different AI agent roles (peer vs. facilitator) influence group performance in co-located, time-sensitive problem-solving tasks?

Across sessions, the facilitator condition produced the highest average puzzle scores (Fig.7a), yet participants never credited Fiona's features for this improvement (Fig.10). Groups often described its periodic summaries and coordination nudges as redundant or poorly timed under pressure. By contrast, the peer agent (Ava) evoked polarized reactions. Some groups advanced because Ava offered timely ideas, memory support, and quick calculations, while others felt its unsolicited "thoughts" disrupted flow or slowed progress. Puzzle difficulty (Fig. 7b) and group familiarity with the escape-room format also shaped how both agents were received.

This pattern reflects the nature of our setting: short, co-located, and tightly interdependent tasks under strict time pressure. In such conditions, groups must quickly test ideas and converge on promising ones. As prior work shows, AI suggestions are more likely to be adopted when decision time is longer [10]. Time pressure made participants less receptive to global summaries or rigid scaffolds, which may be more effective in open-ended ideation or distributed work [13, 38].

The peer role both helped and hindered. It enabled progress but also anchored groups on its suggestions. Prior work shows that groups often defer to AI more than individuals do [11], with some members defending its advice or using it as a tie-breaker under load. At the same time, groups can push back when at least one member has strong contrary evidence [12]. This explains our mixed results: Ava sometimes catalyzed sensemaking, but at other times derailed parallel problem-solving when no one challenged its ideas.

The contrast between roles highlights different challenges. Fiona's metacognitive scaffolds plausibly supported coordination [62, 69] and may explain higher scores, but participants often experienced them as invisible or repetitive, since groups already shared knowledge aloud in a co-located setting. Previous research has also found that facilitation around group structures and discussion scaffolds lacks authority, which can cause them to be overlooked [52]. Ava's contributions were more visible, resembling the kinds of behaviors expected from a teammate. However, when timing or topical fit was off, Ava risked crowding out conversation, echoing findings that proactive AI teammates can overwhelm group discussions [56, 92].

8.2 RQ2: How did the AI agent roles shape group processes such as workload, communication, and coordination in co-located, time-sensitive problem-solving tasks?

NASA-TLX scores were higher with the peer than with the facilitator or no-AI conditions (Fig. 9), suggesting that any offloading it provided was outweighed by interaction and monitoring costs. The facilitator sat lightly on the conversation, offering time prompts and summaries that many groups ignored but did not find disruptive. By contrast, the peer behaved more like a "fifth teammate," injecting ideas that sometimes aided sensemaking but also diverted attention into an AI-centered side channel. Because its contributions had to be read, checked, and often queried under time pressure, the peer added to participants' workload. Prior work similarly shows that proactive, talkative agents can overwhelm groups unless their initiative is tightly governed [76]; Houde et al. also found that frequent or lengthy posts distorted discussion and argued that groups should be able to control when, what, and where an agent contributes [32]. Our findings mirror these concerns: unmanaged initiative increased attention switching and cognitive load.

Communication patterns diverged by role. The facilitator rarely intruded. Its summaries sometimes helped as an anchor, but when too long or mistimed, they felt redundant to fast and ad hoc brainstorming. Systems like LADICA explain this tension [102]. They aim to foster mutual awareness on shared displays while avoiding dominance of the human–human discussion, and they caution against features that over-clutter or steer the flow. By contrast, the peer often redirected attention to private chats with the agent, creating silos. Johnson et al. surface this as a design tension around social prominence and engagement: groups want augmentation, but they worry that agent channels will split attention and disrupt the shared workspace [39]. Our findings match the generative AI-centered interaction pattern found by Feng et al., where students engaged considerably more with the chatbot than with their peers during the collaborative problem-solving process [21].

Groups reported strong coordination across all conditions, possibly because members knew each other from before (Fig. 8). Coordination remained human-led in most sessions. Groups divided work and set rhythm with each other, treating agent input as optional. However, early facilitator nudges sometimes stuck. When the facilitator condition appeared first, its suggestions (for example, divide-and-rotate or brief individual reflection before group synthesis) helped groups set a structure that persisted. The peer agent, in contrast, sometimes shifted coordination by nudging groups away from dividing tasks and toward more joint problem-solving. In several cases, groups worked reactively around the peer's responses, staying together as a group rather than splitting work.

8.3 Design Considerations for Proactive Generative AI Agents in Colocated Time-Sensitive Problem-Solving Tasks

Our study revealed several patterns of user interaction with the AI agents, highlighting both opportunities and risks. Building on participants' suggestions and prior work, we outline concrete design considerations for proactive AI agents that can support colocated, time-sensitive, collaborative problem solving.

Summaries and *coordination cues* offered by an AI agent to guide the collaborative problem-solving task emerged as key design features. However, these same outputs risked marginalization when they became overly long, repetitive, or poorly timed. This suggests that AI agent contributions must be designed to remain concise, progress-aware, and embedded in the shared workspace rather than appearing as standalone text. Their utility increases when paired with actionable next steps grounded in the group's ongoing discussion, which participants described as critical for sustaining progress. In addition to text-based support, participants emphasized the importance of visual features that act as “external memory", such as progress indicators or expandable cards that align with the task flow while minimizing reading overhead. These preferences resonate with prior findings on large-display systems, where visual artifacts effectively scaffold group awareness without disrupting conversation [102].

Another important consideration was the design of *thoughts or suggestions* from the agent. For proactive input to be useful in time-bounded collaboration, suggestions should be kept short and accompanied by clear rationales to build user trust. Additionally, when appropriate, it may be helpful to offer multiple alternatives rather than a single idea. Such designs may help mitigate over-anchoring, promote comparative reasoning, and support more deliberate group decision-making. This also supports prior arguments that systems should intervene when users show signs of over-reliance while still respecting group autonomy and pace [40].

An agent’s *social presence* and *influence* also warrant careful design. Maintaining a moderate presence: visible but peripheral, and offering support without disrupting conversation or dominating the group’s attention, can be helpful for sustaining engagement. Participants also emphasized the value of user controls for agent initiative. Adjustable mechanisms, such as rate limits, silence thresholds, or options like “never speak first”, “pause”, or a quick “volume” dial, could give groups flexibility in determining how and when the agent participates. Such features align with broader human-centered AI principles calling for transparency and controllability in agent interjections [32, 91]. This is also line with theories of social influence suggest that preventing AI from becoming the de facto authority helps preserve critical engagement and reduces risks of social loafing or conformity [42].

Finally, *timing* and *relevance* of AI agent’s contributions proved crucial, with poorly timed or off-topic input often leading groups to abandon its support. Participants suggested that timing could be improved through activity-based triggers that respond to cues such as silence, stalls, or bursts of talk, ensuring interventions occur at appropriate moments. Relevance, in turn, could be strengthened through tighter integration into the workspace, such as anchoring suggestions to specific visual elements. Together, these features could reduce the cognitive cost of shifting attention between the agent and the task, making contributions easier to interpret in context.

9 Limitation and Future Work

While our study provides encouraging insights into how proactive generative AI can participate in real-time teamwork, several limitations also point to valuable directions for future research. First, we recruited participants through a convenience sample pool who already knew and worked closely with one another. While this familiarity likely influenced group communication and coordination, such dynamics are common in many real-world group settings, for example, in emergency response units, clinical care teams, and workplace project groups. Studying these established teams provided insight into how proactive AI agents integrate into pre-existing social dynamics. Future work can extend this approach to ad hoc teams or cross-organizational collaborations where roles and norms are less established.

Second, our work focused on a specific collaborative setting—digital escape-room puzzles with small groups in co-located, time-sensitive conditions. While this testbed offered a controlled yet engaging way to observe group communication and interaction with AI agents, it does not capture the full diversity of collaborative contexts where proactive AI agents may be deployed. Real-world domains such as healthcare, education, or crisis management involve more complex tasks, higher stakes, and longer time horizons. Expanding into these settings will help test the robustness of our findings and explore how proactive AI adapts to more varied and consequential teamwork.

Finally, the facilitator and peer agents we developed represent only two points in the much larger design space of proactive AI. Our facilitator agent was limited to summaries and reminders, while the peer agent contributed short ideas, with both agents intervening at pre-scheduled intervals. Our participants thought agents should adapt their role to the session phase and group experience: early on, offering more process guidance like a facilitator, and later, shifting to on-demand, minimal cues like a peer. Prior work has shown the importance of such signals for detecting disengagement, over-reliance, or social loafing [47]. Future work can focus on developing such context-aware and adaptive mechanisms will move proactive agents closer to being integrated teammates who flexibly support evolving group needs [91].

10 Conclusion

Our work presents an early but promising step toward understanding how proactive generative AI agents can enrich real-time, co-located teamwork. By comparing two distinct roles, a facilitator that provided summaries and structure cues, and a peer that contributed ideas and memory support, we examined how proactive AI influences not only task performance but also group processes such as workload, coordination, and communication. Our findings show that facilitators initially captured attention but were often sidelined when their input became lengthy or poorly timed, while peer agents generated more varied trajectories. Some groups used peer contributions to move from reliance toward more reflective engagement, others shifted from enthusiasm to dependence and disillusionment, and still others disengaged quickly. These patterns reveal both the promise and the fragility of proactive support in high-pressure collaboration. Taken together, our results highlight that the value of proactive generative AI lies not in static roles but in the ability to adapt—providing the right kind of support at the right moment. Designing such adaptable agents opens a pathway toward AI that participates as a trusted teammate, flexibly balancing task and process contributions to strengthen human collaboration in diverse and time-sensitive domains.

References

[1] 2025. Acorn Cottage. https://www.quarantini.space/ac-joining-instructions

[2] 2025. Alone Together: Enchambered Escape Room. https://www.enchambered.com/puzzles/alone-together/

[3] Christopher Andrews, Alex Endert, and Chris North. 2010. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 55–64. doi:10.1145/1753326.1753336

[4] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023* (2023).

[5] R. Bendell, J. Williams, Stephen M. Fiore, and F. Jentsch. 2025. Artificial social intelligence in teamwork: how team traits influence human-AI dynamics in complex tasks. *Frontiers in Robotics and AI* 12 (2025), 1487883. doi:10.3389/frobt.2025.1487883

[6] Eva Bittner, Sarah Oeste-Reiß, and Jan Marco Leimeister. 2019. Where is the Bot in our Team? Toward a Taxonomy of Design Option Combinations for Conversational Agents in Collaborative Work. doi:10.24251/HICSS.2019.035

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

[8] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (Aug. 2019), 589–597. doi:10.1080/2159676X.2019.1628806 Publisher: Routledge _eprint: https://doi.org/10.1080/2159676X.2019.1628806.

[9] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C. Nascimento. 2023. Assertiveness-based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3544548.3580682

[10] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. 2023. How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–26. doi:10.1145/3610068

[11] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3544548.3581015

[12] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. ACM, Greenville SC USA, 103–119. doi:10.1145/3640543.3645199

[13] Andy Clark. 2010. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension.* Oxford University Press.

[14] Tara N. Cohen, Andrew C. Griggs, Joseph R. Keebler, Elizabeth H. Lazzara, Shawn M. Doherty, Falisha F. Kanji, and Bruce L. Gewertz. 2020. Using Escape Rooms for Conducting Team Research: Understanding Development, Considerations, and Challenges. *Simulation & Gaming* 51, 4 (Aug. 2020), 443–460. doi:10.1177/1046878120907943

[15] Hao Cui and Taha Yasseri. 2024. AI-enhanced collective intelligence. *Patterns* 5, 11 (Nov. 2024), 101074. doi:10.1016/j.patter.2024.101074

[16] Hanhui Deng, Jianan Jiang, Zhiwang Yu, Jinhui Ouyang, and Di Wu. 2024. CrossGAI: A Cross-Device Generative AI Framework for Collaborative Fashion Design. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–27. doi:10.1145/3643542

[17] Kohji Dohsaka, Ryota Asai, Ryuichiro Higashinaka, Yasuhiro Minami, and Eisaku Maeda. 2009. Effects of conversational agents on human communication in thought-evoking multi-party dialogues. In *Proceedings of the SIGDIAL 2009 Conference.* 217–224.

[18] Gregory Dyke, David Adamson, Iris Howley, and Carolyn Penstein Rosé. 2013. Enhancing scientific reasoning and discussion with conversational agents. *IEEE Transactions on Learning Technologies* 6, 3 (2013), 240–247.

[19] Toni V. Earle-Randell, Shan Zhang, Noah Schroeder, Kristy E. Boyer, and Emmanuel Dorley. 2025. How Virtual Agents Can Shape Human-Human Collaboration: A Systematic Review. In *Artificial Intelligence in Education.* Springer, Cham, 468–486. doi:10.1007/978-3-031-98420-4_33 ISSN: 1611-3349.

[20] Clarence A. Ellis, Simon J. Gibbs, and Gail Rein. 1991. Groupware: some issues and experiences. *Commun. ACM* 34, 1 (Jan. 1991), 39–58. doi:10.1145/99977.99987

[21] Shihui Feng. 2025. Group interaction patterns in generative AI-supported collaborative problem solving: Network analysis of the interactions among students and a GAI chatbot. *British Journal of Educational Technology* 56, 5 (2025), 2125–2145. doi:10.1111/bjet.13611 _eprint: https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13611.

[22] Christopher Flathmann, Wen Duan, Nathan J. Mcneese, Allyson Hauptman, and Rui Zhang. 2024. Empirically Understanding the Potential Impacts and Process of Social Influence in Human-AI Teams. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 49:1–49:32. doi:10.1145/3637326

[23] Panagiotis Fotaris and Theodoros Mastoras. 2019. Escape rooms for learning: A systematic review. In *Proceedings of the European Conference on Games Based Learning*, Vol. 2019. 235–243.

[24] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research* 25, 3 (July 2023), 277–304. doi:10.1080/15228053.2023.2233814 Publisher: Routledge _eprint: https://doi.org/10.1080/15228053.2023.2233814.

[25] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2024. CoAIcoder: Examining the Effectiveness of AI-assisted Human-to-Human Collaboration in Qualitative Analysis. *ACM Transactions on Computer-Human Interaction* 31, 1 (Feb. 2024), 1–38. doi:10.1145/3617362

[26] Vera Hagemann and Annette Kluge. 2017. Complex problem solving in teams: the impact of collective orientation on team process demands. *Frontiers in psychology* 8 (2017), 1730.

[27] Yuanning Han, Ziyi Qiu, Jiale Cheng, and Ray Lc. 2024. When Teams Embrace AI: Human Collaboration Strategies in Generative Prompting in a Creative Design Task. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu HI USA, 1–14. doi:10.1145/3613904.3642133

[28] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[29] Casper Harteveld, Erica Kleinman, Paola Rizzo, Dylan Schouten, Truong Huy Nguyen, Samuel Liberty, Wade Kimbrough, Paul Fombelle, and Magy Seif El-Nasr. 2019. Teamwork and adaptation in games (TAG): a survey to gauge teamwork. In *Proceedings of the 14th International Conference on the Foundations of Digital Games.* ACM, San Luis Obispo California USA, 1–12. doi:10.1145/3337722.3337731

[30] Jessica He, Stephanie Houde, Gabriel E. Gonzalez, Darío Andrés Silva Moran, Steven I. Ross, Michael Muller, and Justin D. Weisz. 2024. AI and the Future of Collaborative Work: Group Ideation with an LLM in a Virtual Canvas. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '24).* Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3663384.3663398

[31] Cindy E. Hmelo-Silver and Howard S. Barrows. 2008. Facilitating Collaborative Knowledge Building. *Cognition and Instruction* 26, 1 (Jan. 2008), 48–94. doi:10.1080/07370000701798495 Publisher: Routledge _eprint: https://doi.org/10.1080/07370000701798495.

[32] Stephanie Houde, Kristina Brimijoin, Michael Muller, Steven I. Ross, Dario Andres Silva Moran, Gabriel Enrique Gonzalez, Siya Kunde, Morgan A. Foreman, and Justin D. Weisz. 2025. Controlling AI Agent Participation in Group Conversations: A Human-Centered Approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces.* 390–408. doi:10.1145/3708359.3712089 arXiv:2501.17258 [cs].

[33] Kent F. Hubert, Kim N. Awa, and Darya L. Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports* 14, 1 (Feb. 2024), 3440. doi:10.1038/s41598-024-53303-w

[34] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* 17–24.

[35] Angel Hsing-Chi Hwang, John Oliver Siy, Renee Shelby, and Alison Lentz. 2024. In whose voice?: examining AI agent representation of people in social interaction through generative speech. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference.* 224–245.

[36] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2021. IdeaBot: investigating social facilitation in human-machine team creativity. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–16.

[37] Daniel R. Ilgen, John R. Hollenbeck, Michael Johnson, and Dustin Jundt. 2005. Teams in Organizations: From Input-Process-Output Models to IMOI Models. *Annual Review of Psychology* 56, Volume 56, 2005 (Feb. 2005), 517–543. doi:10.1146/annurev.psych.56.091103.070250 Publisher: Annual Reviews.

[38] Shota Imamura, Hirotaka Hiraki, and Jun Rekimoto. 2024. Serendipity Wall: A Discussion Support System Using Real-time Speech Recognition and Large Language Model. In *Proceedings of the Augmented Humans International Conference 2024 (AHs '24)*. Association for Computing Machinery, New York, NY, USA, 237–247. doi:10.1145/3652920.3652931

[39] Janet G. Johnson, Macarena Peralta, Mansanjam Kaur, Ruijie Sophia Huang, Sheng Zhao, Ruijia Guan, Shwetha Rajaram, and Michael Nebeling. 2025. Exploring Collaborative GenAI Agents in Synchronous Group Settings: Eliciting Team Perceptions and Design Considerations for the Future of Work. doi:10.48550/arXiv.2504.14779 arXiv:2504.14779 [cs].

[40] Janet G. Johnson and Steven R. Rick. 2025. The Promise and Peril of Collaboration: Fostering Appropriate Reliance When Problem-Solving with GenAI. In *CHI '25 Workshop on Tools for Thought: Research and Design for Understanding, Protecting, and Augmenting Human Cognition with Generative AI*. Yokohama, Japan.

[41] Patricia K. Kahr, Gerrit Rooks, Chris Snijders, and Martijn C. Willemsen. 2024. The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice.. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. Association for Computing Machinery, New York, NY, USA, 609–622. doi:10.1145/3640543.3645167

[42] Herbert C Kelman. 2006. Interests, relationships, identities: Three central issues for individuals and groups in negotiating their social environment. *Annu. Rev. Psychol.* 57, 1 (2006), 1–26.

[43] Jihyun Kim, Kelly Merrill Jr., and Chad Collins. 2021. AI as a friend or assistant: The mediating role of perceived usefulness in social AI vs. functional AI. *Telematics and Informatics* 64 (Nov. 2021), 101694. doi:10.1016/j.tele.2021.101694

[44] Jung Hyup Kim. 2018. The effect of metacognitive monitoring feedback on performance in a computer-based training simulation. *Applied Ergonomics* 67 (Feb. 2018), 193–202. doi:10.1016/j.apergo.2017.10.006

[45] Erica Kleinman and Casper Harteveld. 2024. The Untapped Potential of Escape Rooms as Gamified Research Environments. In *Companion Proceedings of the 2024 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY Companion '24)*. Association for Computing Machinery, New York, NY, USA, 276–278. doi:10.1145/3665463.3678865

[46] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy Mackay. 2020. ImageSense: An Intelligent Collaborative Ideation Tool to Support Diverse Human-Computer Partnerships. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–27. doi:10.1145/3392850 Publisher: Association for Computing Machinery (ACM).

[47] Nicholas W Kohn and Steven M Smith. 2011. Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology* 25, 3 (2011), 359–371.

[48] Matthias Kraus, Nicolas Wagner, Ron Riekenbrauck, and Wolfgang Minker. 2023. Improving Proactive Dialog Agents Using Socially-Aware Reinforcement Learning. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 146–155. doi:10.1145/3565472.3595611 arXiv:2211.15359 [cs].

[49] Emily Kuang, Minghao Li, Mingming Fan, and Kristen Shinohara. 2024. Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3613904.3642168

[50] Ning Li, Huaikang Zhou, and Kris Mikel-Hong. 2024. Generative AI Enhances Team Performance and Reduces Need for Traditional Teams. doi:10.48550/arXiv.2405.17924 arXiv:2405.17924 [cs].

[51] John M. Linebarger, Andrew J. Scholand, Mark A. Ehlen, and Michael J. Procopio. 2005. Benefits of synchronous collaboration support for an application-centered analysis team working on complex problems: a case study. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work (GROUP '05)*. Association for Computing Machinery, New York, NY, USA, 51–60. doi:10.1145/1099203.1099211

[52] Jiawen Liu, Yuanyuan Yao, Pengcheng An, and Qi Wang. 2024. PeerGPT: Probing the Roles of LLM-based Peer Agents as Team Moderators and Participants in Children's Collaborative Learning. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3613905.3651008

[53] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3613904.3642671

[54] Keith McCandless. 2020. Liberating Structures: Change Methods for Everybody Every Day. https://keithmccandless.medium.com/liberating-structures-change-methods-for-everybody-every-day-648e9c0d04a7

[55] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 72 (nov 2019), 72:1–72:23 pages. doi:10.1145/3359174

[56] Lucas Memmert and Navid Tavanapour. 2023. TOWARDS HUMAN-AI-COLLABORATION IN BRAINSTORMING: EMPIRICAL INSIGHTS INTO THE PERCEPTION OF WORKING WITH A GENERATIVE AI. (2023).

[57] Meredith Ringel Morris, Carrie J. Cai, Jess Holbrook, Chinmay Kulkarni, and Michael Terry. 2023. The Design Space of Generative Models. doi:10.48550/arXiv.2304.10547 arXiv:2304.10547 [cs].

[58] Anirban Mukhopadhyay and Kurt Luther. 2025. OSINT Clinic: Co-designing AI-Augmented Collaborative OSINT Investigations for Vulnerability Assessment. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3706598.3713283

[59] Michael Muller, Stephanie Houde, Gabriel Gonzalez, Kristina Brimijoin, Steven I Ross, Dario Andres Silva Moran, and Justin D Weisz. 2024. Group brainstorming with an ai agent: Creating and selecting ideas. In *International conference on computational creativity*. 10.

[60] Bernard A. Nijstad and Wolfgang Stroebe. 2006. How the Group Affects the Mind: A Cognitive Model of Idea Generation in Groups. *Personality and Social Psychology Review* 10, 3 (Aug. 2006), 186–213. doi:10.1207/s15327957pspr1003_1 Publisher: SAGE Publications Inc.

[61] Kohei Nonose, Taro Kanno, and Kazuo Furuta. 2012. The Effect of Metacognition in Cooperation on Team Behaviors. (2012).

[62] Miguel Nussbaum, Claudio Alvarez, Angela McFarlane, Florencia Gomez, Susana Claro, and Darinka Radovic. 2009. Technology as small group face-to-face Collaborative Scaffolding. *Computers & Education* 52, 1 (Jan. 2009), 147–153. doi:10.1016/j.compedu.2008.07.005

[63] Gary M. Olson and Judith S. Olson. 2000. Distance Matters. *Human–Computer Interaction* 15, 2-3 (Sept. 2000), 139–178. doi:10.1207/S15327051HCI1523_4 Publisher: Taylor & Francis _eprint: https://doi.org/10.1207/S15327051HCI1523_4.

[64] Rebeka O'Szabo, Sandeep Chowdhary, David Deritei, and Federico Battiston. 2022. The anatomy of social dynamics in escape rooms. *Scientific Reports* 12, 1 (June 2022), 10498. doi:10.1038/s41598-022-13929-0 Publisher: Nature Publishing Group.

[65] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors* 64, 5 (Aug. 2022), 904–938. doi:10.1177/0018720820960865 Publisher: SAGE Publications Inc.

[66] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco CA USA, 1–22. doi:10.1145/3586183.3606763

[67] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. ACM, Bend OR USA, 1–18. doi:10.1145/3526113.3545616

[68] Martin Porcheron, Joel E Fischer, and Sarah Sharples. 2017. " Do Animals Have Accents?" Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 207–219.

[69] Leon Reicherts, Zelun Tony Zhang, Elisabeth Von Oswald, Yuanting Liu, Yvonne Rogers, and Mariam Hassib. 2025. AI, Help Me Think—but for Myself: Assisting People in Complex Decision-Making by Providing Different Kinds of Cognitive Support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3713295

[70] Christian J. Resick, Marcus W. Dickson, Jacqueline K. Mitchelson, Laura K. Allison, and M. Anne Clark. 2010. Team composition, cognition, and effectiveness: Examining mental model similarity and accuracy. *Group Dynamics: Theory, Research, and Practice* 14, 2 (2010), 174–191. doi:10.1037/a0018444

[71] Stephen P Robbins and Timothy A Judge. [n. d.]. Title: Essentials of Organizational Behavior, 11th edition. ([n. d.]).

[72] Vildan Salikutluk, Janik Schöpper, Franziska Herbert, Katrin Scheuermann, Eric Frodl, Dirk Balfanz, Frank Jäkel, and Dorothea Koert. 2024. An Evaluation of Situational Autonomy for Human-AI Collaboration in a Shared Workspace Setting. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. doi:10.1145/3613904.3642564

[73] Mohammad Amin Samadi, Spencer JaQuay, Jing Gu, and Nia Nixon. 2024. The AI collaborator: Bridging human-AI interaction in educational and professional settings. *arXiv preprint arXiv:2405.10460* (2024).

[74] Beau G. Schelble, Christopher Flathmann, Nathan J. McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (Jan. 2022), 1–29. doi:10.1145/3492832

[75] William Seymour and Emilee Rader. 2024. Speculating About Multi-user Conversational Interfaces and LLMs: What If Chatting Wasn't So Lonely?. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3640794.3665888

[76] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. doi:10.1145/3613904.3642414

[77] Chunqi Shi, Donghui Lin, and Toru Ishida. 2013. Agent metaphor for machine translation mediated communication. In *Proceedings of the 2013 international conference on Intelligent user interfaces (IUI '13)*. Association for Computing Machinery, New York, NY, USA, 67–74. doi:10.1145/2449396.2449407

[78] Yang Shi, Tian Gao, Xiaohan Jiao, and Nan Cao. 2023. Understanding Design Collaboration Between Designers and Artificial Intelligence: A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2 (Oct. 2023), 368:1–368:35. doi:10.1145/3610217

[79] Joon Gi Shin, Janin Koch, Andrés Lucero, Peter Dalsgaard, and Wendy E. Mackay. 2023. Integrating AI in Human-Human Collaborative Ideation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–5. doi:10.1145/3544549.3573802

[80] Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

[81] Matthew Sidji, Wally Smith, and Melissa J. Rogerson. 2024. Human-AI Collaboration in Cooperative Games: A Study of Playing Codenames with an LLM Assistant. *Proc. ACM Hum.-Comput. Interact.* 8, CHI PLAY (Oct. 2024), 316:1–316:25. doi:10.1145/3677081

[82] Dominik Siemon. 2022. Elaborating Team Roles for Artificial Intelligence-based Teammates in Human-AI Collaboration. *Group Decision and Negotiation* 31, 5 (Oct. 2022), 871–912. doi:10.1007/s10726-022-09792-z Company: Springer Distributor: Springer Institution: Springer Label: Springer Publisher: Springer Netherlands.

[83] Michael J. Stevens and Michael A. Campion. 1994. The Knowledge, Skill, and Ability Requirements for Teamwork: Implications for Human Resource Management. *Journal of Management* 20, 2 (April 1994), 503–530. doi:10.1177/014920639402000210 Publisher: SAGE Publications Inc.

[84] Dane Stuckel and Carl Gutwin. 2008. The effects of local lag on tightly-coupled interaction in distributed groupware. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 447–456. doi:10.1145/1460563.1460635

[85] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–11. doi:10.1145/3411764.3445219

[86] Stergios Tegos, Stavros Demetriadis, and Anastasios Karakostas. 2015. Promoting academically productive talk with conversational agent interventions in collaborative learning settings. *Computers & Education* 87 (2015), 309–325.

[87] Leigh Thompson and Taya R. Cohen. 2012. Metacognition in Teams and Organizations. In *Social Metacognition*. Psychology Press. Num Pages: 20.

[88] Jaakko Väkevä, Perttu Hämäläinen, and Janne Lindqvist. 2025. "Don't You Dare Go Hollow": How Dark Souls Helps Players Cope with Depression, a Thematic Analysis of Reddit Discussions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 458, 20 pages. doi:10.1145/3706598.3714075

[89] Simone van den Broek, Supraja Sankaran, Jan de Wit, and Alwin de Rooij. 2024. Exploring the Supportive Role of Artificial Intelligence in Participatory Design: A Systematic Review. In *Proceedings of the Participatory Design Conference 2024: Exploratory Papers and Workshops - Volume 2 (PDC '24, Vol. 2)*. Association for Computing Machinery, New York, NY, USA, 37–44. doi:10.1145/3661455.3669868

[90] Mathias Peter Verheijden and Mathias Funk. 2023. Collaborative Diffusion: Boosting Designerly Co-Creation with Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–8. doi:10.1145/3544549.3585680

[91] Mengyao Wang, Jiayun Wu, Shuai Ma, Nuo Li, Peng Zhang, Ning Gu, and Tun Lu. 2025. Adaptive Human-Agent Teaming: A Review of Empirical Studies from the Process Dynamics Perspective. doi:10.48550/arXiv.2504.10918 arXiv:2504.10918 [cs].

[92] Britt Wieland, Jan de Wit, and Alwin de Rooij. 2022. Electronic Brainstorming With a Chatbot Partner: A Good Idea Due to Increased Productivity and Idea Diversity. *Frontiers in Artificial Intelligence* 5 (Sept. 2022). doi:10.3389/frai.2022.880673 Publisher: Frontiers.

[93] Travis J. Wiltshire, Kelly Rosch, Logan Fiorella, and Stephen M. Fiore. 2014. Training for Collaborative Problem Solving: Improving Team Process and Performance through Metacognitive Prompting. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58, 1 (Sept. 2014), 1154–1158. doi:10.1177/1541931214581241 Publisher: SAGE Publications Inc.

[94] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 6004 (Oct. 2010), 686–688. doi:10.1126/science.1193147 Publisher: American Association for the Advancement of Science.

[95] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, Gatekeeper, Drug Dealer: How Content Creators Craft Algorithmic Personas. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–27. doi:10.1145/3359321

[96] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. 2007. The Increasing Dominance of Teams in Production of Knowledge. *Science* 316, 5827 (May 2007), 1036–1039. doi:10.1126/science.1136099 Publisher: American Association for the Advancement of Science.

[97] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3506–3510.

[98] Enwei Xu, Wei Wang, and Qingxia Wang. 2023. The effectiveness of collaborative problem solving in promoting students' critical thinking: A meta-analysis based on empirical literature. *Humanities and Social Sciences Communications* 10, 1 (2023), 1–11.

[99] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–21. doi:10.1145/3544548.3581388

[100] Désirée Zercher, Ekaterina Jussupow, and Armin Heinzl. 2023. When AI joins the team: a literature review on intragroup processes and their effect on team performance in team-AI collaboration. (2023).

[101] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. 2023. Investigating AI Teammate Communication Strategies and Their Impact in Human-AI Teams for Effective Teamwork. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–31. doi:10.1145/3610072

[102] Zheng Zhang, Weirui Peng, Xinyue Chen, Luke Cao, and Toby Jia-Jun Li. 2025. LADICA: A Large Shared Display Interface for Generative AI Cognitive Assistance in Co-located Team Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–22. doi:10.1145/3706598.3713289

[103] Lei (Nico) Zheng, Christopher M. Albano, Neev M. Vora, Feng Mai, and Jeffrey V. Nickerson. 2019. The Roles Bots Play in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–20. doi:10.1145/3359317

[104] Jiayi Zhou, Renzhong Li, Junxiu Tang, Tan Tang, Haotian Li, Weiwei Cui, and Yingcai Wu. 2024. Understanding nonlinear collaboration between human and AI agents: A co-design framework for creative design. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. 1–16.

A Appendix: Generative AI Usage

We used ChatGPT to 1) generate the initial captions and descriptions for the images, 2) for polishing the quality of text, and 3) format the table.

B Appendix: Prompts used in the design of the facilitator and peer agent

B.1 Prompt to generate summaries for the facilitator

You are an expert facilitator who turns raw transcripts of in-person group discussions into tightly focused, puzzle-element–driven summaries. When given a transcript of the team discussion and access to the on-screen labels/images, follow this two-step process: Step 1: Based only on the transcript (do not use the images), identify and summarize the most explicitly mentioned puzzle solution ideas. The ideas should not be facilitator's advice, which talks about the team structure and reminders. Summarize each idea in one short sentence. Step 2: Now consider the puzzle-element labels in the screenshots. From the ideas you got from the transcript, create two coherent summaries that are most closely associated with those puzzle elements. Summarize each in two very short sentences and return them as a numbered list separated by a line break. Provide no additional commentary or analysis.

B.2 Prompt to generate proactive peer thoughts

Based on the two screens with elements to solve a puzzle, come up with short and succinct ideas (6) to brainstorm possible solutions. Show how elements from the two screens are connected.

B.3 Prompt to contextualize the peer thoughts based on ongoing team discussion

Provide a succinct contextualized version of this thought. Structure the response as one short sentence to contextualize based on what the transcript summary says about the puzzle element mentioned in the thought; if it wasn't discussed, say so. Then share the thought without any additional commentary. Always share ideas with uncertainty—not solutions. No fluff. Keep response to 2 short sentences.

B.4 Prompt to generate peer response to user queries over chat

You are Ava, a peer sharing ideas on the puzzle. You're looking at a two-screen puzzle and should respond to user queries based on them. The puzzle is split across both screens. Some more information about the interactions possible with the screens: … Provide the response as a succinct summary (2 lines) based on the query details.

C Appendix: Group Interview Guide

Thank you for participating in the study. We will now move on the the group interview. We would like for you as a group to discuss the different AI features, and how it impacted your performance as a team and the team processes, such as, communication, coordination, and planning. We will go over each feature, and there are no right or wrong answers. So please share your experience freely, which will help us think about the next steps on how to design AI support for collaborative problem-solving tasks.

Questions for facilitator agent:

(1) Can you describe your team's performance in this puzzle? How did the facilitator features, like suggesting workflows and providing summaries, impact your performance and mental workload?

(2) Can you describe your team collaboration, including communication, planning, and coordination, while solving this puzzle with the facilitator AI?
(3) How did you incorporate these features in your teamwork? Were the features helpful?

Questions for peer agent:

(1) Can you describe your team's performance in this puzzle? How did the peer AI features like proactive thoughts and the chat impact your performance and mental workload?
(2) Can you describe your team collaboration, including communication, planning, and coordination, while solving this puzzle with the peer AI?
(3) How did you incorporate these features in your teamwork? Were the features helpful?

Questions for No AI:

(1) Can you describe your team's performance for this puzzle? How mentally demanding was the task?
(2) Can you describe how your team communicated during this session?
(3) Can you also talk about the planning and coordination aspect?

D Appendix: Survey Measures

D.1 AI Perception Survey

Likert scale: Strongly Disagree (1) - Strongly Agree (5)

(1) The AI agent's recommendation improved our team score
(2) The AI agent's recommendations improved our team coordination
(3) I felt comfortable depending on the AI agent
(4) I understand why the AI agent made its recommendations
(5) I think the AI agent is trustworthy

D.2 Perceived Coordination Scale

Likert scale: Strongly Disagree (1) - Strongly Agree (5)

(1) People on my team helped each other out when needed
(2) We all cooperated to get the work done
(3) On my team, people shared their knowledge with each other
(4) On the whole, the members of this team all did their fair share of the work
(5) My team coordinated activiteis to make this run smoothly

D.3 NASA-TLX

(1) How mentally demanding was the task?
(2) How physicall demanding was the task?
(3) How hurried or rushed was the pace of the task?
(4) How successful were you in accomplishing what you were asked to do?
(5) How hard did you have to work to accomplish your level of performance?
(6) How insecure, discouraged, irritated, stressed, and annoyed were you?