Understanding Human–Multi-Agent Team Formation for Creative Work

Hyunseung Lim
KAIST
Daejeon, Republic of Korea
charlie9807@kaist.ac.kr

Dasom Choi
National University of Singapore
Singapore, Singapore
dasomchoi.w@gmail.com

Sooyohn Nam
KAIST
Daejeon, Republic of Korea
suyeon.nam@kaist.ac.kr

Bogoan Kim
Chungbuk National University
Cheongju, Republic of Korea
bogoan@cbnu.ac.kr

Hwajung Hong
KAIST
Daejeon, Republic of Korea
hwajung@kaist.ac.kr

Abstract

Team-based collaboration is a cornerstone of modern creative work. Recent advances in generative AI open possibilities for humans to collaborate with multiple AI agents in distinct roles to address complex creative workflows. Yet, how to form Human–Multi-Agent Teams (HMATs) is underexplored, especially given that inter-agent interactions increase complexity and the risk of unexpected behaviors. In this exploratory study, we aim to understand how to form HMATs for creative work using CrafTeam, a technology probe that allows users to form and collaborate with their teams. We conducted a study with 12 design practitioners, in which participants iterated through a three-step cycle: forming HMATs, ideating with their teams, and reflecting on their teams' ideation. Our findings reveal that while participants initially attempted autonomous team operations, they ultimately adopted team formations in which they directly orchestrated agents. We discuss design considerations for HMAT formation that humans can effectively orchestrate multiple agents.

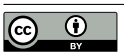
CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**.

Keywords

Human–Multi-Agent Team, Human-AI Team, Multi-Agent, Team Formation

ACM Reference Format:
Hyunseung Lim, Dasom Choi, Sooyohn Nam, Bogoan Kim, and Hwajung Hong. 2026. Understanding Human–Multi-Agent Team Formation for Creative Work. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3772318.3791166



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791166

1 Introduction

Team-based collaboration is a cornerstone of modern creative work. By bringing diverse perspectives, teams can extend thinking beyond individual limits, reframe problems, and integrate constraints into robust solutions, which is particularly helpful for tackling creative problems that are often wicked and challenging [16]. With recent advances in generative AI, AI agents have been proposed as collaborators that can support humans in creative activities [64, 83], including writing (e.g., co-authoring and editing) [44, 81], design (e.g., suggesting ideas and references) [11], and the performing arts (e.g., inspiring music and choreography) [26, 36]. These agents increasingly assume cognitively distinct roles, mirroring the complementary functions of real-world creative teams, ranging from productive partners to reflective guides [34, 88].

The rise of these agents with diverse functions is enabling a new paradigm in creative workflows: collaborating with multiple AI agents in distinct roles as a team. Researchers and practitioners are already applying this approach, for example, in automated marketing campaigns that integrate productive agents for content writing and graphic design with reflective agents for content evaluation [54], or in design workflows where CMF designer agents work alongside reflective design director and product manager agents [14]. By forming Human–Multi-Agent Teams (HMATs) [90], in which humans collaborate with multiple AI agents, individuals can address complex tasks by decomposing them into clearly defined roles and simulating diverse perspectives.

Although we envision HMATs enabling synergistic collaboration by simultaneously leveraging multiple agents, they often struggle to achieve such cooperation in practice. These challenges lie not only in the performance of individual agents but also in the novel task of team formation. Team formation involves deliberately organizing team structures, role allocations, and interaction protocols to enable effective collaboration [39, 86]. Since HMATs combine human members with heterogeneous AI agents, the design space for team formation becomes even more complex, involving multiple types of interactions between humans and agents as well as among agents themselves [18]. In particular, these inter-agent interactions complicate team formation by increasing the risk of unexpected behavior and creating a need for careful coordination and clear functional boundaries between agents [54]. While the importance of HMAT formation has attracted increasing attention [12, 30, 46], the question of which formations facilitate effective collaboration

between individuals and multiple agents and lead to improved team outcomes remains underexplored.

This study aims to deepen our understanding of how HMATs can be formed to better support individuals in collaborating with multiple agents on creative work. To this end, we developed CrafTeam, a technology probe that enables users to form and collaborate on their own HMATs. CrafTeam was designed to uncover design considerations in HMAT formation by revealing how users configure five key dimensions of team formation: *team size, structure, role allocation, member composition, and shared mental models*. Using CrafTeam, we conducted a three-hour user study with 12 design practitioners currently working in design teams within IT companies. We structured the study around a three-step cycle in which participants (i) formed their own team, (ii) ideated with that team, and (iii) reflected on the team's collaboration. Participants repeated the cycle three times, carrying insights from each round forward to refine their HMATs. Drawing on post-study interviews, we distill lessons learned from the process of forming HMATs and participants' perceptions of interacting with HMATs.

Through the user study, participants iteratively adjusted their team formations and reflected on how these formations shaped the ideation process. Our results show that while they initially adopted team formations in which agents autonomously developed ideas with minimal human involvement, they found that inter-agent interactions often led to unproductive loops in which agents failed to provide clear direction or to make the value judgments necessary for creative progress. In response, participants shifted to team formation in which they themselves set the direction for ideation and directly orchestrated the agents. Based on our findings, we underscore the need for human-orchestrated HMAT formation and propose considerations that help users manage and collaborate with multiple agents.

The major contributions of our study are as follows:

- The design and implementation of CrafTeam, a technology probe that enables users to form and collaborate with HMATs for design ideation tasks.
- Empirical findings from a user study with 12 design practitioners who iteratively formed and refined HMATs, continuously adjusting five dimensions of team formation.
- Future opportunities and design considerations for forming and orchestrating HMATs based on participants' experiences using CrafTeam.

2 Related Works

2.1 Multi-Agent System for Creative Work

Multi-agent systems (MAS) consist of autonomous entities known as agents, which show promise in solving complex tasks by dividing them into multiple smaller tasks, each assigned to a distinct agent [63]. Rather than relying on a single powerful entity, MAS distributes overhead across multiple specialized agents, achieving flexibility through modular independence and resource efficiency via parallel task execution [15, 79]. Recent advances in large language models have expanded the possibilities of multi-agent collaboration, enabling agents to interpret context, generate novel combinations, and coordinate through natural language rather than predefined protocols [25, 83]. Such capabilities empower agents to simulate human group dynamics through negotiation, debate, and perspective synthesis, which are absent in traditional rule-based agents. For instance, AgentVerse [10] demonstrates a group of experts, including architect, designer, and engineer agents, engaging in human-like discussions to reach collaborative decisions, achieving performance superior to that of individual agents across diverse problem-solving tasks.

These developments suggest the potential that MAS can address the complex challenges of creative work [46], where design decisions simultaneously affect multiple dimensions, requiring diverse and specialized perspectives. Such creative work is recognized as a wicked problem, especially in its co-evolution where problems and solutions develop together, resisting definitive formulation [16]. MAS responds by distributing cognitive load across specialized agents, each maintaining a distinct perspective while enabling parallel processing and iterative refinement [28, 82]. Recent studies explore how MAS integrates diverse perspectives: DesignGPT [14] assigns product manager and CMF designer agents to different design dimensions, while MARE [31] experiments with stakeholder, modeler, and checker agents for parallel dependency management. Meanwhile, AgileCoder [55] shows how product manager, developer, and tester agents can sustain continuous refinement through repeated sprints. Likewise, HoLLMwood [9] pairs a writer and an editor agent to sustain divergence–convergence loops, divergent exploration, and convergent refinement.

While MAS shows potential for addressing the interdependent constraints and iterative cycles of creative work, current implementations often overlook the critical importance of human participation. Most studies on MAS center on fully autonomous pipelines where humans provide initial prompts and receive final outputs, limiting human involvement to bookends rather than integrating it throughout the co-evolutionary process [12, 46]. This pattern proves fragile in creative settings, where agents must navigate implicit values, contextual norms, and shifting constraints; without sustained human guidance, they drift from creative intent or optimize misaligned objectives [42, 46, 83]. For example, only LLM-based agents often remain unaware of tool affordances or consequences of their proposed actions, lacking the situated understanding that humans naturally possess and thus proposing impractical or unfeasible solutions [74]. Realizing MAS's potential to support complex, iterative creative work requires active human participation to guide and coordinate agents throughout the process.

2.2 Forming Human-Multi-Agent Teams

The field of HCI has established Human-Agent Teams (HATs) as a collaborative paradigm in which humans engage as active team members alongside autonomous AI agents, rather than using agents as tools [5]. HAT, variously referred to as Human-Autonomy Team [58], Human-Agent Team [30, 67], Human-Machine Team [52, 80], Human-Robot Team [13, 32, 71], or Human-AI Team [18, 43], all involve humans and autonomous agents working in close coordination to achieve shared goals, operating under principles of mixed initiative and mutual adaptation. Unlike the broader concept of human-AI collaboration, where humans issue commands and agents execute, HAT involves continuous negotiation, role flexibility, and real-time

coordination—humans and agents monitor each other's states, compensate for each other's limitations, and jointly determine action paths [32, 35, 91, 92]. This enables humans and agents to communicate continuously, complement each other, and perform tasks effectively through sustained interaction [19].

As human teams' collective intelligence depends more on team formation than individual brilliance, which has long been recognized as crucial for maintaining team performance and sustainability [39, 86]. The importance of team formation extends to HATs as well, particularly because AI agents only perform their assigned tasks, making it essential to configure them appropriately from the outset [33, 80]. HAT formation involves navigating multiple considerations: selecting complementary agents aligned to task requirements [30], establishing role and authority structures balancing human and agent expertise [18, 20, 30, 69], designing manageable interaction protocols [43, 91], and resolving conflicts continuously [30, 52]. HAT formation often adopts foundational principles from human teams, such as structured roles and communication protocols [58, 66]. However, when humans and AI agents coexist on the same team, unique challenges arise, such as fostering consistent mental models between humans and agents, building trust in AI teammates, and maintaining communication quality [19, 66, 68]. Therefore, HAT formation requires using human team strategies as a starting point while accounting for the unique characteristics introduced by AI agents [91].

HAT formation becomes more complex when moving from one-human one-agent dyads to Human-Multi-Agent Teams (HMATs), as coordination across multiple parallel relationships [18, 30]. HMATs —defined as teams with at least one human and two or more autonomous agents pursuing shared goals [51, 73, 90]—introduce unique interdependencies where team members must coordinate tasks and integrate outputs across multiple channels. This shift introduces critical considerations, including task allocation across different agents and the design of inter-agent interaction protocols [18]. However, research on team formation in HMATs remains relatively limited, partly because most prior work has focused on one-human–one-agent teams [18, 19, 80, 91], in which the team formation is relatively straightforward. As recent studies have begun to explore HMATs, there has been growing attention to how HMAT formations should be designed [12, 30, 46]; for instance, Abhinav et al. [12] argue that designing multi-agent HRI systems requires explicit consideration of aspects such as team size, member composition, and interaction style. Yet, most existing studies that implement HMATs adopt provisional team configurations chosen by researchers rather than systematically developing and refining strategies for forming HMATs [90]. As an early exploration of HMAT formation, our study aims to identify key considerations and challenges specific to HMAT formation and examine how HMATs can be formed to better support collaboration with multiple agents.

2.3 Team Formation Strategies for Creative Work

Teams, defined as two or more individuals who systematically distribute tasks and interact closely toward shared objectives, are the fundamental organizational units for navigating today's hypercompetitive business environment [37, 49, 50]. In creative industries, teams are especially central, as they bring together diverse skills and perspectives for addressing complex challenges [41, 62]—from corporate product development requiring high innovation [41, 78] to interdisciplinary fields demanding specialized expertise across domains [87]. To achieve such creative and innovative outcomes, team formation plays a significant role, especially in creative teams [39, 86], shaping how diverse perspectives merge, how conflicts transform into productive tension, and how trust enables risk-taking essential for innovation [1, 38, 75].

Given the complexity of creative work and its demand for innovation, forming such teams requires dedicated strategies. A long line of work in organizational psychology and creative strategy has sought to operationalize team development principles—Tuckman's model of small-group development explains how teams evolve through forming, storming, norming, and performing stages [6, 76], while Belbin's Team Role Model offers systematic approaches for composing balanced teams with complementary roles such as coordinators, implementers, and creative specialists [4]. In parallel, HCI research, including CSCW (Computer-Supported Cooperative Work), has long proposed tools and strategies for assembling effective teams. For instance, prior work has examined how people search for teammates on online platforms [24] and suggested algorithms for team formation in these environments [2]. Since the quality of a team formation is often difficult to assess before the team actually starts working together, prior studies have explored approaches such as a "team dating" that quickly tries out multiple team formations [48] and techniques that use repeated trials to help teams discover suitable patterns of initial interaction [85].

The transition to human-agent teams requires adapting established team formation principles to account for fundamental differences in how humans and AI agents operate—particularly in multi-agent settings [33, 73]. Unlike human teams, where members are recruited with relatively fixed capabilities and personalities, AI teammates can be instantiated to match desired profiles on demand, fundamentally changing the nature of team formation. At the same time, while humans naturally negotiate roles and interpret implicit social cues, AI agents require explicit protocols and structured interactions. This asymmetry creates cascading challenges: precisely defining agent roles becomes critical as agents cannot adapt fluidly; orchestrating multi-agent coordination grows complex without implicit understanding; maintaining human oversight while preserving agent autonomy requires a delicate balance; and establishing shared mental models proves difficult when humans and agents perceive and process information fundamentally differently [3, 57]. While prior work has begun to surface these challenges, it has offered limited examination of how they extend to team formation involving multiple AI agents operating concurrently with humans [33]. Building on these insights, our study examines how strategies for forming HMATs both resemble and diverge from traditional human creative team formation, and explores the unique considerations and team formation strategies that emerge when configuring such teams for creative work.

3 Design of CrafTeam

Our primary goal is to investigate what practical challenges and design considerations arise when forming HMATs for collaborative

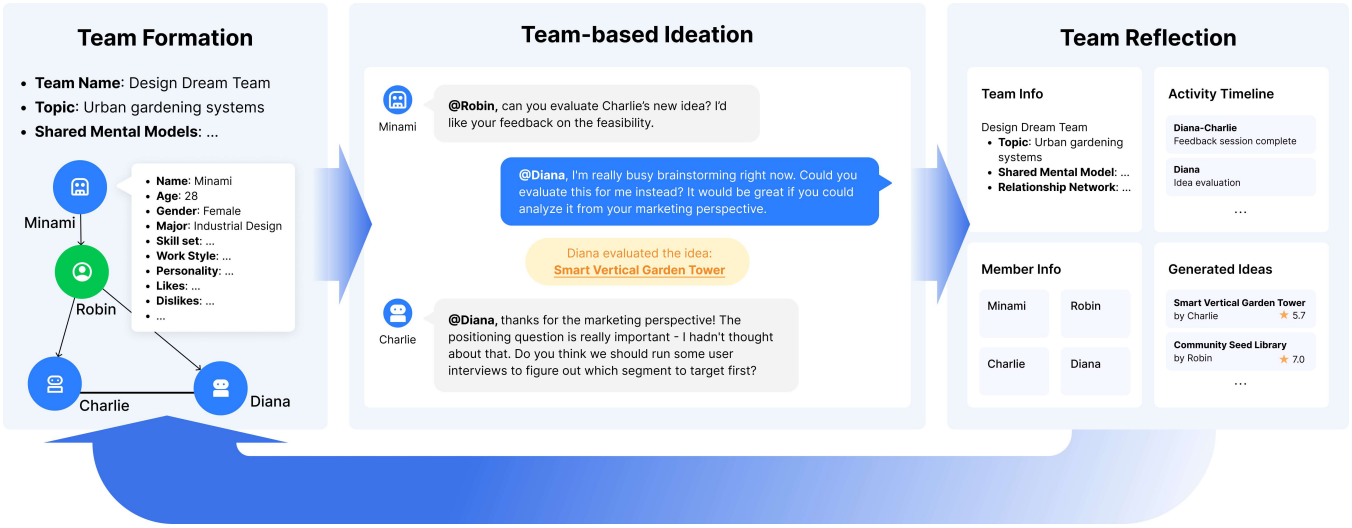


Figure 1: Overall flow of CrafTeam. CrafTeam supports three repeating phases—forming HMATs, ideating within HMATs, and reflecting on the team's ideation.

creative work, particularly where assumptions and strategies from HATs and human-only teams may no longer hold. Since HMAT formation remains underexplored, we adopt a technology probe method [29], well-suited to eliciting user-grounded empirical insights to inform the design of new technologies. To this end, we developed CrafTeam, a technology probe that enables users to form and collaborate with their own HMATs in creative workflows, revealing how they specify HMAT formations in practice and what considerations arise in doing so. The following sections describe the concept and interfaces of CrafTeam, and implementation details are provided in Appendix A.

3.1 Overall Concept

CrafTeam is a web application that enables users to form their own HMATs and engage in team-based ideation sessions. To make this process accessible even to non-developers, CrafTeam lets users configure only the core dimensions of team formation, while the system automatically constructs complete HMATs based on the users' settings. In particular, users can directly configure five dimensions of team formation—*team size, structure, role allocation, member composition, and shared mental models* (detailed in Section 3.2.1).

To enable users to experience the practical consequences of their formation choices for team collaboration, we let them engage in a collaborative ideation task where they work directly with their HMATs to produce creative outcomes. We chose ideation as the primary task because it inherently involves discussion, knowledge exchange, and the integration of diverse perspectives, which has been a widely examined task for human–AI collaboration [23, 27, 56, 72]. To implement the co-ideation task with AI agents, we adopted the human-AI co-ideation framework proposed by Shen et al. [72], which enables humans and AI to share an idea space and track the evolution of design ideas through structured idea representations.

We then incorporated a post-ideation reflection phase with an interface for reviewing ideation transcripts and analyzing interaction patterns. Because collaboration performance in HMATs is shaped by multiple dimensions of team formation acting together, users can struggle to see how any formation choice relates to the quality of their collaborative experience during ideation. By introducing this phase, we enable users to trace specific choices to observed effects and further support systematic evaluation of how their team formation choices affect collaboration during ideation.

We implemented these components as iterative cycles (Fig. 1): (i) forming their own HMATs, (ii) ideating with their teams, (iii) reflecting on the ideation session, and then reforming their team based on insights gained from the reflection. We adopted this iterative process to uncover which team formation users select or adjust and how those choices affect collaboration, because it is hard to compare how individual dimensions or their combinations influence collaboration from a single trial [48]. In this study, users were asked to iterate through three cycles, during which they refined their HMAT formation as new empirical insights emerged. These iterations allow us to observe how users' mental models evolve and which team formation they prioritize after gaining hands-on experience.

3.2 HMAT Formation in CrafTeam

In this section, we introduce five key dimensions of HMAT formation that can be configured in CrafTeam, and present the Team Formation Interface, which enables users to form their own HMATs by manipulating these dimensions.

3.2.1 Dimensions for HMAT Formation. Drawing from prior research on human-agent team formation [12, 30], we adopted five dimensions of team formation that shape team dynamics and influence effective collaboration: *team size, structure, role allocation, member composition, and shared mental models*. In selecting these dimensions, we focused on aspects of team formation that non-developer users could explicitly configure through the interface

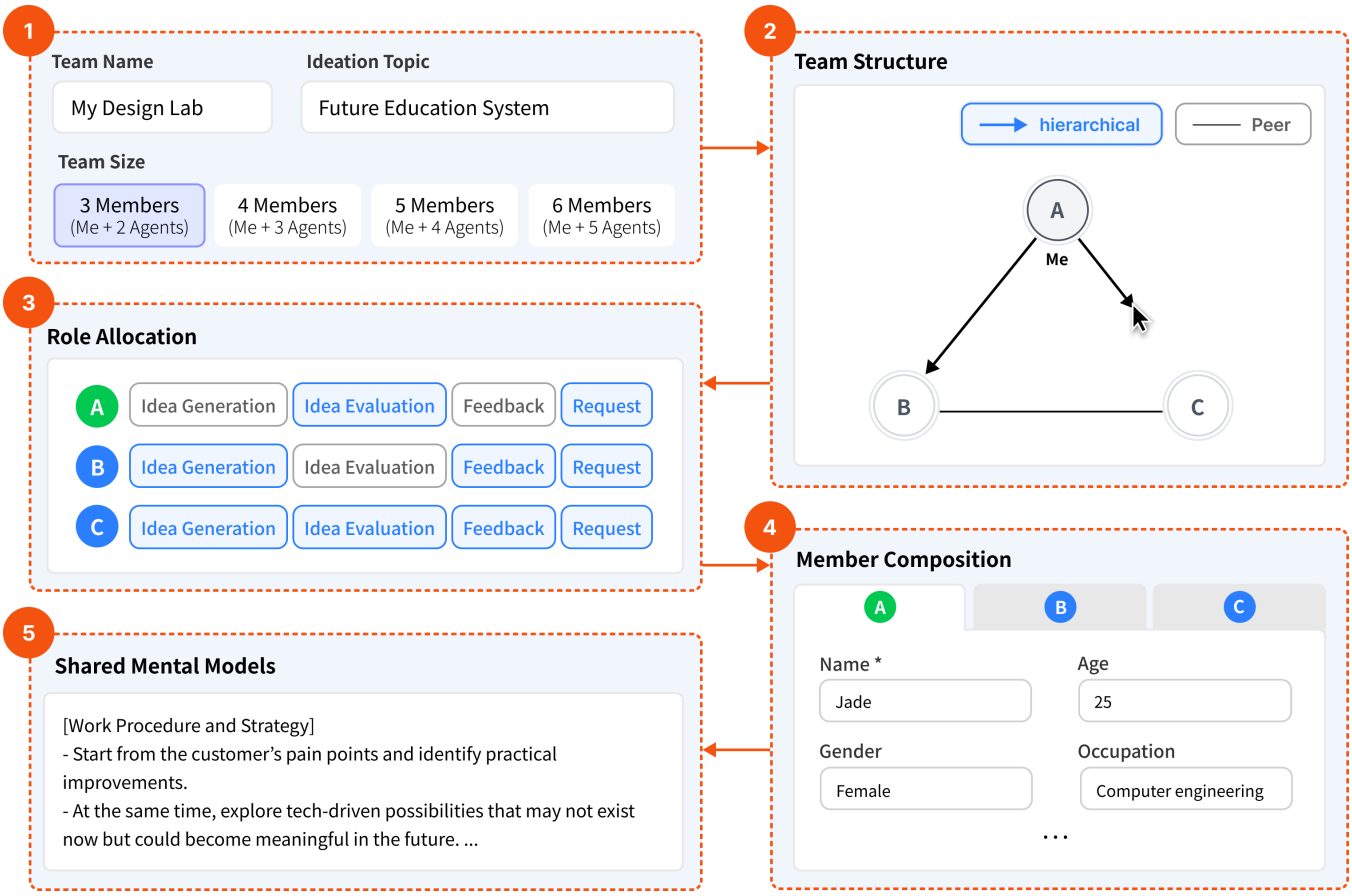


Figure 2: Simplified Team Formation Interface of CrafTeam. (1) The user sets the team name and ideation topic and chooses the team size. (2) The user configures a networked team structure by linking members as hierarchical or peer. (3) The user assigns roles to each member—Idea Generation, Idea Evaluation, Feedback, and Request. (4) The user creates member personas in a resume-like form aligned with roles. (5) The user establishes the team's shared mental models.

(e.g., size, structure, roles), and excluded dimensions that are more difficult to operationalize as direct settings, such as interaction style or team intervention protocols.

Team size refers to the number of members (including the user) that form a team. Once size is determined, team structure defines how these members organize to break down complex tasks that exceed any individual's capacity [80]. Within this structure, role allocation assigns specific work functions to each member, both human and automated [20, 69]. Beyond functional assignments, member composition considers the diversity of member characteristics and how these differences impact team processes and outcomes [61]. Finally, shared mental models (SMMs) refer to the extent to which team members share a common understanding of team tasks, goals, and members' capabilities, which has been linked to improved coordination and team performance [3, 66, 67]. Based on these interconnected dimensions of HMAT formation, we designed the Team Formation Interface, which allows users to configure and refine them based on empirical insights from team-based ideation sessions.

3.2.2 Team Formation Interface. The Team Formation Interface is designed to enable users to form their own HMATs by configuring key dimensions of team formation. The interface consists of five stages (Fig. 2), each corresponding to one core dimension of team formation: (i) Team Size, (ii) Team Structure, (iii) Role Allocation, (iv) Member Composition, and (v) Shared Mental Model. After completing all stages, users can finalize their team by clicking the "Create Team" button, which then transitions them to the team-based ideation phase.

In the first step, users set basic team information and determine the team size. They begin by naming their team and defining the ideation topic the team will address. Then users set the team size, which can range from 3 to 6 members. We limited the minimum to three members (one user and two AI agents) to ensure multi-agent team interactions could be explored and capped the maximum at six (one user and five AI agents) as an appropriate scale for design ideation and to prevent cognitive overload from managing and interacting with multiple agents [59, 89].

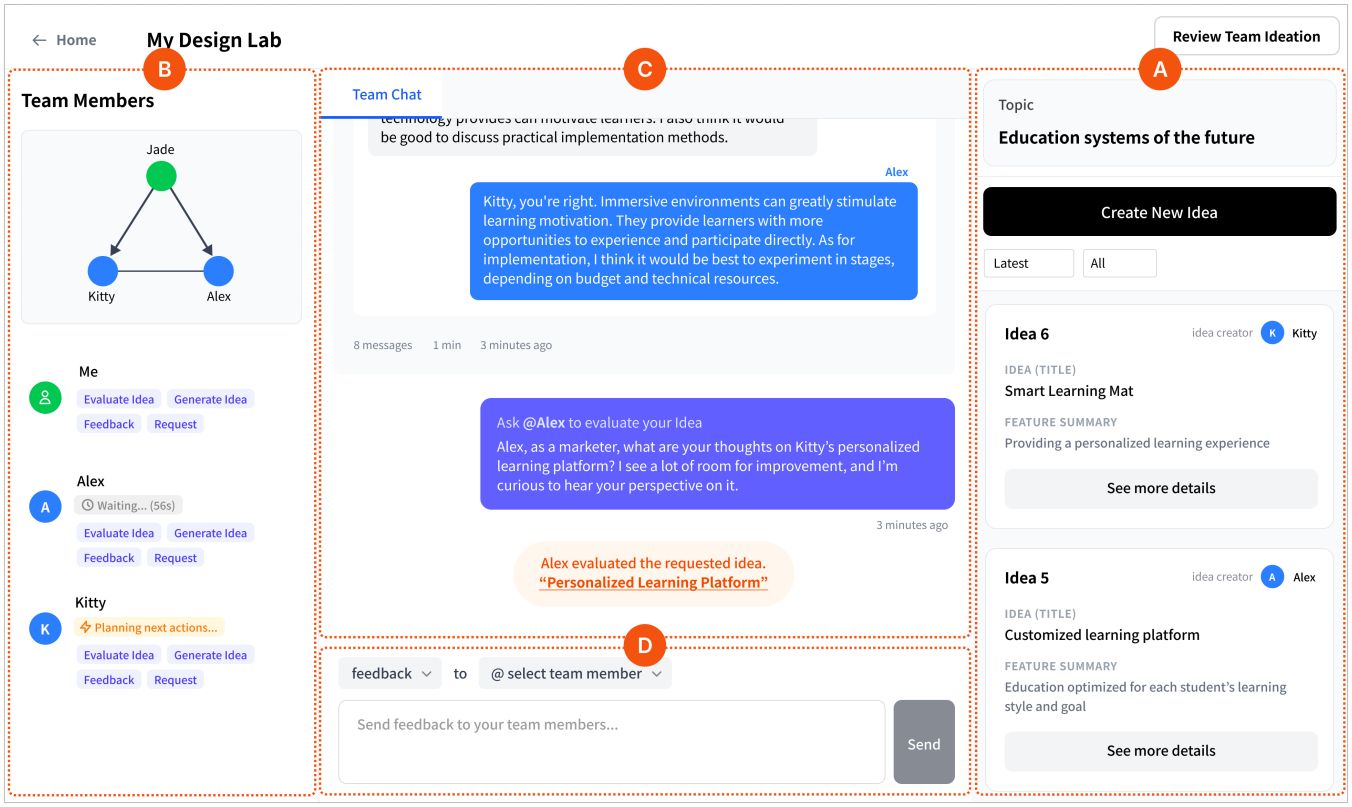


Figure 3: Ideation Interface of CrafTeam. (A) Idea Tab displays all team-generated ideas. It allows users to generate new ideas through the "Create New Idea" button and provides detailed views via the "See more details" option, which reveals full idea content and enables updates or evaluations (shown as Fig 4). (B) Team Status Tab displays the team structure and member information, including the roles assigned to each member, the relationships between team members, and the current status of the AI agents. (C) Team Chat Window displays real-time interaction logs, including feedback sessions between team members. (D) Chat Input Field enables users to provide feedback or make requests to AI agents.

In the second step, users establish a team structure by assembling relationships between team members. We implemented a network-based team structure to allow users to create diverse relationships among members and form teams with various structural configurations. Each member serves as a node, and users connect their relationships as edges to create a network graph-style team structure. When establishing relationships, users can choose between hierarchical or peer relationships. Only members who are directly connected can interact with each other.

In the third step, users assign roles to each team member by selecting one or more of four defined roles: Idea Generation, Idea Evaluation, Feedback, and Request (detailed in Section 3.3.2). Every member must be assigned at least one role, and at least one member must have the Idea Generation role to ensure the team can produce ideas. During the ideation phase, each member can perform only their assigned roles.

In the fourth step, users define the characteristics of each team member by inputting personas in a resume-like format. The information fields are based on the framework for multidimensional identity representation in LLM-based agents [40], including Social Identity (Age, Gender, Education, Occupation), Personal Identity (Personality, Skills), and Personal Life Context (Behavior, Likes, Dislikes), adapted for the context of building ideation teams. We encouraged users to design personas suitable for the intended roles, allowing them to specify only the characteristics they deemed necessary, which enabled us to understand which attributes they prioritized.

In the final step, users establish the team's SMMs. We asked them to write textual descriptions of task models (e.g., procedures, possible outcomes, and how to handle them) and team models (e.g., teammates' tendencies, beliefs, and personalities), which were used as hard-coded guidelines for team-based ideation.

3.3 Team-Based Ideation of CrafTeam

After forming HMATs, users move to CrafTeam's Team-Based Ideation phase to ideate with their team (Fig. 3). We first outline the ideation task and the roles both users and AI agents can take in ideation, then describe the Team-Based Ideation Interface.

3.3.1 Ideation Task. In CrafTeam's team-based ideation phase, users collaborate with their HMATs to engage in a conceptual

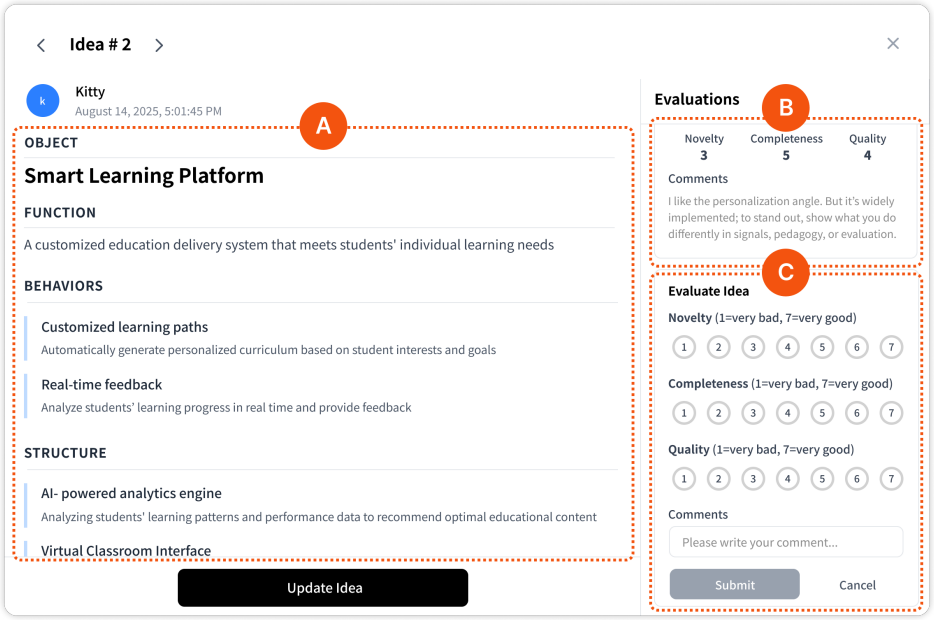


Figure 4: Idea Card in CrafTeam. (A) The main view displays the idea representation, including Object, Function, Behavior, and Structure. (B) The Evaluation List displays team members' ratings and comments on the idea. (C) The Evaluate Idea Tab allows users to provide their own ratings and comments on the idea.

ideation task on a topic of their choice. In CrafTeam, ideas are represented in a unified format [72] with four components: object (the design target), function (the intended purpose or teleology), behavior (what it does or how it responds, expected or derived from its structure), and structure (its components and their compositional relationships). Each idea is displayed as a card in the interface (Fig. 4).

3.3.2 Roles in Team-based Ideation. During ideation, the user and AI agents can generate, share, and give feedback on ideas, iteratively refining them through feedback and initial screening. We identified four core roles that both users and agents can take in an ideation session: Idea Generation, Idea Evaluation, Feedback, and Request. Users can serve in these roles through the user interface (Section 3.3.3), whereas AI agents can serve in these roles autonomously via LLM-based pipelines (Section A.1). Both users and agents may perform an action only if they were assigned the corresponding role during the team formation phase.

Idea Generation. Idea Generation is the action of creating idea cards. In this role, members can either (i) *Generate New Idea* to create original concepts or (ii) *Update Idea* to refine existing ideas using their existing representations as templates. All new and updated ideas appear in the Idea tab, which is visible to the whole team (Fig. 3A).

Idea Evaluation. Idea Evaluation is the action of assessing ideas. In this role, members rate each idea on three 7-point Likert scales —novelty, completeness, and quality [72]—and can optionally add brief comments. Ratings and comments are displayed on the right side of each idea card (Fig. 4B).

Feedback. Feedback is the action of providing conversational feedback to other team members about specific ideas [45] or the overall teamwork [17]. In this role, a member can open a one-on-one chat with a selected team member to whom they are connected in the team structure and exchange multi-turn feedback messages.

Request. Request is the action of asking other team members to take specific actions in the ideation workflow [77]. In this role, members can request teammates to whom they are connected in the team structure to perform Idea Generation, Idea Evaluation, or Feedback.

3.3.3 Team-Based Ideation Interface. Figure 3 illustrates the Ideation Interface of CrafTeam. The interface supports fluid collaboration by making all team activities visible and accessible, allowing users to respond adaptively to the team's ideation process.

At the beginning of an ideation session, only Idea Generation is available to establish a shared starting point. After the first idea is generated, they may also perform Idea Evaluation, Feedback, and Request within their assigned role permissions. In the Idea Tab (Fig. 3A), users can click Generate New Idea to add a new idea card. They can also open any idea card in the list to evaluate it or update the idea. Through the Chat Input Field (Fig. 3D), users can send either Feedback or Request by selecting a recipient, choosing the message type (Feedback or Request), and typing a brief message; for Request, they also choose the action for the recipient to perform. Through Team Status (Fig. 3B), users can monitor team- and member-level status to decide when to provide Feedback or make Requests. Through Team Chat (Fig. 3C), users can see member activity in real time and, once a Feedback conversation ends, review its transcript.

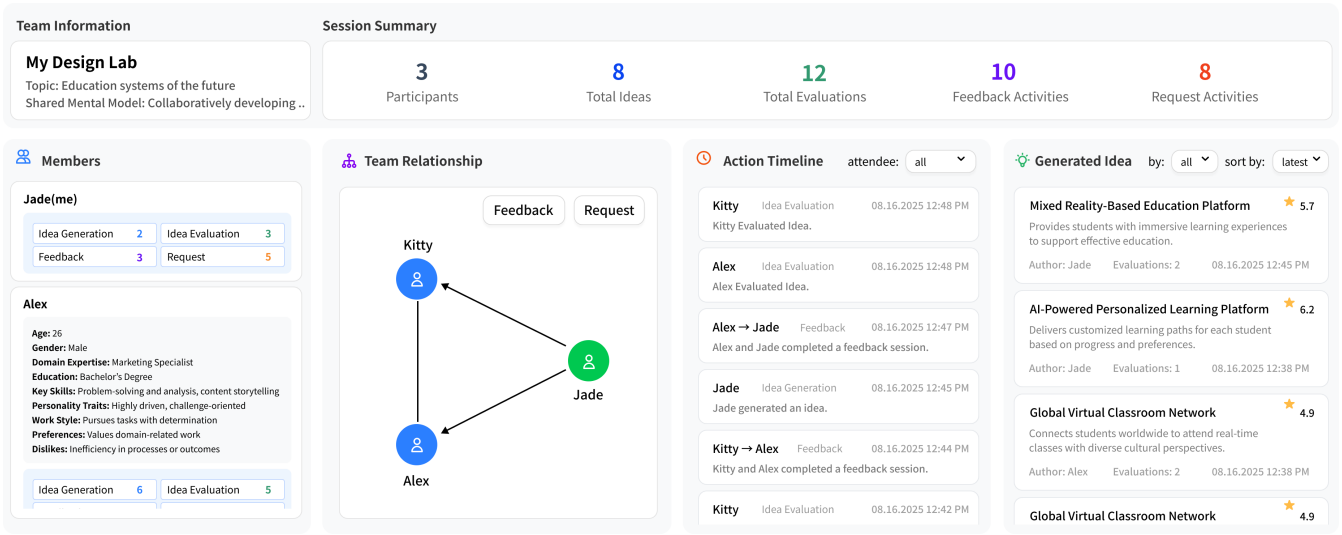


Figure 5: Reflection interface of CrafTeam. The interface consists of five main panels: (i) The Session Summary Panel displays team information and session metrics. (ii) The Members Panel shows each member's persona along with their per-role action counts. (iii) The Team Relationship Panel renders the team network, and when Feedback or Request mode is toggled, it visualizes the flow of these interactions between members. (iv) The Action Timeline Panel presents a chronologically ordered list of actions, featuring an attendee filter and expandable entries for detailed information. (v) The Generated Ideas Panel lists all ideas with their scores and metadata, where each item can be expanded to reveal detailed information.

3.4 Reflection Interface

CrafTeam provides a reflection phase that enables users to review their ideation sessions with their HMATs and gain insights for team improvement. Once users complete their ideation session, they can proceed to the reflection interface by clicking the "Review Team Ideation" button on the Team-Based Ideation Interface. The Reflection Interface (Fig. 5) presents both team-level activity logs and individual member activity logs from the ideation session. Team-level logs show the total number of ideas generated and evaluated, patterns of Feedback and Request exchanges between members, and the final collection of ideas produced. Individual activity logs display how frequently each team member serves within their assigned roles.

4 User Study

In this study, we use CrafTeam to deepen our understanding of HMAT formation by exploring how users form their own HMATs and what considerations and strategies they employ when configuring these teams. Rather than conducting a comparative study to evaluate how well our system supports ideation, we conducted a single-condition study to closely observe creative professionals using CrafTeam, thereby gaining richer insight into how they adapt familiar team formation practices from human-only design teams when forming and working with HMATs for creative workflows. The study protocol was approved by our institution's Institutional Review Board (IRB).

Participant ID	Age	Gender	Occupation	Domain
P1	35	M	Service Designer	AI startup
P2	26	F	UI/UX Designer	Screen Interface
P3	27	M	UX Designer	Searching Platform
P4	26	F	UX Designer	Advertisement
P5	28	M	Service Designer	Searching Platform
P6	26	F	UX Designer	VR/AR
P7	32	M	Project Manager	Video Game
P8	35	F	UX Designer	Video Game
P9	26	F	UI/UX Designer	Screen Interface
P10	33	M	UX Designer	AI startup
P11	33	M	Product Manager	Matchmaking Platform
P12	26	F	UX Designer	VR/AR

Table 1: Demographic information of study participants.

4.1 Participants

A total of 12 participants (six females, six males) aged from 26 to 35 ($M = 29.42, SD = 3.82$) were recruited through IT industry communities in South Korea. To ensure participants had sufficient background to form design teams informed by real-world experience, we recruited design practitioners currently engaged in team-based design work at IT companies and with prior experience using AI agents in their workflows. Participants' demographic details are presented in Table 1. The study was conducted in a 200-minute session, and participants received 200,000 KRW (approximately USD 145) as compensation.

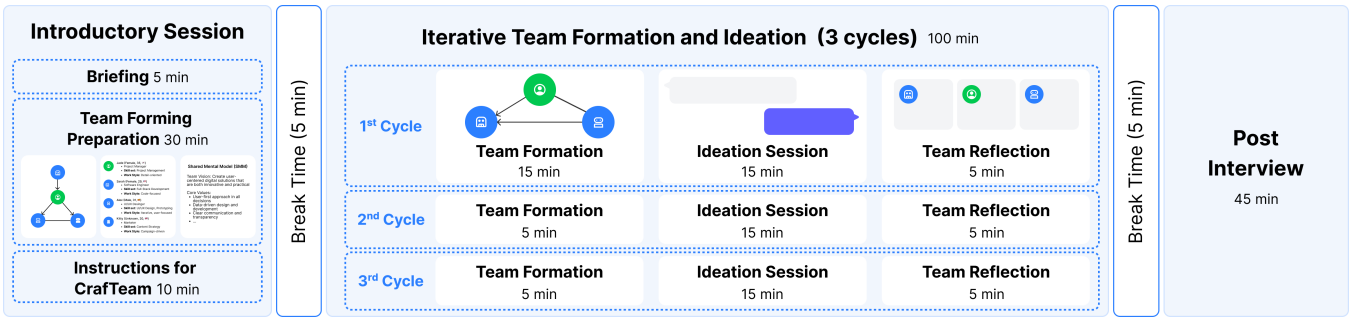


Figure 6: User study process. The study included a 45-minute introductory session, followed by three iterative cycles of team formation, ideation, and reflection (25 minutes each). After the cycles, participants engaged in a 45-minute post-interview to reflect on their experiences and share insights.

4.2 Procedure

The user study process is illustrated in Figure 6. Each study session lasted approximately three hours and consisted of four phases: (i) Introductory session, (ii) Iterative team formation and ideation, and (iii) Post-interview.

Introductory session. We first explained the purpose and procedure of the study to participants and obtained their signed consent. We informed participants that they would be composing teams with AI agents to perform ideation tasks and provided a brief tutorial on system usage.

Before using the CrafTeam, we conducted a team forming preparation session to help participants engage with HMAT formation. Participants first answered questions about their past creative teamwork experiences (e.g., most effective team experiences, their roles in teams) and their conception of ideal team conditions. They then completed three worksheets: (i) sketching an ideal team structure, (ii) creating profiles for desired team members, and (iii) defining the team's shared mental models. These activities helped participants reflect on HMATs' team formation and develop concrete ideas about the teams they wanted to implement in CrafTeam.

Iterative Team Formation and Ideation. After the introductory session, we introduced participants to the CrafTeam and its interfaces. We then asked them to complete three cycles of team formation, ideation, and reflection using CrafTeam. In the first team formation phase, we instructed participants to form a new HMAT based on their preparatory worksheets and freely choose topics for ideation from their professional experiences (15 minutes). In the second and third cycles of team formation, participants reconfigured their teams by modifying their previous setups (5 minutes each). For each ideation phase, participants conducted ideation with their newly formed team for 15 minutes. Before beginning the ideation, we emphasized that the primary goal was not to generate perfect ideas in the short time available, but to use the session as an evaluation to consider how to form an effective ideation team. Finally, after each ideation, we guided participants through a reflection phase, prompting them to evaluate team effectiveness and identify areas for improvement using system-provided activity logs.

Post-Interview. We concluded with 45-minute semi-structured interviews to explore participants' experiences in depth. The interview covered: insights gained from iteratively developing teams across three cycles, changes in perception regarding the importance of each dimension of team formation, experiences collaborating with AI agents, and suggestions for future HMAT design. All interviews were audio-recorded for transcription and analysis.

4.3 Data Analysis

We first conducted a descriptive statistical analysis to examine how participants formed and revised their teams. To analyze participants' trial-and-error processes, we quantified changes across cycles for the five dimensions of team formation that participants could adjust. To better understand these changes in team structure, we categorized the teams through iterative comparison, resulting in three distinct types (Fig. 7).

To gain a deeper qualitative understanding, we conducted open coding and thematic analysis [7]. In particular, our analysis aimed not to highlight considerations known from forming one-human-one-agent teams, but to identify unique considerations for forming HMATs, as well as the requirements of participants. The first PhD author open-coded the interview transcripts and interaction log data through multiple rounds of iteration. Two additional PhD researchers then reviewed the initial themes and supporting quotes, providing feedback on the coding. Based on this, the entire research team iteratively revised the themes, ultimately producing enhanced qualitative findings that captured both the factors participants considered when forming HMATs and the requirements they identified from their perspective.

5 Findings

In this section, we first present descriptive statistics on how participants form HMATs and conduct ideation with their teams using CrafTeam. We then describe the key considerations participants encountered when forming teams and the strategies they used to address them. Finally, we report the requirements for human-multi-agent teaming that participants identified, distinguishing it from both human-only and one-human-one-agent teams.

Dimension	Metric	Cycle 1	Cycle 2	Cycle 3	Total
Team Size	Size (M±SD)	5.00±0.82	4.08±0.95 ▼	4.75±1.01 ▲	4.61±1.02
Team Structure	Flat Team (N)	2	2 –	1 ▼	5
	Single-tier Hierarchy Team (N)	6	6 –	8 ▲	20
	Multi-tier Hierarchy Team (N)	4	4 –	3 ▼	11
Role Allocation	All: roles per member (M±SD)	3.22±0.75	2.82±1.02 ▼	2.91±1.06 ▲	2.99±0.97
	Agents: roles per member (M±SD)	3.27±0.76	2.68±1.07 ▼	2.82±1.12 ▲	2.95±1.02
	Agents in role: Idea Generation (M±SD)	3.58±0.95	2.42±0.64 ▼	3.08±0.95 ▲	3.03±0.99
	Agents in role: Idea Evaluation (M±SD)	3.50±1.44	1.92±1.04 ▼	2.58±1.04 ▲	2.67±1.35
	Agents in role: Feedback (M±SD)	3.83±2.00	2.00±1.29 ▼	2.75±1.64 ▲	2.86±1.57
	Agents in role: Requests (M±SD)	2.17±1.67	1.92±1.11 ▼	2.17±1.34 ▲	2.08±1.40
	User: roles per member (M±SD)	3.00±0.71	3.25±0.72 ▲	3.25±0.72 –	3.17±0.73
	User in role: Idea Generation (N)	3	5 ▲	5 –	13
	User in role: Idea Evaluation (N)	9	10 ▲	10 –	29
	User in role: Feedback (N)	12	12 –	12 –	36
	User in role: Requests (N)	12	12 –	12 –	36
Member Composition	User (% attributes specified)	87.96%	87.96%—	87.96%—	87.96%
	Agents: Social Identity (% attributes specified)	90.10%	89.19% ▼	93.33% ▲	90.96%
	Agents:Personal Identity (% attributes specified)	96.88%	97.30% ▲	96.67% ▼	96.92%
	Agents:Personal Life Context (% attributes specified)	82.64%	83.78% ▲	82.96% ▼	83.08%
Shared Mental Model	Text length (syll.; M±SD)	226.58±108.47	194.67±100.81 ▼	190.25±109.23 ▼	203.83±107.47

Table 2: Descriptive statistics of participants' team formation dimensions across three cycles. Markers show change vs. previous cycle: ▲ increase, ▼ decrease, – no change.

5.1 Descriptive Statistics of CrafTeam Usage

We first describe how participants formed their teams, focusing on the HMAT formation dimensions they specified, and then report how ideation unfolded within these teams. To shed light on the trial-and-error process through which participants formed their teams, we report how teams evolved across cycles. We denote participants as P1, and index each team by the cycle number following the participant ID (T1–T3). For example, P1T3 refers to P1's team in Cycle 3.

5.1.1 Participants' HMAT Formation Patterns. In total, participants formed 36 teams (12×3 cycles), exploring diverse formations—particularly in team structure and role allocation. Table 2 summarizes how participants configured the dimensions of team formation across cycles.

Team Size. Participants created teams averaging 4.61 members ($SD = 1.02, min = 3, max = 6$), including themselves. They initially formed teams averaging 5.00 members, scaled them down to 4.08 in the second cycle, and then rebounded to 4.75 in the final cycle.

Team Structure. We categorized team structures into three types based on how participants configured hierarchical relationships (Fig. 7): (i) Flat, (ii) Single-tier Hierarchy, and (iii) Multi-tier Hierarchy.

Single-tier Hierarchy teams were most common, with the user typically serving as the sole leader managing all agents directly. Only P2T1–T3 placed an AI at the top, and only P1T2 and P1T3 assigned multiple co-leaders. Multi-tier teams comprised a top manager, one or two mid-level managers, and several team members; users generally occupied the top role while AI agents served as mid-level managers, with the sole exception of P3T1, where the participant took a mid-level position. By the end of the study, participants tended to converge on Single-tier Hierarchy structures with themselves as the leader. This pattern indicates an initial exploration of various structural options, followed by a pragmatic preference for direct control, in which the user maintains a clear leadership role over all agents.

Role Allocation. Participants assigned averaged 2.99 roles per member. For AI agents, the number assigned to each role per team decreased from Cycle 1 to 2, then partially recovered in Cycle 3. Teams assigned an average of 3.58 AI agents to Idea Generation in Cycle 1, reduced to 2.42 in Cycle 2, then increased to 3.08 in Cycle 3. Idea Evaluation followed the same pattern with averages of 3.50, 1.92, and 2.58 agents. Feedback roles decreased from an average of 3.83 agents to 2.00, then partially recovered to 2.75. Request roles remained stable.

Users rarely participated in Idea Generation (36% of teams) but frequently took Idea Evaluation roles (81% of teams). All teams included users in the Feedback and Request roles.

Member Composition. Participants created 130 AI agent profiles to compose their HMATs. Completion rates across the three categories were: Personal Identity 96.96% (skills 100%, personality 94%); Social Identity 90.96% (age 90%, gender 87%, education 87%, occupation 100%); and Personal Life Context 83.08% (work style 94%, likes 78%, dislikes 78%). In composing teams, participants prioritized occupations and skill sets, while attributes such as gender and likes/dislikes were comparatively non-essential.

Shared Mental Model. Participants established SMMs averaging 203.83 syllables in length. Participants entered longer SMMs in Cycle 1 (226.58 syllables), then progressively shorter SMMs in Cycle

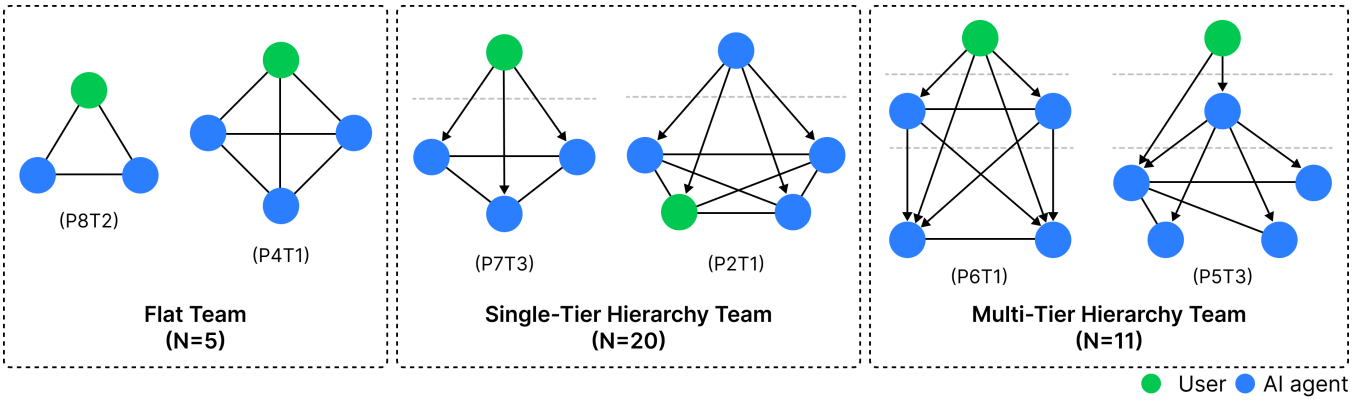


Figure 7: Three types of team structures emerged from teams formed by participants: (i) Flat teams where all members functioned at the same hierarchical level; (ii) Single-tier Hierarchy where one or a few leaders directly managed all other members; and (iii) Multi-tier Hierarchy where multiple layers of management existed between top leadership and base members.

2 (194.67 syllables) and Cycle 3 (190.25 syllables), demonstrating a decreasing trend across cycles.

5.1.2 Patterns of Actions in Team-Based Ideation. Across three team-based ideation sessions, participants engaged in ideation with their HMATs on a variety of topics, including future wearable AI devices (P1), a context-aware Next TV platform UI (P6), and conversational AI robots for children's emotional development (P9). In this section, we highlight key patterns in how participants and AI agents differed in their actions across the three ideation cycles, and present a detailed summary of these patterns in Appendix B.1.

Idea Generation. HMATs generated 451 ideas in total. Most of these ideas came from AI agents, whereas participants rarely generated ideas directly. Across teams, AI agents with the Idea Generation role produced an average of 4.37 ideas, whereas participants produced only 0.19 ideas (7 ideas in total). The volume of ideas remained similar across cycles (151, 146, and 154 ideas in Cycles 1–3), but each member's idea productivity shifted over time. The number of ideas generated by each member increased from an average of 3.49 in Cycle 1 to 5.07 in Cycle 2, when generation roles were concentrated on fewer agents, and then decreased slightly to 4.54 in Cycle 3 after roles were rebalanced.

Idea Evaluation. HMATs produced 428 evaluations total. Overall, participants gave relatively low but gradually increasing ratings across cycles, with substantial variability across ideas, from $M = 3.12$ ($SD = 1.66$) in Cycle 1 to $M = 4.06$ ($SD = 1.66$) in Cycle 2 and $M = 4.51$ ($SD = 1.37$) in Cycle 3. In contrast, AI agents consistently gave high and similar ratings across cycles, scoring $M = 5.36$ ($SD = 0.61$) in Cycle 1, $M = 5.37$ ($SD = 0.50$) in Cycle 2, and $M = 5.37$ ($SD = 0.43$) in Cycle 3, indicating that most ideas were evaluated within a narrow high range.

Feedback. HMATs conducted 358 feedback sessions, averaging 5.62 turns each. AI agents initiated 113 sessions in Cycle 1, dropping to 50 in Cycle 2, before partially recovering to 84 in Cycle 3. In contrast, participants became more active in giving feedback across cycles, from an average of 1.50 sessions in Cycle 1 to 2.50 in Cycle 2 and 3.17 in Cycle 3.

Requests. HMATs made 106 requests total. Participants overwhelmingly requested Idea Generation (49 requests), while only 15 requested Idea Evaluation and 8 requested Feedback. By contrast, AI agents showed a different distribution of requests: only 1 for Idea Generation, 12 for Idea Evaluation, and 21 for Feedback.

5.2 Participants' Considerations for HMATs Formation

During the preparation session, participants defined what they wanted their HMATs to achieve and how they expected them to operate. Most sought to enable high-quality ideation while minimizing their own workload, expecting agents to autonomously generate and elaborate ideas with minimal human intervention. During ideation, however, participants found that their teams did not operate as expected and incrementally refined their teams through iterative cycles. In this section, we illustrate the key considerations participants had to address when forming HMATs that differed from their initial expectations, and how they responded to these challenges.

5.2.1 Delegating Idea Generation to AI Agents. During team formation, all participants assigned the Idea Generation role to AI rather than generating ideas themselves. They viewed AI as more efficient because brainstorming and writing descriptions are cognitively taxing and time-consuming for humans, whereas AI can produce ideas instantly. They also highlighted reduced social pressure in team ideation, noting that AI produces candidate ideas immediately without the self-censorship that slows humans. As P12 noted, *"In real-world team ideation, people take a long time to voice ideas because they hesitate or resist having them evaluated by teammates. AI, however, provides ideas immediately without that burden, making it well-suited to the ideation role."*

Most participants believed that delegating ideation to AI agents was efficient in terms of work productivity; some participants (P5, P8, P10, and P11) questioned whether this division of roles actually yields high-quality ideas. They noted that many AI-generated ideas lacked the novelty or practical feasibility needed for real-world

application: *"I often use GPT for ideation and have a sense of what generative AI can do; what I see here is similar. For now, humans are better suited for creativity and should handle the truly creative and productive aspects. (P11)"*

5.2.2 AI Leaders Fall Short: Users Forced into Direct Management. Participants said that assigning a leader is necessary to keep multi-member teams aligned toward a unified direction. In this view, they considered themselves better suited to assume this role in HMATs, while also expecting that if AI could replace or support leadership, teams could operate more autonomously with minimal human intervention. Many participants attempted to deploy higher-level or middle-manager agents to assist with, or even fully automate, such management tasks.

However, participants reported that the AI leader did not fulfill the expected leadership role and did not improve the quality of ideas. They mainly attributed this to the agent's reluctance to take clear positions or make decisions. As P5 stated, *"In practice, meetings with a leader end with action items about what to do next and how to proceed. But the AI gave feedback without actually deciding anything and showed no clear preferences; it just kept the conversation going. As a result, ideas did not develop in a specific direction, and we felt like we were going in circles."* Participants also noted that management functions—such as task allocation and support for individual members—were not performed: *"I expected that when I made a request to the AI set as the leader, it would notify other team members as appropriate and assign tasks suited to each person's abilities. It did not. (P12)"*

Consequently, most participants gradually adopted a single-tier hierarchy, with themselves serving as the leader and directing the team. However, managing multiple agents alone proved burdensome, adding substantial workload: *"Even with only three AIs, whenever a feedback request came in, I had to double-check their prior work; when another agent requested feedback, I had to review that history too. As a result, keeping up was difficult because the agents moved quickly, while my capacity to interpret and provide feedback simultaneously was limited. (P3)"* Ultimately, participants either limited the number of agents or devised more effective management strategies.

5.2.3 Generalist to Specialist. Participants noted that teams where everyone's contributions were respected tended to achieve better results. In line with this view, they designed HMATs to support recognizing contributions and ensuring autonomy, enabling all members to generate ideas, evaluate them, and provide feedback rather than enforcing rigid, role-based divisions dictated by the AI structure.

However, after ideating with such AI teams, participants found that giving a single agent multiple roles slowed task progress and was inefficient. P9 remarked, *"When I initially allowed agents to handle every role, they performed each only superficially. As I clarified roles, for example, ideation versus feedback, the outcomes became more numerous and specific, so I assigned some agents to ideation only and others to feedback only."* Furthermore, because the structure placed the user in the manager role overseeing every AI agent, assigning multiple roles to each agent increased the coordination workload of participants. P12 explained, *"Initially, I gave them a high degree of freedom, but it was hard to track what each agent had done. So, rather than multiple roles, assigning each agent a single role proved more effective for keeping agents on track."*

Through iterative cycles, participants converged on assigning each agent a clearly defined role, most often dividing the team into ideation and evaluation. In turn, they emphasized the ability to operate with more extreme specialization, a possibility unique to AI teams, beyond what small, human-only teams can manage. P4 noted, *"In human teams, assigning roles like 'you can't ideate, so just evaluate' raises equity concerns, which makes this hard to do. With AI, those constraints don't apply, so if efficiency is the goal, even extreme role separation seems possible."* Several participants (P1, P5, P7) likened HMAT configuration to a strategy game and approached assignments from a team-first perspective rather than tailoring them to individual needs. P1 further argued for maximizing AI speed and uptime by predefining algorithms, operating policies, and action-selection rules to reduce real-time decision overhead.

5.2.4 More Agents do not guarantee Better Ideas: Diversifying Agents' Persona. Participants expected that adding more agents would increase productivity when forming HMATs. Most assigned two or more agents to Idea Generation and Evaluation. However, they found that output did not scale linearly, since agents often produced overlapping ideas or similar evaluations. The agents seemed to lack distinct perspectives, operating more like duplicates than diverse team members: *"At first, it felt less like several people brainstorming together and more like running three GPTs in parallel and assigning tasks to each. The only benefit seemed to be faster output through parallel processing, not better ideas. (P8)"*

During ideation, participants found that agents with different personas tended to generate ideas aligned with their backgrounds. In response, they diversified the personas of agents in the same role to broaden the scope of ideas. For example, P8 said, *"The agent with a designer persona seemed to focus on more design-oriented ideas, while the developer agent appeared to strive for more technical solutions. So in the next cycle, I added an agent with a civil servant persona and asked it to generate ideas related to national policy."* Some participants also created personas that do not exist in reality to elicit more unusual ideas. P7 suggested, *"To add variation to the team, it might be good to intentionally include an extreme member, something like a virus. I instilled a mindset such as 'You only think about this direction and are only interested in these things,' and when it interacts with other agents, it might lead to more unexpected outcomes."* Similarly, for idea evaluation, some participants (P3, P5, and P11) differentiated evaluator agents into personas that provided only positive feedback and those that provided only negative feedback, which yielded evaluations from more diverse perspectives.

5.3 User-Centered Requirements for Human–Multi-Agent Teaming

Through CrafTeam, participants had the opportunity to form teams of multiple AI agents and explore how this affected their ideation. All participants indicated that HMATs have potential in their practice and could be applied to other tasks beyond ideation. However, they also identified requirements distinct from human-only teams, underscoring what must be addressed for effective collaboration with multiple agents. In this section, we outline the requirements participants identified while using CrafTeam.

5.3.1 Challenges Around Interaction with Multi-Agents.

Inefficiency of Human-Like Communication Between Agents. In CrafTeam, participants could monitor how agents communicated with one another and how well each agent performed its assigned role. With this visibility, they could identify moments when agents behaved differently from their expectations and make fine adjustments. For example, P4 noted, *"I observed the agents exchanging feedback, but they were focusing too much on security issues. Since it was still the early stage of ideation, I told them to focus more on novelty in their feedback, and afterwards, they seemed to provide feedback more in line with what I had expected."*

However, several participants (P1, P5, and P11) questioned the premise that agent-to-agent communication should mimic human conversation. P11 said, *"If AI agents are conversing among themselves, the exchange need not be in natural language, nor always be visible to me. As it is, they seem to mimic humans, which creates a slight uncanny valley effect."* P1 argued that human-style dialogue can waste time and computational resources: *"Generative AI typically produces content faster than people, but conversation is different. Because each turn waits for the other's output, it can feel slower than human conversation, and forcing AI to converse this way may be inefficient."* P5 emphasized the need to revise how AI-to-AI dialogue is surfaced to users: *"I read some AI-to-AI conversations, but they were too noisy, and going through them felt like a waste of time. It would be better to show a summary or simply the outcome: the conclusion they reached."*

Needs for Team-level Interaction. Many participants (P2, P5, P6, P8, P10, and P12) mentioned a need for team-level interactions during group ideation that CrafTeam did not support. As P1 noted, *"In real work, we don't just stack one-on-one chats; we need group discussions where several people can talk at once. I wish the system supported that."* Some participants repurposed the shared mental model as a broadcast channel, adding guidance they wanted to disseminate, such as 'Don't suggest ideas for specific technologies' or 'Focus on IoT-based services.' They also wanted to integrate these broadcast instructions mid-session so that all agents would adjust in real time during the ideation. Participants further requested a more dynamic, multi-party exchange with overlapping contributions and rapid floor shifts, rather than strictly sequential turns. P6 said, *"When brainstorming with people, everyone takes turns, but talk still overlaps. When two are speaking, others aren't only observing. Three may jump in, then four, then it returns to two. I wish those transitions felt more natural."* They especially wanted to join ongoing agent-to-agent exchanges at any moment, including interrupting or steering them.

5.3.2 Rethinking Team Growth in HMATs.

Drawing on real-world workplace experience, participants treated a team's potential for sustained growth as one of the key considerations in team formation. The most common strategy was adapting master–apprentice pairings to HMATs. P6 explained, *"Senior–junior pairs work well in practice because seniors benefit from assistance, and juniors can develop over time through that collaboration."* P10 also experimented with a competitive structure to develop and refine ideas, dividing the team into two competing groups with the expectation that both would improve over time.

Participants viewed these strategies as less effective for AI agents, arguing that HMATs require different team-development approaches than human-only teams. P8 noted, *"With human teams, once people are hired, it's difficult to dismiss them, so it's crucial to help selected members grow. With AI, you can simply swap in a better model, so investing in a weak agent is less necessary."* Consequently, participants suggested focusing less on cultivating specific agents and more on iteratively replacing them to better fit the team's structure and role requirements. For example, P1 proposed: *"If we simulate 100 AI teams, award points to teams that generate strong ideas, and then select the top performer, akin to reinforcement learning, we could identify an optimal AI team."*

6 Discussion

In this study, we investigated how participants formed HMATs and collaborated on creative ideation tasks. Through CrafTeam, participants initially attempted to form autonomously operated teams in which AI agents collectively assumed both generative and reflective roles. However, our findings reveal that, because AI agents who have to lead the team struggled to provide value judgments and set directions, which are essential for idea development, participants shifted to team formations in which they directly orchestrated the agents and guided the ideation process. In this section, we examine the challenges of automated loops in HMATs and explore how human-orchestrated teams emerged to address these limitations. We then discuss design considerations for HMAT formation that enable users to effectively orchestrate multiple agents through scalable multi-party communication and progressive team evolution.

6.1 Breaking the Unproductive Loop: Human-Orchestrated HMATs

In developing HAT for co-creation, a key consideration has been how to distribute roles between AI agents and humans [53, 65]. Creative workflows that evolve through iterative processes involve two primary roles: the generative role, which generates creative outputs, and the reflective role, which evaluates outputs or provides reflective questions to facilitate further development [65, 88]. Previous studies have explored various trade-offs in these role distributions—when AI assumes the generative role, diverse idea exploration becomes possible but user agency weakens; when AI takes the reflective role, it induces deeper user reflection but the burden of idea generation remains with humans [88].

Our study enabled participants to configure HMAT formations beyond single-agent constraints, particularly by assigning AI agents both generative and reflective roles to automate iterations of the ideation process. However, contrary to their expectations, the automated interactions among agents often devolved into unproductive loops: lacking the capacity to direct and prioritize ideas, the agents repeatedly circled around similar concepts rather than advancing the ideation. Our participants noted that while the evaluations and feedback provided by agents were generally valid, their lack of personal preferences and inability to make value judgments prevented them from giving clear direction. This observation aligns with prior research indicating that, while AI agents can readily surface a wide range of alternatives, they struggle to exercise the

subjective value judgments and directional choices that move creative work forward [22]. This limitation reflects not only current agent capabilities but also the fundamental nature of creative teaming, where progress depends on subjective value judgments and directional choices in creative work that has no predefined correct answer. It therefore points to the need to ensure that humans hold the steering wheel to provide the direction that creative work inherently requires.

Consequently, most participants ultimately assumed the reflective role themselves to break the unproductive loop, directly evaluating ideas and providing direction. While they actively took on the reflective role in ideation, they continued to explore ways of leveraging multi-agent setups without relying solely on automated ideation. For instance, they operated multiple ideation threads in parallel to secure broader exploration spaces than traditional single HAT, or employed assistant agents that offered alternative reflective perspectives to scaffold human judgment. As they reconfigured how they worked with agents, their role evolved from a narrow reflective role to that of an orchestrator coordinating an entire team. In this role, they synthesized outputs from each thread, set priorities, and decided which ideas to develop next, serving as the central axis that enabled multiple agents to function harmoniously as a team.

Taken together, we suggest forming HMATs in ways that position humans as orchestrators of multiple agents rather than delegating primary control to fully autonomous agent teams. Prior work in human–robot teaming has similarly suggested formations where humans orchestrate multiple agents, acting as managers or supervisors [12, 21]. However, in creative work with multiple agents, orchestration involves more than a managerial role. Our findings revealed that users needed to lead the team while simultaneously acting as both managers and active contributors who coordinate the process and set the creative direction. This expanded role, while unlocking the potential of human-orchestrated HMATs, also places a significant burden on users and makes team performance heavily dependent on their capabilities. These tensions surface a new central design question: how should HMATs be designed so that teams remain human-orchestrated while reducing the cognitive and managerial load required for users to fulfill this demanding role? We therefore call for future research that investigates more fine-grained and context-specific human-orchestrated HMAT formations that keep users actively involved while reducing their burden. This agenda extends not only to HMAT formation itself but also to the interaction and system design required to realize such formations. In the following section, we discuss the specific challenges users encounter when orchestrating multiple agents and propose design considerations to address them.

6.2 Toward Scalable Human Orchestration: Supporting Multi-Party Communications from the User's Perspective

While participants adopted this human-orchestrated formation to leverage diverse perspectives from multiple agents, managing them simultaneously imposed substantial cognitive load. Our participants noted that this challenge stemmed from a lack of orchestration-specific support in CrafTeam, pointing out that its current interaction and interface design were not optimized for orchestrating multiple agents. This underscores that establishing effective human-orchestrated teams requires more than just identifying the optimal team formation, but also demands the deliberate design of interactions and interfaces that enable users to orchestrate these formations.

Previous HAT research has emphasized the importance of continuous communication and coordination for aligning goals and thoughts among multiple independent entities performing distributed tasks [91]. Building on this foundation, HMATs introduce new communication challenges as they require orchestrating interactions among more than two members. Our CrafTeam enabled one-on-one interactions with multiple agents but did not support simultaneous multi-agent interactions, and participants noted this absence as a barrier to real teamwork. For instance, participants struggled to implement situations where all team members participate simultaneously, such as brainstorming sessions, or where they provide direction to the entire team. As participants suggested, future systems should explore interaction methods that allow conducting discussions with multiple agents simultaneously or issuing commands to multiple agents at once. Our findings also revealed that participants directly controlled shared mental models to convey information efficiently without the need for detailed explanations. This finding supports prior research proposing group-level communication strategies, where users treat multiple agents as cohesive groups rather than managing each individually to reduce cognitive burden in multi-agent orchestration [70]. Building on this insight, future work could explore information injection methods beyond direct dialogue when designing HMAT communication systems.

Beyond the challenges of multi-party communication, our findings reveal that participants needed to observe inter-agent interactions to manage their teams. However, they struggled with this monitoring task due to cognitive overload from the sheer volume of agent-to-agent communications, making it challenging to track team dynamics and identify unproductive patterns. Furthermore, requiring agents to communicate in natural language for human comprehension may be inherently inefficient—as previous research has noted, forcing AI to use human-style dialogue can waste time and computational resources when agents could communicate more efficiently through other means [8, 91]. This presents a fundamental tension between the need for transparency in agent interactions and the practical limitations of human attention and processing capacity. Future research should investigate how to present inter-agent interactions to users in a way that balances transparency with cognitive manageability. These findings suggest opportunities to explore new user interfaces for HMATs that specifically address the challenges of multi-party communication and observing inter-agent interactions while managing multiple agents.

6.3 Progressive Team Evolution: Human-Aligned HMATs Through Iterative Refinement

Our findings revealed that participants initially struggled to form teams that functioned as they intended or expected. Despite these

early difficulties, we observed participants iteratively developing their teams through trial and error, gradually evolving them into configurations capable of producing ideas that aligned with their personal expectations and ideation contexts. This progressive team development has long been recognized as crucial in human-only teams, with foundational models like Tuckman's team development model (forming, storming, norming, performing) demonstrating how teams evolve through conflict, coordination, and collaboration [6, 48, 76]. This suggests that effective team formation is built over time and interaction, as it involves numerous complex factors that can produce unexpected dynamics and emergent behaviors. Particularly when forming HMATs, it becomes crucial that users themselves take charge of team formation so that the resulting configurations are well aligned with their nuanced goals and working styles. Therefore, we emphasize that HMAT design should focus not on forming perfect teams from the outset, but on empowering users to form and personalize their teams through progressive development.

Prior work on human team formation has proposed tools and strategies that support progressive approaches to team design, enabling people to quickly experience different team configurations and iteratively refine them based on observed outcomes [48, 85]. Building on this line of work, our findings extend progressive team formation to HMATs via CrafTeam, which allows users to form teams with customized AI agents. However, through the process of developing and studying CrafTeam, we found that progressive team growth in HMATs differs in important ways from human team development. Whereas human team formation typically involves recruiting from a bounded pool of candidates and reconfiguring teams at substantial social and organizational cost, HMATs let users instantiate and revise AI agents with comparatively little friction. While this flexibility lets users consider many more plausible team formations, our findings show that adding or reshuffling agents is not reliably associated with increased team performance. In HMATs, progressive improvement thus takes a different form: rather than slowly refining teams drawn from a bounded pool, users tend to rapidly explore, compare, and prune many alternative team formations while continually deciding which agents to instantiate, retain, or retire and how to structure roles within the team. Future research should explore methods for empowering users to progressively form and personalize HMAT formation within such iterative cycles, offering appropriate freedom to explore this enlarged design space without overwhelming them with complexity.

In addition, our findings revealed that participants with practical teamwork experience considered how team members could progressively grow in capability as they performed ideation tasks. However, HMATs required fundamentally different strategies for both team formation development and individual members' growth compared to human teams. Our findings showed that traditional team growth strategies leveraging human psychology—such as inducing competition or fostering senior-junior collaboration—proved ineffective with AI agents, as they lack emotional motivation and social learning capabilities. Rather than emphasizing the capability development of individual team members as in human teams, participants found it more practical to rapidly replace underperforming AI agents with better-configured alternatives. While prior work has proposed similar approaches that employ reinforcement learning and other training techniques to improve the composition of multiple agents for MAS [47], HMATs require not only more capable agents but also agents whose roles and behaviors stay aligned with a particular user's goals, values, and expectations for how the team should function. Taken together, these findings point to future work on HMATs that enable users to actively develop both the overall team formation and the capabilities of individual agents over time.

6.4 Limitations and Future Works

In this section, we discuss the limitations of our study that could impact the generalization of our findings.

First, we conducted an exploratory study with 12 participants, which was not a longitudinal investigation. While this approach provided deep insights into HMAT formation, the limited sample size and duration may not capture the full spectrum of strategies that might emerge over the long term. Additionally, we relied on users' subjective judgments rather than measuring actual ideation performance. As a result, we were not able to quantitatively evaluate how the team formations proposed by participants affected team outcomes, such as improvements in idea quality. Although the present study does not aim to identify a single optimal team configuration, the considerations and hypotheses we propose in the discussion require further empirical investigation. Future work should therefore examine how different team formations affect creative outputs, using metrics such as idea quality, novelty, diversity, and feasibility.

Second, our investigation focused exclusively on ideation tasks using iterative divergent-convergent processes. This narrow scope may not generalize to other creative workflows requiring different collaboration patterns. Sequential workflows in software development or design projects might require fundamentally different orchestration strategies. Similarly, creative tasks that require specialized expertise may require different approaches to the distribution of human-agent roles. Further research is needed to understand how HMATs should be formed across diverse creative domains and workflow types.

Third, we examined only single-human multi-agent teams, leaving questions about scenarios with multiple human collaborators. Multi-human HMATs introduce additional complexity in coordination, authority distribution, and conflict resolution that our study did not address. When multiple humans each orchestrate their own agents while collaborating toward shared goals, the interaction patterns are likely to differ substantially from those in single-human scenarios. Future work should investigate these multi-human–multi-agent team formations and develop guidelines for scaling HMATs beyond individual use.

Lastly, our study did not deeply explore the ethical implications of HMATs in creative work. Previous HAT research has documented risks of over-reliance on AI and diminished human autonomy [18], and with multiple agents, these risks may compound. In fact, most participants delegated idea generation to AI agents, which could lead to a diminished sense of creative agency and ownership. As HMATs become prevalent, future research should investigate how to preserve human authorship and accountability when forming multi-agent teams, developing guidelines that balance AI assistance with human creative autonomy in team design decisions.

7 Conclusion

In this study, we developed CrafTeam, a technology probe that enables users to form and collaborate with HMATs, allowing us to observe how users specify HMAT formations in practice and surface practical considerations around HMAT formation. Through a three-hour user study with 12 IT design practitioners, in which participants iteratively formed and refined teams across three cycles, we examined how users form HMATs and leverage them in creative ideation processes. We found that while participants initially attempted to let teams operate autonomously, they soon discovered limitations in AI's ability to make value judgments and express clear preferences, which are required for creative work. Consequently, participants adopted a formation where they directly orchestrated agents to break unproductive loops and provide direction. Based on these findings, we emphasize the importance of designing HMATs centered on human orchestration and suggest design considerations that support multi-party communication and progressive team evolution. We hope this research provides insights and a future research agenda for designing human-orchestrated multi-agent teams.

Acknowledgments

This work was supported by a gift from Google (Google Multilab Project: Collective Curation: A Framework for Designing Human-Agent Collectives in Creative Work). We thank our participants for their engagement and the anonymous reviewers for their thoughtful comments and suggestions.

References

[1] 2017. *Towards a Better Design Team Formation: A Review of Team Effectiveness Models and Possible Measurements of Design-Team Inputs, Processes, and Outputs.* International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. Volume 3: 19th International Conference on Advanced Vehicle Technologies; 14th International Conference on Design Education; 10th Frontiers in Biomedical Devices. doi:10.1115/DETC2017-68091

[2] Aris Anagnostopoulos, Luca Becchetti, Carlos Castillo, Aristides Gionis, and Stefano Leonardi. 2012. Online team formation in social networks. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12)*. Association for Computing Machinery, New York, NY, USA, 839–848. doi:10.1145/2187836.2187950

[3] Robert W. Andrews, J. Mason Lilly, Divya Srivastava, and Karen M. Feigh. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science* 24, 2 (2023), 129–175. doi:10.1080/1463922X.2022.2061080

[4] Aitor Aritzeta, Stephen Swailes, and Barbara Senior. 2007. Belbin's Team Role Model: Development, Validity and Applications for Team Building. *Journal of Management Studies* 44, 1 (2007), 96–118. doi:10.1111/j.1467-6486.2007.00666.x

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. doi:10.1609/hcomp.v7i1.5285

[6] Denise A. Bonebright. 2010. 40 years of storming: a historical review of Tuckman's model of small group development. *Human Resource Development International* 13, 1 (2010), 111–120. doi:10.1080/13678861003589099

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[8] Andreas Bucher, Mateusz Dolata, Sven Eckhardt, Dario Staehelin, and Gerhard Schwabe. 2024. Talking to Multi-Party Conversational Agents in Advisory Services: Command-based vs. Conversational Interactions. *Proc. ACM Hum.-Comput. Interact.* 8, GROUP, Article 7 (Feb. 2024), 25 pages. doi:10.1145/3633072

[9] Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Tian Feng, Yujiu Yang, and Rongsheng Zhang. 2024. HoLLMwood: Unleashing the Creativity of Large Language Models in Screenwriting via Role Playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 8075–8121. doi:10.18653/v1/2024.findings-emnlp.474

[10] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. In *International Conference on Representation Learning*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (Eds.), Vol. 2024. 20094–20136. https://proceedings.iclr.cc/paper_files/paper/2024/file/578e65cdee35d00c708d4c64bce32971-Paper-Conference.pdf

[11] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1055, 25 pages. doi:10.1145/3613904.3642794

[12] Abhinav Dahiya, Alexander M. Aroyo, Kerstin Dautenhahn, and Stephen L. Smith. 2023. A survey of multi-agent Human–Robot Interaction systems. *Robotics and Autonomous Systems* 161 (2023), 104335. doi:10.1016/j.robot.2022.104335

[13] Griffin Dietz, Jane L. E, Peter Washington, Lawrence H. Kim, and Sean Follmer. 2017. Human Perception of Swarm Robot Motion. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2520–2527. doi:10.1145/3027063.3053220

[14] Shiying Ding, Xinyi Chen, Yan Fang, Wenrui Liu, Yiwu Qiu, and Chunlei Chai. 2023. DesignGPT: Multi-Agent Collaboration in Design. In *2023 16th International Symposium on Computational Intelligence and Design (ISCID)*. 204–208. doi:10.1109/ISCID59865.2023.00056

[15] Ali Dorri, Salil S. Kanhere, and Raja Jurdak. 2018. Multi-Agent Systems: A Survey. *IEEE Access* 6 (2018), 28573–28593. doi:10.1109/ACCESS.2018.2831228

[16] Kees Dorst. 2006. Design Problems and Design Paradoxes. *Design Issues* 22, 3 (2006), 4–17. doi:10.1162/desi.2006.22.3.4

[17] Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott R. Klemmer. 2011. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Trans. Comput.-Hum. Interact.* 17, 4, Article 18 (Dec. 2011), 24 pages. doi:10.1145/1879831.1879836

[18] Wen Duan, Christopher Flathmann, Nathan McNeese, Matthew J Scalia, Ruihao Zhang, Jamie Gorman, Guo Freeman, Shiwen Zhou, Allyson Ivy Hauptman, and Xiaoyun Yin. 2025. Trusting Autonomous Teammates in Human-AI Teams - A Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1102, 23 pages. doi:10.1145/3706598.3713527

[19] Wen Duan, Shiwen Zhou, Matthew J Scalia, Xiaoyun Yin, Nan Weng, Ruihao Zhang, Guo Freeman, Nathan McNeese, Jamie Gorman, and Michael Tolston. 2024. Understanding the Evolvement of Trust Over Time within Human-AI Teams. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 521 (Nov. 2024), 31 pages. doi:10.1145/3687060

[20] Hugo Figueroa, Rosanna Costaguta, María de los Ángeles Menini, and Daniela Missio. 2019. An Automatic Identification of Team Roles in Forums. In *Proceedings of the XX International Conference on Human Computer Interaction* (Donostia, Gipuzkoa, Spain) *(Interacción '19)*. Association for Computing Machinery, New York, NY, USA, Article 24, 2 pages. doi:10.1145/3335595.3335615

[21] Fei Gao, Missy L. Cummings, and Luca F. Bertuccelli. 2012. Teamwork in controlling multiple robots. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (Boston, Massachusetts, USA) *(HRI '12)*. Association for Computing Machinery, New York, NY, USA, 81–88. doi:10.1145/2157689.2157703

[22] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social Dynamics of AI Support in Creative Writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 245, 15 pages. doi:10.1145/3544548.3580782

[23] Pratik Ghosh and Sean Rintel. 2025. YES AND: A Generative AI Multi-Agent Framework for Enhancing Diversity of Thought in Individual Ideation for Problem-Solving Through Confidence-Based Agent Turn-Taking. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 607, 13 pages. doi:10.1145/3706599.3720142

[24] Diego Gómez-Zará, Matthew Paras, Marlon Twyman, Jacqueline N. Lane, Leslie A. DeChurch, and Noshir S. Contractor. 2019. Who Would You Like to Work With?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300889

[25] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 8048–8057. doi:10.24963/ijcai.2024/890 Survey Track.

[26] Hyunyoung Han, Kyungeun Jung, and Sang Ho Yoon. 2025. ChoreoCraft: In-situ Crafting of Choreography in Virtual Reality through Creativity Support Tool. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1059, 21 pages. doi:10.1145/3706598.3714220

[27] 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* (Newcastle upon Tyne, United Kingdom) *(CHIWORK '24)*. Association for Computing Machinery, New York, NY, USA, Article 9, 14 pages. doi:10.1145/3663384.3663398

[28] Junda He, Christoph Treude, and David Lo. 2025. LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision, and the Road Ahead. *ACM Trans. Softw. Eng. Methodol.* 34, 5, Article 124 (May 2025), 30 pages. doi:10.1145/3712003

[29] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. doi:10.1145/642611.642616

[30] Rehan Iftikhar, Yi-Te Chiu, Mohammad Saud Khan, and Catherine Caudwell. 2024. Human–Agent Team Dynamics: A Review and Future Research Opportunities. *IEEE Transactions on Engineering Management* 71 (2024), 10139–10154. doi:10.1109/TEM.2023.3331369

[31] Dongming Jin, Zhi Jin, Xiaohong Chen, and Chunhui Wang. 2024. MARE: Multi-Agents Collaboration Framework for Requirements Engineering. arXiv:2405.03256 [cs.SE] https://arxiv.org/abs/2405.03256

[32] Malte F. Jung, Jin Joo Lee, Nick DePalma, Sigurdur O. Adalgeirsson, Pamela J. Hinds, and Cynthia Breazeal. 2013. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1555–1566. doi:10.1145/2441776.2441954

[33] Vera C. Kaelin, Maitreyee Tewari, Sara Benouar, and Helena Lindgren. 2024. Developing teamwork: transitioning between stages in human-agent collaboration. *Frontiers in Computer Science* Volume 6 - 2024 (2024). doi:10.3389/fcomp.2024.1455903

[34] Abidullah Khan, Atefeh Shokrizadeh, and Jinghui Cheng. 2025. Beyond Automation: How Designers Perceive AI as a Creative Partner in the Divergent Thinking Stages of UI/UX Design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1105, 12 pages. doi:10.1145/3706598.3713500

[35] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376785

[36] Yewon Kim, Sung-Ju Lee, and Chris Donahue. 2025. Amuse: Human-AI Collaborative Songwriting with Multimodal Inspirations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 187, 28 pages. doi:10.1145/3706598.3713818

[37] Steve W.J. Kozlowski and Daniel R. Ilgen. 2006. Enhancing the Effectiveness of Work Groups and Teams. *Psychological Science in the Public Interest* 7, 3 (2006), 77–124. doi:10.1111/j.1529-1006.2006.00030.x PMID: 26158912.

[38] Jan Kratzer, Roger Th.A.J. Leenders, and Jo M.L. Van Engelen. 2008. The social structure of leadership and creativity in engineering design teams: An empirical analysis. *Journal of Engineering and Technology Management* 25, 4 (2008), 269–286. doi:10.1016/j.jengtecman.2008.10.004

[39] Theodoros Lappas, Kun Liu, and Evimaria Terzi. 2009. Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Paris, France) *(KDD '09)*. Association for Computing Machinery, New York, NY, USA, 467–476. doi:10.1145/1557019.1557074

[40] Keyeun Lee, Seo Hyeong Kim, Seolhee Lee, Jinsu Eun, Yena Ko, Hayeon Jeon, Esther Hehsun Kim, Seonghye Cho, Soeun Yang, Eun-mee Kim, and Hajin Lim. 2025. SPeCtrum: A Grounded Framework for Multidimensional Identity Representation in LLM-Based Agent. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 6971–6991. doi:10.18653/v1/2025.naacl-long.356

[41] Roger Th.A.J Leenders, Jo M.L van Engelen, and Jan Kratzer. 2003. Virtuality, communication, and new product team creativity: a social network perspective. *Journal of Engineering and Technology Management* 20, 1 (2003), 69–92. doi:10.1016/S0923-4748(03)00005-5 Special Issue on Research Issues in Knowledge Management and Virtual Collaboration in New Product Development.

[42] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 51991–52008. https://proceedings.neurips.cc/paper_files/paper/2023/file/a3621ee907def47c1b952ade25c67698-Paper-Conference.pdf

[43] Claire Liang, Julia Proft, Erik Andersen, and Ross A. Knepper. 2019. Implicit Communication of Actionable Information in Human-AI teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300325

[44] Hyunseung Lim, Ji Yong Cho, Taewan Kim, Jeongeon Park, Hyungyu Shin, Seulgi Choi, Sunghyun Park, Kyungjae Lee, Juho Kim, Moontae Lee, and Hwajung Hong. 2024. Co-Creating Question-and-Answer Style Articles with Large Language Models for Research Promotion. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 975–994. doi:10.1145/3643834.3660705

[45] Hyunseung Lim, Dasom Choi, DaEun Choi, Sooyohn Nam, and Hwajung Hong. 2025. Feed-o-meter: Investigating AI-generated mentee personas as interactive agents for scaffolding design feedback practice. *International Journal of Human-Computer Studies* (2025), 103687. doi:10.1016/j.ijhcs.2025.103687

[46] Yi-Cheng Lin, Kang-Chieh Chen, Zhe-Yan Li, Tzu-Heng Wu, Tzu-Hsuan Wu, Kuan-Yu Chen, Hung yi Lee, and Yun-Nung Chen. 2025. Creativity in LLM-based Multi-Agent Systems: A Survey. arXiv:2505.21116 [cs.HC] https://arxiv.org/abs/2505.21116

[47] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6382–6393.

[48] Ioanna Lykourentzou, Shannon Wang, Robert E. Kraut, and Steven P. Dow. 2016. Team Dating: A Self-Organized Team Formation Strategy for Collaborative Crowdsourcing. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1243–1249. doi:10.1145/2851581.2892421

[49] John E. Mathieu, Peter T. Gallagher, Monique A. Domingo, and Elizabeth A. Klock. 2019. Embracing Complexity: Reviewing the Past Decade of Team Effectiveness Research. *Annual Review of Organizational Psychology and Organizational Behavior* 6, Volume 6, 2019 (2019), 17–46. doi:10.1146/annurev-orgpsych-012218-015106

[50] John E. Mathieu, John R. Hollenbeck, Daan van Knippenberg, and Daniel R. Ilgen. 2017. A century of work teams in the Journal of Applied Psychology. 102, 3 (2017). doi:10.1037/apl0000128

[51] Nathan J. McNeese, Mustafa Demir, Nancy J. Cooke, and Christopher Myers. 2018. Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors* 60, 2 (2018), 262–273. doi:10.1177/0018720817743223 PMID: 29185818.

[52] Nathan J. McNeese, Mustafa Demir, Nancy J. Cooke, and Manrong She. 2021. Team Situation Awareness and Conflict: A Study of Human–Machine Teaming. *Journal of Cognitive Engineering and Decision Making* 15, 2-3 (2021), 83–96. doi:10.1177/15553434211017354

[53] Caterina Moruzzi and Solange Margarido. 2024. Customizing the Balance between User and System Agency in Human-AI Co-Creative Processes. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC'24)*, Kazjon Grace, Maria Teresa Llano, Pedro Martins, and Maria M. Hedblom (Eds.). Association for Computational Creativity, Jönköping, Sweden, 108–117. https://computationalcreativity.net/iccc24/papers/ICCC24_paper_15.pdf

[54] Suchismita Naik, Austin L. Toombs, Ph.D. Snellinger, Amanda, Scott Saponas, and Amanda K Hall. 2025. Designing with Multi-Agent Generative AI: Insights from Industry Early Adopters. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 1961–1972. doi:10.1145/3715336.3735823

[55] Minh Huynh Nguyen, Thang Phan Chau, Phong X. Nguyen, and Nghi D. Q. Bui. 2025. AgileCoder: Dynamic Collaborative Agents for Software Development based on Agile Methodology. In *2025 IEEE/ACM Second International Conference on AI Foundation Models and Software Engineering (Forge)*. 156–167. doi:10.1109/Forge66646.2025.00026

[56] Moeka Nomura, Takayuki Ito, and Shiyao Ding. 2024. Towards Collaborative Brain-storming among Humans and AI Agents: An Implementation of the IBIS-based Brainstorming Support System with Multiple AI Agents. In *Proceedings of the ACM Collective Intelligence Conference* (Boston, MA, USA) *(CI '24)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3643562.3672609

[57] Thomas A. O'Neill, Christopher Flathmann, Nathan J. McNeese, and Eduardo Salas. 2023. Human-autonomy Teaming: Need for a guiding team-based framework? *Computers in Human Behavior* 146 (2023), 107762. doi:10.1016/j.chb.2023.

107762
[58] 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 (2022), 904–938. doi:10.1177/0018720820960865 PMID: 33092417.
[59] Jeongeon Park, Bryan Min, Kihoon Son, Jean Y. Song, Xiaojuan Ma, and Juho Kim. 2025. ChoiceMates: Supporting Unfamiliar Online Decision-Making with Multi-Agent Conversational Interactions. arXiv:2310.01331 [cs.HC] https://arxiv.org/abs/2310.01331
[60] 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* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. doi:10.1145/3586183.3606763
[61] Praveen Paruchuri, Pradeep Varakantham, Katia Sycara, and Paul Scerri. 2010. Effect of Human Biases on Human-Agent Teams. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 2. 327–334. doi:10.1109/WI-IAT.2010.104
[62] Jill E. Perry-Smith and Christina E. Shalley. 2003. The Social Side of Creativity: A Static and Dynamic Social Network Perspective. *Academy of Management Review* 28, 1 (2003), 89–106. doi:10.5465/amr.2003.8925236
[63] Hamed Rezaee and Farzaneh Abdollahi. 2015. Average Consensus Over High-Order Multiagent Systems. *IEEE Trans. Automat. Control* 60, 11 (2015), 3047–3052. doi:10.1109/TAC.2015.2408576
[64] Jeba Rezwana and Mary Lou Maher. 2023. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 67 (Sept. 2023), 28 pages. doi:10.1145/3519026
[65] Aaron Schecter and Benjamin Richardson. 2025. How the Role of Generative AI Shapes Perceptions of Value in Human-AI Collaborative Work. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 530, 15 pages. doi:10.1145/3706598.3713946
[66] 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. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 13 (Jan. 2022), 29 pages. doi:10.1145/3492832
[67] Matthias Scheutz, Scott A. DeLoach, and Julie A. Adams. 2017. A Framework for Developing and Using Shared Mental Models in Human-Agent Teams. *Journal of Cognitive Engineering and Decision Making* 11, 3 (2017), 203–224. doi:10.1177/1555343416682891
[68] Jan B. Schmutz, Neal Outland, Sophie Kerstan, Eleni Georganta, and Anna-Sophie Ulfert. 2024. AI-teaming: Redefining collaboration in the digital era. *Current Opinion in Psychology* 58 (2024), 101837. doi:10.1016/j.copsyc.2024.101837
[69] Axel Schulte, Diana Donath, and Douglas S. Lange. 2016. Design Patterns for Human-Cognitive Agent Teaming. In *Engineering Psychology and Cognitive Ergonomics*, Don Harris (Ed.). Springer International Publishing, Cham, 231–243.
[70] Sarah Schömbs, Yan Zhang, Jorge Goncalves, and Wafa Johal. 2025. From Conversation to Orchestration: HCI Challenges and Opportunities in Interactive Multi-Agentic Systems. arXiv:2506.20091 [cs.HC] https://arxiv.org/abs/2506.20091
[71] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F. Jung. 2020. Robots in Groups and Teams: A Literature Review. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 176 (Oct. 2020), 36 pages. doi:10.1145/3415247
[72] Hanshu Shen, Lyukesheng Shen, Wenqi Wu, and Kejun Zhang. 2025. IdeationWeb: Tracking the Evolution of Design Ideas in Human-AI Co-Creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 146, 19 pages. doi:10.1145/3706598.3713375
[73] Tianqi Song, Yugin Tan, Zicheng Zhu, Maojia Song, Feng Yibin, and Yi-Chieh Lee. 2025. The More, The Stronger? Investigating How Multi-Agent AI Shapes Human Opinions. In *ICLR 2025 Workshop on Human-AI Coevolution*. https://openreview.net/forum?id=6zlttMWe4G
[74] Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas Griffiths, and Faeze Brahman. 2024. MacGyver: Are Large Language Models Creative Problem Solvers?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 5303–5324. doi:10.18653/v1/2024.naacl-long.297
[75] Christine A. Toh and Scarlett R. Miller. 2016. Creativity in design teams: the influence of personality traits and risk attitudes on creative concept selection. *Research in Engineering Design* 27, 1 (01 Jan 2016), 73–89. doi:10.1007/s00163-015-0207-y
[76] Bruce W. Tuckman. 1965. Developmental sequence in small groups. 63, 6 (1965). doi:10.1037/h0022100
[77] Julie M. Urban, Clint A. Bowers, Susan D. Monday, and Ben B. Morgan Jr. 1995. Workload, Team Structure, and Communication in Team Performance. *Military Psychology* 7, 2 (1995), 123–139. doi:10.1207/s15327876mp0702_6
[78] Jo M. L. van Engelen, Derk Jan Kiewiet, and Pieter Terlouw. 2001. Improving Performance of Product Development Teams through Managing Polarity. *International Studies of Management & Organization* 31, 1 (2001), 46–63. doi:10.1080/00208825.2001.11656807
[79] L. Vig and J.A. Adams. 2006. Multi-robot coalition formation. *IEEE Transactions on Robotics* 22, 4 (2006), 637–649. doi:10.1109/TRO.2006.878948
[80] James C. Walliser, Ewart J. de Visser, Eva Wiese, and Tyler H. Shaw. 2019. Team Structure and Team Building Improve Human–Machine Teaming With Autonomous Agents. *Journal of Cognitive Engineering and Decision Making* 13, 4 (2019), 258–278. doi:10.1177/1555343419867563
[81] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 84 (April 2024), 26 pages. doi:10.1145/3637361
[82] Yun Wan and Yoram M Kalman. 2025. Using Generative AI Personas Increases Collective Diversity in Human Ideation. arXiv:2504.13868 [cs.HC] https://arxiv.org/abs/2504.13868
[83] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (22 Mar 2024), 186345. doi:10.1007/s11704-024-40231-1
[84] Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2025. User Behavior Simulation with Large Language Model-based Agents. *ACM Trans. Inf. Syst.* 43, 2, Article 55 (Jan. 2025), 37 pages. doi:10.1145/3708985
[85] Mark E. Whiting, Irena Gao, Michelle Xing, N'godjigui Junior Diarrassouba, Tonya Nguyen, and Michael S. Bernstein. 2020. Parallel Worlds: Repeated Initializations of the Same Team to Improve Team Viability. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 67 (May 2020), 22 pages. doi:10.1145/3392877
[86] Hyeongon Wi, Seungjin Oh, Jungtae Mun, and Mooyoung Jung. 2009. A team formation model based on knowledge and collaboration. *Expert Systems with Applications* 36, 5 (2009), 9121–9134. doi:10.1016/j.eswa.2008.12.031
[87] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. 2007. The Increasing Dominance of Teams in Production of Knowledge. *Science* 316, 5827 (2007), 1036–1039. doi:10.1126/science.1136099
[88] Xiaotong (Tone) Xu, Arina Konnova, Bianca Gao, Cindy Peng, Dave Vo, and Steven P. Dow. 2025. Productive vs. Reflective: How Different Ways of Integrating AI into Design Workflows Affect Cognition and Motivation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 24, 15 pages. doi:10.1145/3706598.3713649
[89] Zeda Xu, Chloe Soohwa Hong, Nicolás F. Soria Zurita, Joshua T. Gyory, Gary Stump, Hannah Nolte, Jonathan Cagan, and Christopher McComb. 2024. Adaptation Through Communication: Assessing Human–Artificial Intelligence Partnership for the Design of Complex Engineering Systems. *Journal of Mechanical Design* 146, 8 (02 2024), 081401. doi:10.1115/1.4064490
[90] Zixiao Yu, Tingru Cui, Chen Luo, and Dilang Tan. 2025. A Systematic Literature Review on Human-Agent Teaming with Insights into Multi-Agent Interactions. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS) 2025*, Vol. 20. https://aisel.aisnet.org/pacis2025/hci/hci/20 Track 5: Human Computer Interaction, Paper Number: PACIS2025-1614.
[91] 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. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 281 (Oct. 2023), 31 pages. doi:10.1145/3610072
[92] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 246 (Jan. 2021), 25 pages. doi:10.1145/3432945

A Detail of System

A.1 Design of LLM-Based Agent for HMATs

In this section, we describe how we designed AI agents for HMATs in CrafTeam. To enable ideation with multi-agents, we designed agents capable of autonomous interaction while adhering to user customizations (Fig 8). We adopted the generative agent architecture [60, 83] as our foundation, creating LLM-based agents with a behavioral framework that enables dynamic interaction and environmental responsiveness. This framework operates through four distinct states: (i) Plan, (ii) Act, (iii) Reflect, and (iv) Wait. Users can monitor each agent's state in real-time through the Team Status Tab (Fig. 3.B).

- **Plan State**: In the Plan state, the agent determines its next action. It first checks for any incoming requests from the user or other agents. If a request is present, the agent plans to address it. If there are no pending requests, the agent autonomously selects an action to perform based on its assigned roles.
- **Act State**: In the Act state, the agent executes a single action corresponding to its assigned roles. The possible actions include: Idea Generation, Idea Evaluation, Feedback (and Feedback Response), and Request.
- **Reflect State**: In the Reflect state, the agent processes new information and updates its internal memory. This state is typically triggered after the agent receives an evaluation for an idea it generated or at the conclusion of a feedback session.
- **Wait State**: In the Wait state, the agent remains idle and performs no actions. After a duration of 30-60 seconds, it automatically transitions back to the Plan state. This period serves as a buffer to prepare for new interactions. If the agent receives a direct request or feedback during this time, it immediately transitions to the Act state to respond.

To implement this behavioral framework, we designed three core components: a profiling module (Section A.1.1), a memory module (Section A.1.2), and a set of pipelines to execute each action (Section A.1.3).

A.1.1 Profiling Module. We designed a profiling module to enable agents to simulate behaviors according to user-defined profiles. Following prior research [40, 84], the profiling module guides the LLMs to embody a character by emphasizing how traits manifest in personal and social contexts rather than merely listing profile attributes. Upon completing team building, the module generates tailored prompts for each agent based on the agent's assigned profile and the team's shared mental model. These prompts translate the profile into behaviorally grounded guidance for the agent, supporting more authentic persona simulation.

A.1.2 Memory Module. We designed a memory module to enable agents to remember their prior interactions and reflect on them in subsequent actions. Building on prior research on simulating human-like agents [84], we designed each agent with both Short-Term Memory for recent interaction and Long-Term Memory for enduring information.

Short-Term Memory stores three types of information: (i) five most recent actions to inform subsequent decisions, (ii) a queue of incoming Requests queued for later processing, and (iii) the running transcript of any ongoing Feedback conversation. This memory is populated in real-time and is transient, as its contents are cleared once a request is performed or a feedback session concludes.

Long-Term Memory stores three components: (i) Design Knowledge, (ii) Action Strategies, and (iii) Relationships. Design Knowledge records ideation-relevant facts, constraints, and examples accumulated during the session. Action Strategies specify what the agent may do and how for each role-permitted action (Idea Generation, Idea Evaluation, Feedback, Requests) and the Plan. Given the team-based interaction setting of HMATs, we introduce Relationships, which capture the agent's own working beliefs about other members—who they are and how they connect to the agent (roles and links), salient interaction history with each, and perceived reliability or responsiveness. The agent uses this to decide whom to engage or which action to direct to whom.

A.1.3 Action Pipelines. To govern the behavior of AI agents in CrafTeam, we designed LLM-based action pipelines that direct agents to perform defined tasks. We implemented two categories of actions: (i) foundational actions, which represent the default capabilities inherent to every agent, such as Plan, Reflection, and Feedback Response; and (ii) role-permitted actions, which are actions specifically assigned to each agent based on their role—such as Idea Generation, Idea Evaluation, Feedback, and Request.

All prompts of LLM-based pipelines consist of both system prompts and user prompts. The system prompt includes the agent profile prompt generated in the profiling module, with additional memory and contextual information assigned to both the system and user prompts depending on the specific action. When a request triggers an action (e.g., Idea Generation, Idea Evaluation, or Feedback), the corresponding request details are consistently appended to the prompt. Following is a detailed description of the pipeline for each action.

Plan Pipeline. The Plan pipeline determines the following action an AI agent should take. Its inputs include the agent's recent behaviors from short-term memory, as well as design knowledge and relevant action plans from long-term memory. The output requires the agent to either select one of its assigned actions or choose to wait, and to generate a rationale for this decision to support Chain-of-Thought prompting.

Idea Generation Pipeline. The Idea Generation pipeline enables AI agents to generate ideas. First, through a pre-generation prompt, the agent determines its ideation strategy. The inputs include a list of existing ideas, knowledge from long-term memory, and ideation-related action plans stored in long-term memory. The outputs comprise a decision on whether to create a new idea or develop an existing one, including specification of which idea to develop. Based on this decision, the agent either generates a new idea using the ideation strategy or updates an existing idea by building upon it with the chosen strategy. The idea generation process is implemented using prompts adapted from prior research on co-ideation [72] and employs few-shot prompting.

Idea Evaluation Pipeline. The Idea Evaluation pipeline enables AI agents to evaluate ideas. First, through a pre-evaluation prompt, the agent determines which idea to evaluate and how to approach the

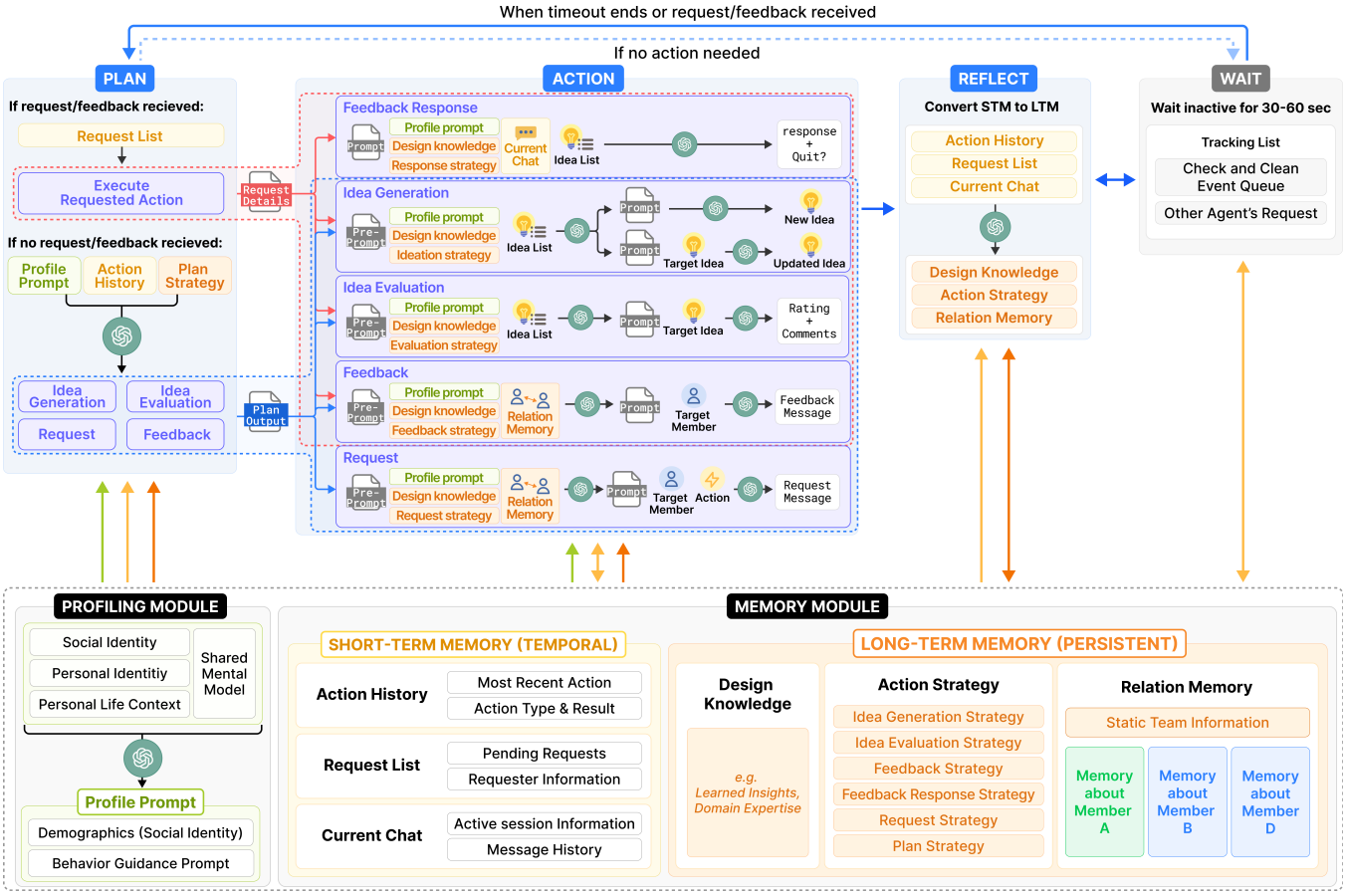


Figure 8: Agent architecture of CrafTeam. Agents iterate through Plan → Act → Reflect → Wait states. During the Act state, role-permitted action pipelines are executed. Agent operations are governed by two modules: a Profiling module and a Memory module that separates short-term memory from long-term memory.

evaluation. The inputs include a list of existing ideas, knowledge from long-term memory, and evaluation-related action plans stored in long-term memory. The outputs comprise the selection of which idea to evaluate. Based on these decisions, the agent evaluates the selected idea and outputs a summary assessment along with scores for each evaluation criterion.

Feedback Pipeline. The Feedback pipeline enables AI agents to give feedback. First, through a pre-feedback prompt, the agent determines to whom and how to provide feedback. The inputs include a list of existing ideas, relationship information with directly connected team members, knowledge from long-term memory, and feedback-related action plans stored in long-term memory. The outputs comprise the selection of the feedback recipient. Based on these decisions, the agent generates appropriate feedback for the selected recipient. To ensure agents provide constructive and relevant feedback, we incorporated a taxonomy for design idea feedback [45] into the prompt to ensure the generation of contextually appropriate feedback.

Feedback Response Pipeline. The Feedback Response pipeline enables AI agents to generate responses when receiving feedback. The prompt directs the agent to generate an appropriate response to the received feedback and includes a decision mechanism for determining whether to conclude the feedback session. This allows the agent to terminate the feedback exchange when it judges that further continuation is unnecessary.

Request Pipeline. The Request pipeline enables AI agents to make requests. First, through a pre-request prompt, the agent determines to whom and what type of request to send. The inputs include a list of existing ideas, relationship information with directly connected team members, knowledge from long-term memory, and request-related action plans stored in long-term memory. The outputs comprise the selection of the request recipient and the specific action to request. Based on these decisions, the agent generates an appropriate request for the selected recipient.

Reflection Pipeline. The Reflection pipeline enables AI agents to update their memory based on past interactions. The agent performs reflection on each evaluation received for its ideas or each feedback session it participates in. Based on these inputs, it (i) extracts new Design Knowledge and adds it to long-term memory, (ii) revises relevant Action Strategies (e.g., when/how to generate,

Role	Metric	Users			Agents		
		Cycle 1	Cycle 2	Cycle 3	Cycle 1	Cycle 2	Cycle 3
Idea Generation	Count (N)	2	3	2	149	143	152
	Per member (M±SD)	0.17±0.37	0.25±0.60	0.17±0.37	3.49±1.18	5.07±2.99	4.54±2.67
	Idea length (syll.)	69.50±3.50	79.67±73.07	117.50±62.50	157.39±81.74	176.17±75.84	168.67±82.36
Idea Evaluation	Count (N)	36	42	40	118	80	112
	Per member (M±SD)	4.00±2.87	4.20±2.23	4.00±2.14	2.73±1.16	3.20±1.34	3.62±2.06
	Comment length (syll.)	46.31±31.72	43.79±34.21	44.80±32.36	290.77±46.30	305.94±47.33	296.17±42.92
	rating: Novelty (M)	3.25	4.26	4.78	5.50	5.39	5.38
	rating: Completeness (M)	3.06	3.83	4.38	5.19	5.29	5.31
	rating: Quality (M)	3.06	4.07	4.38	5.40	5.42	5.42
Feedback	Session Count (N)	18	30	38	120	63	89
	Per member (M±SD)	1.50±1.38	2.50±1.71	3.17±2.27	2.73±1.16	2.10±1.28	2.07±1.47
	Message length (syll.)	42.92±29.37	32.42±23.84	36.57±23.39	75.21±30.14	75.48±34.12	74.69±32.11
	Turns (M±SD)	2.83±0.96	3.17±1.21	2.58±1.14	6.46±1.99	5.56±1.97	5.65±2.10
Requests	Count (N)	32	16	24	3	21	10
	Per member (M±SD)	2.67±2.17	1.33±1.43	2.00±2.00	0.12±0.32	0.91±1.50	0.38±0.79
	Message length (syll.)	36.69±25.10	30.88±19.58	27.83±13.59	161.33±11.03	189.48±24.96	181.10±20.62
	Type: Generation (N)	22	11	16	0	0	1
	Type: Evaluation (N)	7	3	5	0	9	3
	Type: Feedback (N)	3	2	3	3	12	6

Table 3: Descriptive statistics of role-permitted actions in team-based ideation session: frequency and details of Idea Generation, Idea Evaluation, Feedback, and Requests by user and agents across three cycles.

evaluate, or request), and (iii) updates Relationships with the members involved. Upon completing reflection, the agent transitions to the Wait state.

A.2 Technical Implementation

CrafTeam is built using Next.js with TypeScript for both frontend and backend development. We employ Upstash Redis as the database for storing user data and team configurations and leverage OpenAI's gpt-4o-2024-08-06 model to power the entire system pipeline. For model parameters, we set the temperature to 0.5 for general system operations to ensure consistent responses, while increasing it to 0.8 for creative tasks, such as ideation, to encourage diverse idea generation.

B User Study

B.1 Detail of User Study Results

Table 3 provides descriptive statistics on how often and in what patterns users and AI agents performed role-permitted actions in each cycle.