Towards AI as Colleagues: Multi-Agent System Improves Structured Professional Ideation

ANONYMOUS AUTHOR(S)

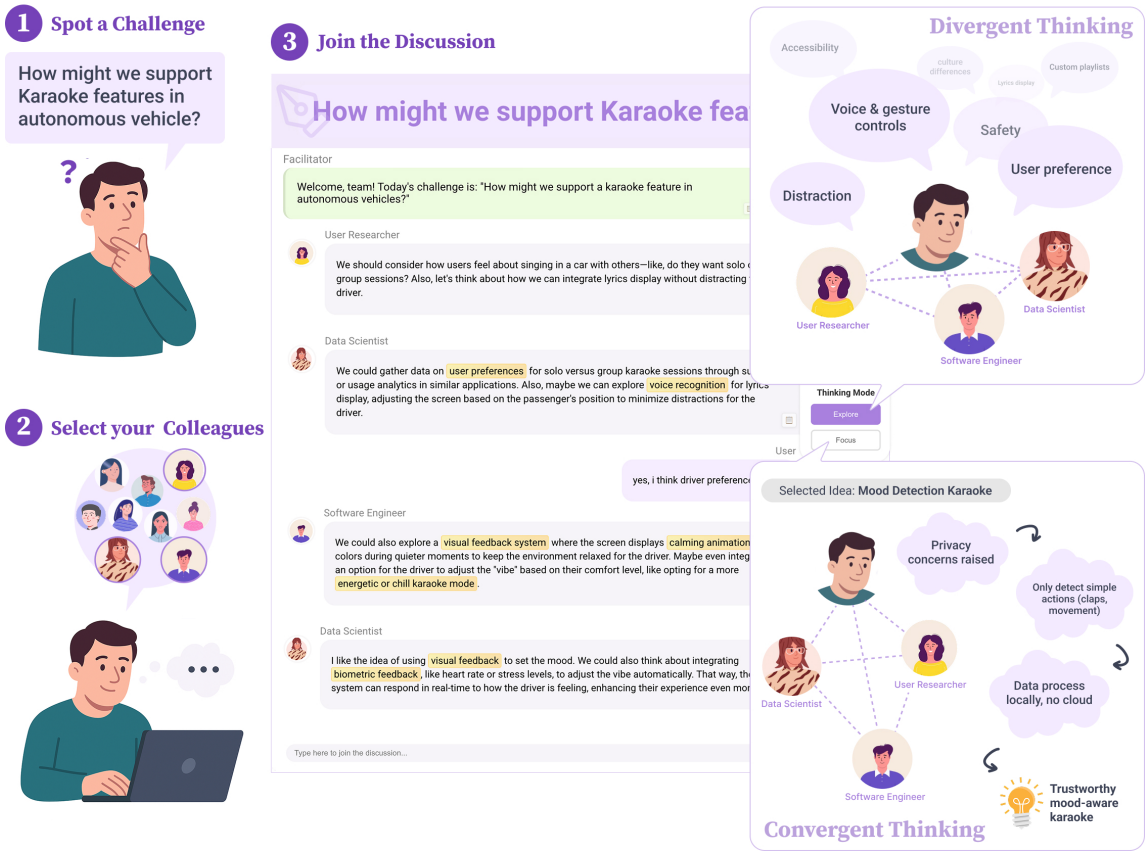

Fig. 1. **MultiColleagues: An Illustrative User Scenario of AI-Supported Ideation.** Example of P10 engages in an ideation session with three role-differentiated AI colleagues. In Step 1, the challenge is introduced. In Step 2, P10 selects colleagues (User Researcher, Data Scientist, and Software Engineer). In Step 3, the discussion unfolds: in *Explore* mode, the team generates diverse ideas; in *Focus* mode, they converge on *Trustworthy mood-aware Karaoke* that balances user experience with privacy, simplicity, and local data processing.

Most AI systems today are designed to manage tasks and execute predefined steps. This makes them effective for process coordination but limited in their ability to engage in joint problem-solving with humans or contribute new ideas. We introduce MultiColleagues, a multi-agent conversational system that shows how AI agents can act as colleagues by conversing with each other, sharing new ideas,

Manuscript submitted to ACM

and actively involving users in collaborative ideation. In a within-subjects study with 20 participants, we compared MultiColleagues to a single-agent baseline. Results show that MultiColleagues fostered stronger perceptions of social presence, produced ideas rated significantly higher in quality and novelty, and encouraged deeper elaboration. These findings demonstrate the potential of AI agents to move beyond process partners toward colleagues that share intent, strengthen group dynamics, and collaborate with humans to advance ideas.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**; **Interaction design**.

Additional Key Words and Phrases: AI as Colleagues, Multi-agent Systems, Large Language Models, Interaction Design, Collaborative Ideation

ACM Reference Format:

Anonymous Author(s). 2026. Towards AI as Colleagues: Multi-Agent System Improves Structured Professional Ideation. In *CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 37 pages. https://doi.org/10.xxxx/xxxxxxx.xxxxxxx

1 INTRODUCTION

Recent adoption of large language models (LLMs) has moved from deployments as "copilot" tools toward more dynamic roles in collaborative settings. Early applications positioned LLMs as tools or judges [14, 68], automating tasks such as summarization, programming assistance, or quality assessment that humans found repetitive or peripheral [60, 64]. Although useful, these applications positioned artificial intelligence (AI) as automation rather than as a partner in collaboration [54], whereas emerging multi-agent frameworks suggest a broader transition [10, 28]. Systems such as AutoGen [101] and CrewAI [1] demonstrate how multiple LLMs can be coordinated under structured protocols, with agents adopting complementary roles such as planner, critic, or explainer. LLMs are moving beyond single-task execution to acting as participants that contribute perspectives in a team process.

The growing use of multi-agent frameworks prompts a deeper question of whether LLMs can be experienced not only as tools but as colleagues in collaborative work. Collaboration between humans and models is often most productive when it leverages asymmetries in capability, with people contributing judgment, values, and imagination while LLMs provide scale, recall, and breadth [6, 23]. Moving beyond copilot metaphors requires examining whether users can experience these systems as team members with complementary strengths.

Ideation provides a representative context to probe this question. Research in creativity shows that human strengths remain decisive in the early stages of idea generation. At the same time, studies show that LLM assistance can expand the number and diversity of ideas, while also risking homogenization and over-reliance [15, 57, 69]. Multi-agent personas present a promising approach, as they can emulate interdisciplinary team dynamics, introduce diverse perspectives, and enable humans to retain strategic oversight. Interaction paradigms play a central role in shaping this experience. Roundtable exchanges, hierarchical supervision, or progressive disclosure influence how people engage with multiple agents and how cognitive load is managed [58, 109]. Besides, role-playing structures task allocation and heightens social presence, which makes collaboration feel closer to working with teammates [8, 48].

Recent work has examined distinct facets of this design space, with systems emphasizing either divergent–convergent structuring through staged debate and role play [57] or orchestration through contrasting roles and rotating perspectives to guide reflection [70, 110]. These approaches demonstrate the value of both dimensions but have largely been explored in isolation. Building on these directions, we introduce MultiColleagues, a multi-agent conversational system that brings together diverse AI personas for co-ideation while centering the human as facilitator-in-chief. The system

pursues three design goals: (1) supporting shifts between divergent and convergent thinking, (2) engaging diverse viewpoints to expand the idea space, and (3) providing clear features that preserve human oversight in collaboration.

Guided by these goals, we evaluate MultiColleagues through a within-subjects study against a single-agent ChatGPT baseline, focusing on three research questions:

(1) **RQ1 - Experience**: How do role-taking patterns and perceptions of social presence shape the collaborative atmosphere and user engagement?
(2) **RQ2 - Outcomes**: How does exposure to multiple AI-Colleague perspectives affect the creative outcomes, and what underlying dynamics may explain these differences?
(3) **RQ3 - System Design**: How do system design features support or constrain support during creative ideation?

2 RELATED WORK

2.1 From Tools to Colleagues: The Evolution of Human–AI Collaboration

Over recent years, large language models (LLMs) have emerged as transformative tools across a wide range of applications, demonstrating state-of-the-art performance in natural language processing and knowledge-intensive tasks [47, 105]. They have been widely adopted in domains such as software development [22, 102], writing and creativity support [19, 100], scientific research and literature review [32, 62], and scientific experimentation [12]. Across these domains, LLMs have primarily functioned as assistants or evaluators, supporting human work by improving efficiency and facilitating decision-making [76, 108]. Nevertheless, human-in-the-loop (HITL) involvement remains indispensable. Research has shown that relying on humans solely as "reviewers" can introduce risks, such as decision-making risks [33, 82], reliability and transparency risks [7], systemic risks [79], and ethical risks [3, 46, 74]. Additional HITL studies across domains reinforce the irreplaceable role of humans and confirm that human involvement enhances accuracy and reliability [52, 92]. Building on this foundation, human-centered and mixed-initiative frameworks argue that the future of AI lies in augmentation rather than replacement, coupling higher levels of automation with sustained human control [6, 87]. As LLMs grow more capable, systems are gradually shifting toward proactive collaboration. Multi-agent frameworks such as AutoGen and CAMEL operationalize LLMs as differentiated collaborators [51, 101], where models can assume specific roles, engage in mutual critique, and coordinate planning activities. Through such structured interactions, they approximate team-like collaboration and move toward systems where LLMs function less as tools and more as partners [27, 56].

2.2 Multi-Agent LLMs and Teamwork Dynamics

The growing capabilities of recent years' LLMs have motivated increasing interest in multi-agent frameworks as a way to extend the scope and complexity of applications. A first major direction examines *task-decomposed collaboration*, where models are assigned complementary roles such as planning [101], execution [97], or debugging [25]. Within this strand, researchers have proposed different coordination paradigms, including dialogic debate [27, 53] and hierarchical supervision [35, 102]. While these systems demonstrate gains in reasoning, factuality, and coding ability, their evaluations remain largely confined to benchmark settings [20, 38] rather than interactive use.

A second line of work explores *persona-driven role play*, showing that LLMs can convincingly simulate interdisciplinary teamwork by adopting distinct identities. This stream highlights how personas shape more human-like interaction styles [50, 71, 86, 99], enhance engagement [21, 98], and diversify task performance through structured impersonation [51]. Examples range from predefined expert roles in CAMEL [51], which bring complementary perspectives

System	System			User Study			Remarks
	Turn	Div–Conv	Orch.	In-situ	Pick	Ctrl.	
LLM Discussion (COLM'24) [57]		✓		—	—	—	Agents as process partners in agent–agent debate with convergence to boost model creativity on benchmarks.
SWTW (CHI'24) [110]			✓	✓			Agents as a guidance panel for progressive exposure in media reading.
Weaver (CHI EA'25) [70]			✓	✓	✓		Advisory round-table with next-speaker and summaries to surface impacts.
MultiColleagues (our work)	✓	✓	✓	✓	✓	✓	**Agents as colleagues for co-ideation with Explore/Focus and human-paced facilitation.**

Table 1. Comparison of multi-agent systems. Icons: ✓ yes, — not applicable. **Turn** = dynamic turn selection; **Div–Conv** = divergent–convergent phases; **Orch.** = human-facing orchestration; **In-situ** = interactive study; **Pick** = user chooses agents; **Ctrl.** = controlled (within-subjects) vs. baseline.

to task execution, to generative agents that exhibit emergent social behaviors in daily scenarios [71, 85]. Collectively, this work points to the potential of role differentiation for strengthening social presence and aligning collaboration with human expectations.

A third line of research focuses on *coordination mechanisms* that sustain coherence across longer or more complex interactions. Efforts here include shared memory, scheduling, and blended model outputs. Skeleton-guided reasoning [66], model fusion [44], and collaborative decoding strategies [91] exemplify how multiple agents can combine strengths to tackle tasks beyond the capacity of a single model. At the same time, across all three directions, prior work consistently emphasizes the importance of human oversight for aligning decisions with user intent and preventing failures under high autonomy [5, 11, 18, 81].

As shown in Table 1, recent work has begun to explore multi-agent systems in interactive settings. LLM Discussion [57] adopts a three-phase agent–agent debate with role-play to enhance originality and elaboration of model outputs. SWTW [110] introduces progressively contrasting roles in media reading, using orchestration and gamified puzzles to mitigate filter bubbles and deepen reflection. Weaver [70] organizes advisory-style roundtables with speaker rotation and summaries to anticipate broader social impacts. While each contributes valuable orchestration strategies, they do not combine dynamic turn-taking, divergent–convergent phases, and user-facing facilitation within controlled, in-situ studies. Motivated by these gaps, MultiColleagues advances co-ideation by combining dynamic turn-taking, explicit divergent–convergent shifts, and interactive orchestration, enabling a controlled study of how role differentiation and multi-agent dynamics shape collaborative ideation.

2.3 GenAI-Assisted Ideation

Research on AI-assisted ideation has evolved from retrieval-based systems to generative, structured support. Early tools such as IdeaHound [88], ProbMap [59], and IdeateRelate [103] used semantic similarity to surface relevant ideas but offered limited scaffolding beyond recall. Recent systems embed generative models into interactive environments: Jamplate organizes reflection through templates [104], BioSpark clusters LLM concepts into analogical "inspiration cards" [45], Scideator recombines scientific paper facets to propose novel research ideas [77], and CausalMapper visualizes causal relations among concepts to guide systematic exploration [42]. These systems demonstrate how generative outputs can be transformed into higher-level structures that promote reflection, analogy, and cross-domain thinking. Complementary approaches emphasize role-based scaffolding. PersonaFlow [55] simulated multiple expert personas to

inject disciplinary perspectives, while Rayan et al. [78] leveraged generative chat to stimulate collaboration. Yet purely textual interfaces risk fixation: listing example solutions can anchor users to narrow trajectories [16]. Spatial or hierarchical arrangements mitigate this risk by situating ideas in broader solution spaces, as demonstrated by tree-structured diagrams for structured exploration [26, 107].

More recently, attention has shifted to the limitations of single-agent LLMs. Their autoregressive nature tends toward convergence, limiting diversity [95]. Multi-agent approaches distribute perspectives across distinct personas to sustain divergence and reframe models as collaborators rather than tools [31, 33]. Building on this trajectory, our work positions LLMs as colleagues in a multi-agent environment, embedding the Double Diamond design model [24] to balance divergence and convergence, user control and agent autonomy, toward more coherent yet diverse ideation.

3 SYSTEM DESIGN

3.1 Design Goals and Implementation

Our reflections on prior work in conversational agents for human–AI collaboration led us to articulate the following design goals for our MultiColleagues system:

(1) DG1: Support adaptive Human-AI co-ideation dynamics. The system should enable users to fluidly navigate between exploratory and evaluative phases of collaborative ideation while maintaining strategic control over the creative process. This design goal was primarily motivated by prior research on the double diamond design process and human creative cognition patterns, which conceptualize innovation as an iterative progression through phases of divergent exploration and convergent refinement [24].

(2) DG2: Enable rich, multi-perspective co-ideation. The system should facilitate engagement with diverse viewpoints and expertise domains to avoid narrow ideation patterns and encourage comprehensive exploration of solution spaces. This design goal was motivated by prior research on interdisciplinary collaboration and the limitations of single-agent interactions.

(3) DG3: Facilitate purposeful and transparent Human-AI collaborative control. To promote effective human–AI synergy, the interface should provide clear interaction points and user-friendly visual interfaces that enable users to maintain strategic oversight while leveraging AI capabilities for ideation support. This design goal emerged primarily from prior research on human-AI teaming, highlighting the need for balanced human agency and intuitive interface design in AI-mediated creative processes.

To address our design goals, we designed and developed **MultiColleagues**, a human–AI collaborative conversational platform where multiple AI personas participate in structured brainstorming alongside the user. The system architecture follows a two-tier design pattern: presentation layer (React frontend) and application layer (Flask API). The system integrates OpenAI GPT-4o for natural language generation (see Figure 2).

3.1.1 Adaptive Thinking Transition Mechanisms. To support adaptive thinking transitions in collaborative ideation (DG1), we implemented a *dual-mode switching framework* grounded in the double diamond design methodology, explore mode and focus mode, as illustrated within the usage flow (Fig. 3). The Explore mode emphasizes breadth and diversity of viewpoints. During this mode, different colleague experts are encouraged to expand the space of possibilities, each approaching the problem from their own perspectives and generating ideas freely without immediate concern for constraints. Focus mode emphasizes depth, clarity, and actionable outcomes. In this stage, colleagues shift toward evaluating, filtering, and aligning ideas, working from their own roles to concentrate on feasibility and actionable outcomes.

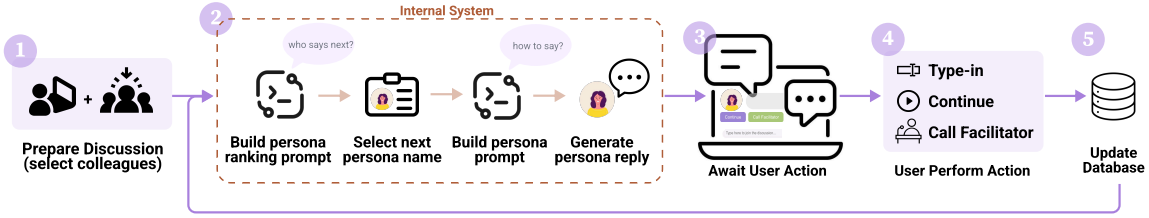

Fig. 2. **System Workflow for Persona-Guided Discussions.** This diagram illustrates the end-to-end workflow of the persona-guided discussion system across five steps. The process begins with preparing the discussion by selecting AI colleagues (Step 1). The internal system first builds a persona ranking prompt, next selects the next persona, then builds a persona prompt and generates the corresponding reply (Step 2). The generated output is presented to the user, and the system awaits user action (Step 3). The user performs one of three actions: type a response, continue with the system-generated reply, or call a facilitator for support (Step 4). All logs are stored in the database (Step 5).

Our system provides effective ideation through systematic alternation between divergent exploration phases (expanding problem and solution spaces) and convergent synthesis phases (evaluating, refining, and consolidating ideas). This template-based approach ensures that thinking mode transitions extend beyond interface changes to actually modify AI thinking processes in alignment with human creative phases.

Strategic interaction points are integrated after each AI response to preserve human agency and prevent cognitive saturation, as we found that uninterrupted AI generation creates information overload and diminishes human creative contribution and strategic oversight capacity [2]. These pause points enable participants to make explicit choices to "Continue" autonomous AI discussion or "Call Facilitator" for guided intervention, which preserves human cognitive control by preventing information saturation and enabling reflective engagement with AI-generated content before proceeding to subsequent ideation phases (DG3).

An AI Facilitator functions as a meta-cognitive regulator that monitors conversation dynamics and intervenes at calculated intervals when discussions deviate from productive ideation patterns, lack adequate synthesis, or fail to incorporate diverse participant perspectives [30]. The facilitator employs real-time conversation analysis to identify critical transition points between diamond phases. It explicitly prompts users with a quick overview of the current conversation history, and to reflect on whether the team should continue divergent exploration or transition toward convergent evaluation, and provides structured progress synthesis to prevent cognitive fragmentation across extended ideation sessions.

3.1.2 Multi-Persona Orchestration System. To enable rich, multi-perspective co-ideation (DG2), the system implements a persona orchestration framework that instantiates a diverse roster of AI colleagues with distinct professional backgrounds, communication styles, and domain expertise (see Figure 8 in Appendix D.1). Each persona is constructed through structured configurations defining behavioral instructions for communication patterns, specialized knowledge domains that establish topical authority, and participation patterns that govern engagement frequency. These persona templates are drawn from prior works on multi-agents' interdisciplinary collaboration [40], which show that encoding standardized workflows into multi-agent prompts improves coordination and reduces cascading errors. We designed the personas to interact through a multi-stage selection process. First, all selected personas generate preliminary thoughts on the participant's problem using individualized prompts. Then, an AI-driven algorithm evaluates these responses for relevance and engagement potential, selecting the first speaker to establish a natural conversation flow. For subsequent turns, a dynamic ranking mechanism selects the next speaking persona by scoring candidates on

contextual relevance to prior user comments, conversation history, and unexpressed perspectives, while adding a 20% randomization factor to prevent deterministic patterns. The ranking algorithm instructs the language model to identify which persona would have "the strongest urge or most relevant comment to share next" given current discussion dynamics, ensuring diverse perspectives emerge organically rather than through artificial rotation. Once a persona is selected, the system retrieves its prompt template and combines it with the full conversation context, the user's most recent input comments, and the current orchestration state (focus vs. explore). Guided by this structured input, the language model generates the persona's output. To sustain longer dialogues, a conversational history compression pipeline is applied when the message count exceeds a threshold. As shown in Figure 9, recent turns' history is kept in full while older persona contributions are summarized, allowing the system to maintain immediate context while compactly representing earlier perspectives. This compression design ensures efficiency and coherence in multi-colleague conversations (see Appendix D.2). Detailed prompt templates for persona creation, first-speaker selection, persona ranking, and response generation are provided in Appendix C.

3.1.3 User-Friendly Interface and Interaction Design. To maintain strategic oversight in multi-agent ideation while preventing passive consumption of AI output, we implement user-friendly interfaces and interaction mechanisms that enable clear control and integration points throughout collaborative discussions. Beyond the conversation flow controls described in DG1, the system provides clear visual cues through distinctive persona profile pictures and role-based message styling that enable users to quickly identify different AI perspectives and track individual contributions across extended discussions. An intelligent highlighting system with user-controlled visibility allows participants to manage information density by toggling keyword emphasis on or off, which supports cognitive load management and preserves access to AI-generated insights. The interface enforces each persona's conversation wording limits and structured message threading that enhances readability during multi-agent exchanges. Besides, the thinking mode controls enable users to explicitly switch between "explore" mode (divergent thinking) and "focus" mode (convergent thinking), with clear visual indicators showing the current cognitive state and immediate effects on subsequent AI behavior.

3.2 Usage Scenario

To illustrate how participants engage with the multi-agent co-ideation system, we present a typical user journey (see Figure 3). The participant begins by selecting three AI personas (User Researcher, System Architect, and Market Analyst) and submitting the problem statement *"How might we design an AI system to help remote teams collaborate more effectively?"* The system generates initial thoughts from each persona and presents the User Researcher as the first speaker, who raises questions about user pain points in remote collaboration. During the initial exploration phase, the participant primarily operates in "explore" mode, alternating between clicking "Continue" to observe autonomous AI discussions and actively joining the conversation by submitting their own insights. After approximately six AI responses covering topics ranging from technical infrastructure concerns to user experience considerations, the participant clicks "Call Facilitator" to request guidance on discussion direction. The AI facilitator provides a synthesis of perspectives shared so far and prompts the participant to consider whether the team should continue exploring the problem space or begin focusing on specific solution approaches. At this point, the participant switches to "focus" mode and submits a message, narrowing the scope to "real-time collaboration tools for creative teams." The AI personas now operating in convergent thinking mode, begin evaluating and synthesizing the discussed ideas, with the System Architect proposing specific technical architectures while the User Researcher focuses on user-related principles. The

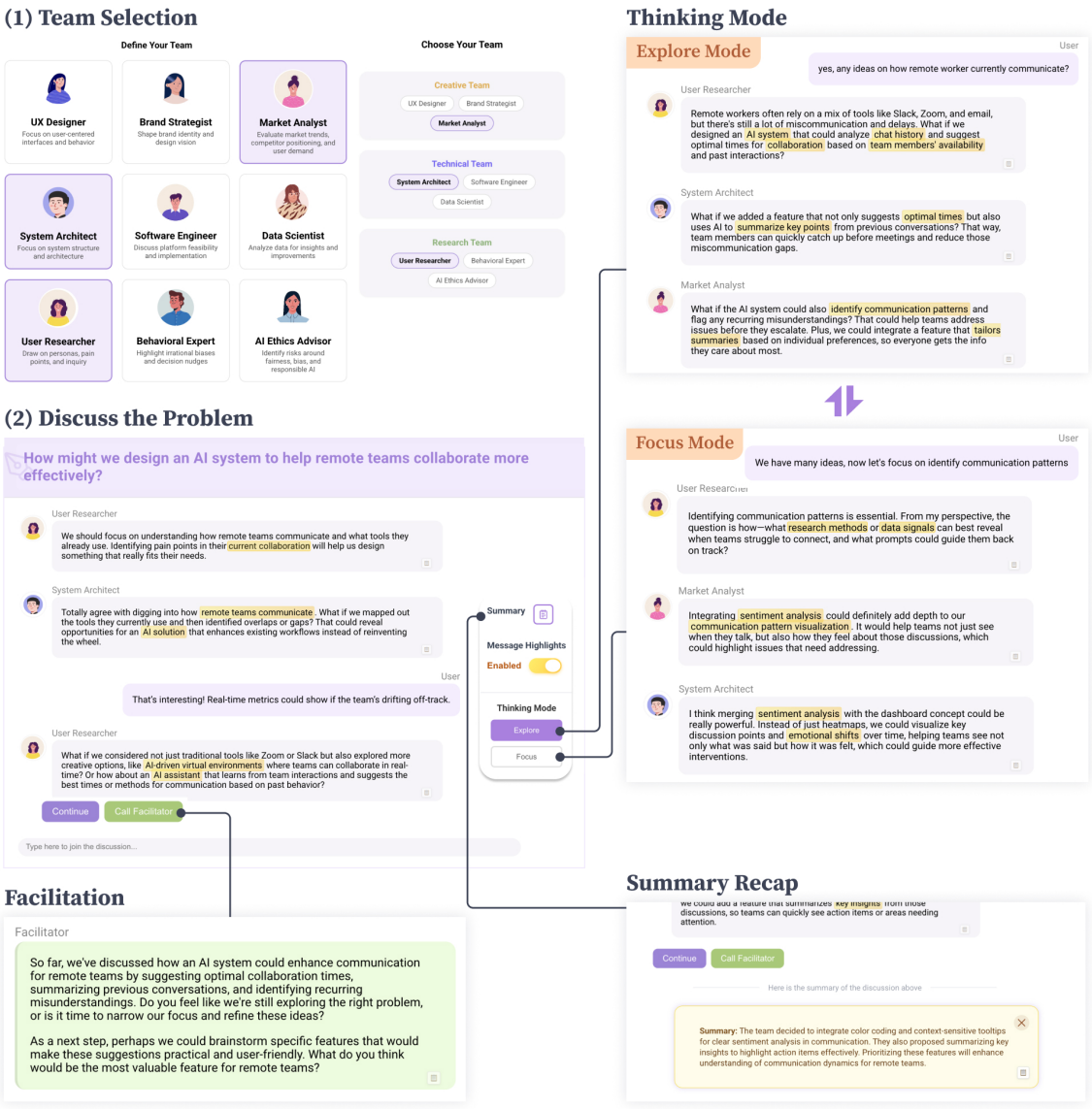

Fig. 3. **Usage flow of the Multi-agent Conversational System for co-ideation.** (1) Users first *select a team* of persona experts. (2) They then *discuss the problem*, with personas contributing from their own perspectives. The system supports two *thinking modes*: **Explore mode** expands ideas broadly, while **Focus mode** refines and synthesizes them. A *summary recap* highlights key points, and *facilitation* is triggered either by user request or automatically by the system when guidance is needed.

participant carefully reviews the intelligent highlighting phrases to identify key concepts to consolidate the emerging solution framework.

ID	Age	Occupation	Specialization	Creativity
P1	25-29	Master Student	Communication	2.09
P2	25-29	PhD Student	Material Science & Engineering	4.91
P3	25-29	Professional	Computer Science	4.91
P4	20-24	Master Student	AI Research	5.18
P5	25-29	Professional	Information Science	7.00
P6	25-29	PhD Student	Information Science	6.27
P7	20-24	Master Student	NLP	4.73
P8	25-29	Professional	UX Research	4.64
P9	25-29	PhD Student	Computer Science	6.55
P10	25-29	Professional	Design	5.27
P11	20-24	PhD Student	Computer Science	4.64
P12	30-34	PhD Student	Voice Interaction	6.27
P13	20-24	Undergraduate Student	HCI	5.36
P14	25-29	PhD Student	Virtual Reality	6.09
P15	35-39	PhD Student	AI Research	5.27
P16	30-34	PhD Student	HCI	6.45
P17	20-24	Master Student	Virtual Reality	6.00
P18	25-29	Professional	Design	6.00
P19	25-29	PhD Student	Machine Learning	5.18
P20	30-34	PhD Student	Information Science	6.55

Table 2. Participant demographics and creativity scores.

4 STUDY DESIGN

To examine how different AI interaction paradigms influence creative collaboration, we designed our study structured as follows: participant recruitment and characteristics (Section 4.1), detailed procedure including task design (Section 4.2), and data collection and analysis methodology (Section 4.3).

4.1 Participants

As shown in Table 2, our study recruited 20 participants through university mailing lists and professional networks (9 males, 11 females), aged 20-39 years (M = 26.7, SD = 3.8). 15 participants were undergraduate to PhD students pursuing degrees in fields such as computer science, information science, HCI, AI research, and communication, while the remaining 5 were early-career professionals working in technology, research, and design-related roles. All participants met our eligibility criteria of having backgrounds in relevant backgrounds and prior experience with creative problem-solving or AI-driven tools. This study received the university's IRB approval. Each participant was compensated with $20. Table 2 summarizes participants' demographics along with their creativity scores (M = 5.39, SD = 1.16). Scores were calculated as the average of 11 items from a 7-point Likert-scale creativity assessment (see Appendix A.1 for the full questionnaire).

4.2 Study Procedure

Each study session lasted approximately 70 minutes and was conducted remotely via Zoom with a researcher present to observe interactions and provide technical support when necessary. All participants experienced both systems, MultiColleagues and ChatGPT (GPT-4o), in counterbalanced order to control for potential order effects [75]. The detailed procedure comprised the following phases (see Figure 4 for visualized study workflow):

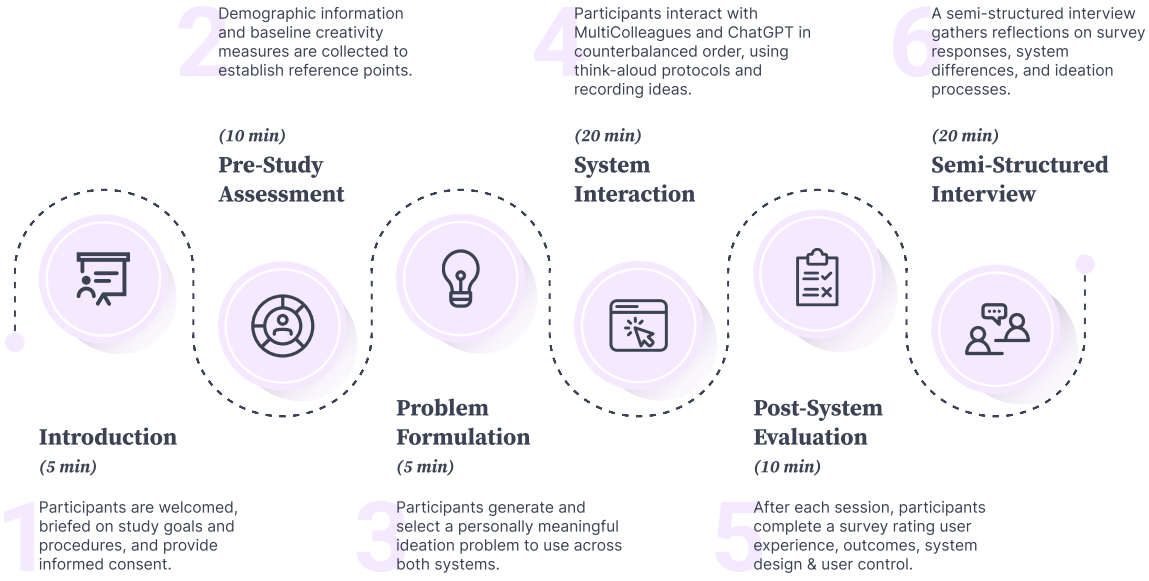

Fig. 4. User Study Workflow.

Introduction and informed consent (5 minutes): Participants were briefed on study objectives, procedures, and data handling protocols before providing written informed consent.

Pre-study assessment (10 minutes): Participants completed demographic questionnaires capturing age, gender, educational background, and prior experience with AI-assisted creativity tools. Additionally, participants completed established creativity assessment instruments adapted from validated scales [80, 93] to establish baseline creative capabilities (complete items provided in Appendix A.2).

Problem formulation (5 minutes): Before experiencing systems, participants received a brief prompt on what constitutes a suitable ideation problem and chose a topic they were familiar with or personally interested in. Previous works revealed that predefined problem themes reduced participants' engagement [63], therefore in the main study, participants generated their own topics of interest. The same problem was used with both systems, with a counterbalanced order to minimize order effects [75].

System interaction phase (20 minutes total; 10 minutes per system): Participants were randomly assigned to one of two condition orders, with equal distribution across sequences to control order effects. Think-aloud protocols [94] were employed throughout both interactions to capture real-time cognitive processes. For both conditions, a split-screen setup was provided with a Google Doc on the right side for recording generated ideas following each interaction. The allocated 10-minute timeframe could be extended if deemed insufficient, and early termination was permitted upon idea exhaustion or fatigue.

- **MultiColleagues condition**: Participants received a structured tutorial (approximately 2 minutes) covering interface navigation, persona selection mechanisms, thinking mode transitions, and facilitator function utilization. Following orientation, participants engaged in a 10-minute guided brainstorming session employing the multi-agent system architecture with their self-selected problem.
- **Baseline condition**: Participants unfamiliar with ChatGPT received a brief interface orientation focusing on conversation initiation and prompt formulation strategies. To ensure model consistency, all participants accessed ChatGPT through the web interface using the GPT-4o model. Participants then conducted unrestricted text-based ideation sessions using identical problem parameters.

Post-system evaluation (10 minutes total; 5 minutes per system): Following each system interaction, participants completed a 12-item survey on 7-point Likert scales (1 = strongly disagree, 7 = strongly agree). The survey evaluated three key dimensions: *Experience* (3 items), *Outcomes* (3 items), and *System Design & User Control* (4 items). For example, *Outcomes* evaluates if "I reached lots of valuable or actionable ideas that felt better than what I might have generated alone." The complete set of 12 items is provided in Appendix A.

Semi-structured comparative interview (20 minutes): The concluding interview phase employed a structured protocol designed to elicit detailed comparative reflections on participants' experiences across both systems. Participants first reviewed and elaborated on their quantitative survey responses, providing contextual explanations for their ratings. Subsequently, three targeted follow-up questions were explored: (1) the extent to which different personas provided distinctive perspectives and fulfilled unique collaborative roles, (2) how participants' perceived relationship and interaction patterns with AI agents differed between single-agent and multi-agent configurations, and (3) whether exposure to multiple personas influenced participants' ideation structuring processes or evaluative criteria for generated concepts.

4.3 Evaluation Methods

To answer our research questions, we adopted a mixed-methods approach that integrates both behavioral and perceptual measures. Quantitative data were collected from conversation histories and system interaction logs, complemented by structured post-system evaluations of usability and creativity. Qualitative data came from semi-structured interviews, capturing participants' experiences with collaboration strategies, the outcomes of navigating different interaction paradigms, and the utility of multi-agent versus single-agent approaches. The following sections detail our evaluation methods.

4.3.1 Comparative Analysis of User Ratings. To statistically evaluate the 12 survey data, we employed a structured analytical approach that began with thematically grouping the questions according to our research questions, followed by a non-parametric comparison of the two systems. The 12 questions were first organized into three core dimensions: (1) **Experience (RQ1)**, which captured the subjective quality of the collaboration and user engagement, as well as factors shaping those outcomes; (2) **Outcomes (RQ2)**, which measured the quality and novelty of the creative output; and (3) **System Design & Control (RQ3)**, which evaluated the user's perceived sense of control and flexibility. To streamline the analysis, closely related survey questions were merged into broader, more robust metrics, such as *Outcome Quality & Novelty*. Given the ordinal nature of the Likert scale data and the within-subjects design of the study, we used the non-parametric Wilcoxon signed-rank test to compare the paired ratings of MultiColleagues and Baseline for each metric. Effect sizes were calculated as the magnitude of observed differences between conditions.

4.3.2 Thematic Analysis. Two researchers conducted thematic analysis [13] of interview transcripts, independently coding the data before reconciling differences and refining the codebook. Themes were organized around the study's research questions on collaborative experience, creative outcomes, and system support, with sub-themes (e.g., "Traceable Perspectives") emerging through iterative discussion and grouping into role- or condition-specific insights.

4.3.3 Topics Analysis and Conversational Structures. To complement engagement-level metrics, we examined the *semantic structure* of conversations by analyzing the number and distribution of discussion topics. We used GPT-5 with role-based prompts to segment transcripts into *main topics* (high-level themes) and *sub-topics* (supporting details). We chose GPT-5 as evaluator for its state-of-the-art reasoning and judgment capabilities across domains to ensure reliable idea assessment [96]. To mitigate stochastic variability, each conversation was processed three times and the results averaged. All outputs were subsequently reviewed by a researcher for alignment with coding criteria. Based on these annotations, we calculated a *branching ratio*, defined as the mean number of sub-topics associated with each main topic. This metric characterizes whether conversations progressed more linearly (lower branching) or expanded into multi-threaded explorations (higher branching). To ensure comparability across sessions of different lengths, topic counts were normalized by conversation duration, producing *topics per minute* and *sub-topics per minute* as indicators of topical density. In addition, we calculated the average *time per main topic* and *time per sub-topic* to reflect the extent of conversational investment in individual ideas.

5 RESULTS

We organize the results into two parts. First, we report log analyses that establish overall interaction statistics in the MultiColleagues condition (Section 5.1). Our descriptive results capture how participants engaged with the system in terms of colleague selection patterns and message distributions, providing context for interpreting subsequent findings. The remainder of the results sections (Section 5.1, 5.2, 5.3) are structured around our three research questions, combining survey responses, conversation logs, and interviews. For RQ1 (Experience), we examine how role-taking patterns and perceptions of social presence shaped participants' collaborative atmosphere and engagement. For RQ2 (Outcomes), we assess how exposure to multiple AI-Colleague perspectives influenced the quality, novelty, and organization of creative outputs. For RQ3 (System Support), we evaluate how system design features affected participants' sense of agency, control, and flexibility during ideation.

5.1 Overview of Participants' Interaction Statistics

To contextualize participants' engagement with MultiColleagues, we first examined colleague selection patterns, followed by interaction statistics across sessions. The analysis revealed substantial variance in AI colleague choices, with 95% unique persona combinations across 20 participants. This diversity reflects strong individual differences in how participants valued expertise for ideation, suggesting that effective collaboration benefits from accommodating varied teaming preferences.

Turning to conversational dynamics, multi-chat sessions included an average of M = 4.00 personas (SD = 1.00), ranging from three to seven colleagues. On average, each session contained 31.3 utterances, while participants contributed M = 8.3 utterances (26.7%). To avoid overweighting the AI presence by summing across multiple colleagues, we examined the average contribution per AI colleague. Individual colleagues produced M = 5.04 utterances (SD = 1.63). This indicates that while AI colleagues collectively sustained much of the conversational flow, each colleague's contribution was comparable in magnitude to that of the human participant, suggesting a more balanced distribution of interaction.

5.2 RQ1: Collaborative Experience with Multi-AI Colleagues

To address RQ1, we operationalized collaborative experience into three dimensions: complementary strengths, teammate-like relationships, and engagement. We observed significant differences across all, with MultiColleagues rated more positively than Baseline. We present results for each dimension below, drawing on both post-system evaluation responses (Table 3) and interviews.

Metric	Q#	MC (M ± SD)	Baseline (M ± SD)	W	p-value	Effect Size (r)
Teammate-like Feel	Q6	5.75 ± 1.02	5.05 ± 1.15	17.5	**.046***	0.49
Complementary Strengths	Q8	6.05 ± 1.00	5.05 ± 1.64	2.5	**<.01****	0.71
Engagement & Flow	Q10	5.70 ± 1.38	4.45 ± 1.57	30.0	**.014***	0.63

Table 3. Statistical Comparison for Process & Experience Metrics between **MultiColleagues (MC)** and **Baseline**. Results show significantly higher ratings for MC across all 3 metrics.

5.2.1 Distributed Collaboration Enhances Complementary Strengths. Survey results and interview reflections consistently emphasized that MultiColleagues enabled a stronger sense of distributed collaboration, where different AI colleagues contributed complementary strengths to the discussion from distinct angles. During the interview, we first invited participants to reflect on their collaborations whether they felt colleagues served distinctive roles and offered perspectives from different angles on a 7-point Likert scale. Findings confirm that most participants perceived colleagues as differentiated contributors (M = 5.85, SD = 1.09). Moreover, survey comparisons showed that participants rated the MultiColleagues condition (M = 6.05, SD = 1.00) significantly higher than the Baseline condition (M = 5.05, SD = 1.64; W = 2.5, p <.01) in terms of contributing different views and strengths to the ideation process. These quantitative findings indicate that participants more often framed the multi-agent system as a team of collaborators rather than a single tool. Echoing the survey results, rather than relying on a single expert voice, participants described the system as a collection of distributed roles that encouraged them to participate more actively in the interview. As P3 reflected, "it actually encourages me to join in the thinking process… like one of the people in the conversation," while P19 valued how "[AI colleagues'] contributions complement each other." This role separation often gave participants the impression of working with a team of specialists, with P10 noting that the role-playing "makes you feel more that everyone is contributing their strength."

At the same time, participants recognized trade-offs in this distributed setup. Some noted that AI colleagues tended to remain bounded within their professional domains, with P2 observing that "they're limited to their professional aspects," and P18 describing the system as expanding by "giving you another island" rather than filling out an existing territory. This metaphor reflected how colleagues generated separate contributions without consolidating expertise and broader topical coverage. By contrast, participants characterized the Baseline condition's responses as "more academic and useful with wider scope" (P2) and emphasized its ability to fill in missing details to make ideas more complete (P18). In this sense, Baseline offered more integrated coverage within a single agent, while MultiColleagues emerged from the aggregate of multiple narrower perspectives, offering breadth across roles.

5.2.2 Colleague Roles Enable Facilitative Leadership. Findings from the interview showed that MultiColleagues generated facilitative leadership through distributed expertise. Since understanding multiple AI colleagues' voices required oversight and integration, participants often stepped into coordinator or facilitator roles. P3 explained that it "encourages me to join in the thinking process… like one of the people in the conversation," while P5 reflected, "I really feel

like I'm in front of this team, and they have to deliver their ideas to me, so I might feel that my sense of power is a little higher." Importantly, this authority felt collaborative rather than authoritarian, with P16 noting, "I felt somewhere between leader, facilitator, but I definitely had the feeling of a collaborator." The distribution of roles across AI colleagues demanded that users manage the collaborative process, as P17 described: "feel like chatting with my team in Slack, I can freely let whoever I want to speak out. I have a role feeling — like a PM [project manager]." The Baseline condition, in contrast, consolidated expertise into a single authoritative voice, reinforcing structured authority relations. Some participants described it as "a very senior-level expert who can immediately give you a very complete, very detailed plan" (P3), while others positioned themselves as supervisors with an assistant: "I feel like a boss. I'm just asking my assistant to fetch something for me" (P19). Yet even in supervisory roles, participants still noted how the Baseline condition shaped their thinking, with P12 reflecting, "I feel like I'm the boss… but [Baseline] reshapes how I think, so it has more power." These accounts illustrate how MultiColleagues' distributed roles fostered facilitative leadership, while Baseline's consolidated expertise reinforced stable hierarchical structures.

Metric	MC (M ± SD)	Baseline (M ± SD)	W	p-value	Effect Size (r)
Linguistic Cohesion Metrics					
Narrativity	24.40 ± 14.50	18.62 ± 15.62	59.0	.090	1.58
Syntactic Simplicity	28.62 ± 15.72	28.83 ± 19.62	92.0	.648	2.46
Word Concreteness	15.13 ± 17.71	19.13 ± 19.24	82.0	.409	2.19
Referential Cohesion	27.25 ± 18.55	25.42 ± 20.63	96.0	.756	2.57
Pragmatic / Interaction Style Metrics					
Sentiment	4.44 ± 0.41	4.46 ± 0.33	104.0	.985	2.78
Formality	4.00 ± 0.42	4.08 ± 0.63	80.0	.368	2.14
Directness	4.76 ± 0.23	5.05 ± 0.44	30.0	**.009****	0.80
Relationship	4.39 ± 0.34	4.17 ± 0.44	88.0	.545	2.35
Participation	4.29 ± 0.67	4.09 ± 1.07	87.0	.521	2.33

Table 4. Statistical Comparison of Linguistic Cohesion and Pragmatic Style Metrics between **MultiColleagues (MC)** and **Baseline**. Results show a significant difference in *Directness* of conversational style.

5.2.3 Team-Like Atmospheres Shape Collaborative Experience. Survey results indicated that the distribution of roles in MultiColleagues fostered a stronger sense of team-like interaction, with participants rating MultiColleagues (M = 5.75, SD = 1.02) significantly higher than the Baseline condition (M = 5.05, SD = 1.15; W = 17.5, p = .046). These results suggest that the system's role differentiation encouraged participants to experience the interaction as more socially collaborative. Interview reflections further highlighted how role differentiation created a team-like dynamic. P5 noting that "everyone has their own role… it's a very social state" and P10 adding that role-playing "makes you feel more that everyone is contributing their strength." Others described a heightened sense of immersion, as P19 reflected that the team atmosphere "naturally facilitate[d] or control[led] the direction of the discussion." For a few, this immersion was sufficiently strong to diminish the perceived boundary between human-AI contributions, with P9 recalling, "I felt like I forgot they were AI [colleagues]." By contrast, interview accounts of the Baseline condition consistently depicted it as neutral and instrumental. Participants described it as a source to "seek an answer and take the answer away" (P5). P9 echoed this perspective, "When I use it, I am a human being and [Baseline] is just a tool. I don't feel it is my teammate."

Overall, while the Baseline condition was regarded as efficient and authoritative, MultiColleagues' role differentiation cultivated stronger social immersion, reinforced the sense of presence, and encouraged more collaborative engagement.

To examine whether this heightened sense of team-like atmosphere was also reflected in participants' own language, we analyzed linguistic cohesion metrics from Coh-Metrix [34, 61] alongside pragmatic style ratings across conditions (see detailed methods in Appendix B.1). Results presented in Table 4 showed a significant difference in Directness, with the Baseline condition (M = 5.82, SD = 0.91) rated significantly higher than the MultiColleagues condition (M = 5.21, SD = 0.88; W = 30.0, p = .009), suggesting that interactions with Baseline elicited more straightforward, task-focused language. Other contrasts did not reach statistical significance.

Metric	MC (M ± SD)	Baseline (M ± SD)	W	p-value	Effect Size (r)
# of Utterances	8.35 ± 5.79	4.10 ± 2.45	9.5	**.001****	0.71
Total User Words	104.70 ± 55.85	51.30 ± 42.47	18.0	**<.001*****	0.73
# of Utterances per Minute	0.65 ± 0.38	0.42 ± 0.23	125.0	**.044***	0.32
Total User Words per Minute	8.12 ± 4.25	5.12 ± 3.67	119.0	**.029***	0.35
Average Word Count per Message	13.47 ± 4.65	12.83 ± 9.12	68.0	.177	0.31
Session Duration (minutes)	12.90 ± 2.80	9.80 ± 2.30	33.0	**.006****	0.60

Table 5. Statistical Comparison of User Interaction Metrics between **MultiColleagues (MC)** and **Baseline**.

5.2.4 Engagement and Flow Enhance Collaborative Immersion. Participants reported experiencing stronger engagement and conversational flow in MultiColleagues, with significantly higher ratings in the MultiColleagues condition (M = 5.70, SD = 1.38) compared to the Baseline condition (M = 4.45, SD = 1.57; W = 30.0, p = .014). Behavioral interaction measures further reinforced this pattern (Table 5). Participants contributed nearly twice as many utterances in MultiColleagues (M = 8.35, SD = 5.79) than in the Baseline condition (M = 4.10, SD = 2.45; W = 9.5, p = .001) and produced substantially more words overall (W = 18, p < .001). Sessions also lasted longer in MultiColleagues (M = 12.90, SD = 2.80) compared to Baseline (M = 9.80, SD = 2.30; W = 33, p = .006), suggesting that participants remained more cognitively and temporally invested when coordinating multiple voices. Interview reflections echoed these dynamics. P5 explained, "I feel like in my discussions, I'm constantly prompting new information and asking for new information, kind of wanting to lead the discussion. I feel I'm more like a facilitator." Others emphasized how the incremental delivery kept them attentive and invested (P3, P10, P12). While MultiColleagues fostered ongoing involvement, Baseline condition's one-shot responses were often described as efficient but less engaging. P19 described the interaction as "relatively one-way, like a presentation… I just pick what I want," and P8 admitting, "I didn't feel like we were collaborating… I just wanted to hear its perspective." These findings show that MultiColleagues' structured rhythm sustained deeper immersion and engagement, while Baseline delivered efficiency at the cost of shallower participation.

5.3 RQ2: Impact of Multiple AI Perspectives on Creative Outcomes

To address RQ2, we examined participants' evaluations of creative outcomes along three dimensions: creative exploration, process enrichment, and outcome quality and novelty. Results indicate that MultiColleagues was consistently rated more favorably than Baseline, with significant differences observed for creative exploration as well as for outcome quality and novelty (Table 6).

Metric	Q#	MC (M ± SD)	Baseline (M ± SD)	W	p-value	Effect Size (r)
Creative Exploration	Q1, Q4	6.00 ± 0.87	4.95 ± 1.55	31.5	**.018***	0.63
Process Enrichment	Q3	5.80 ± 1.20	5.00 ± 1.72	22.5	.054	0.45
Outcome Quality & Novelty	Q5, Q7	5.95 ± 0.92	4.97 ± 1.16	17.0	**<.01****	0.69

Table 6. Statistical Comparison for Performance & Integration Metrics between **MultiColleagues (MC)** and **Baseline**. Results show significantly higher ratings for MC on *Creative Exploration* and *Outcome Quality & Novelty*.

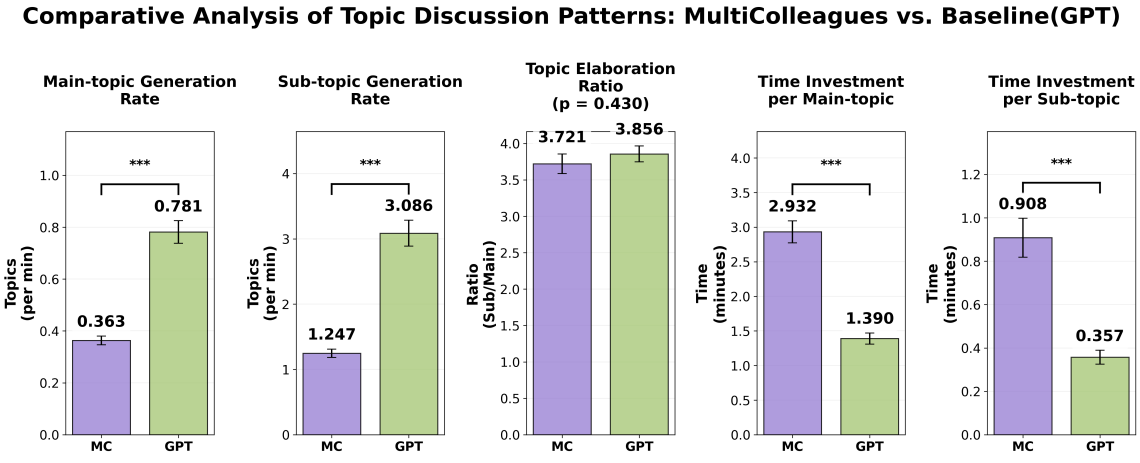

Fig. 5. This figure compares discussion patterns between MultiColleagues and Baseline. Baseline generated main topics (0.78 vs. 0.36 per minute) and sub-topics (3.09 vs. 1.25 per minute) at a faster rate, but branching ratios were comparable. MultiColleagues invested significantly more time per topic (2.93 vs. 1.39 minutes) and per sub-topic (0.91 vs. 0.36 minutes), supporting a slower, more deliberate style that enabled sustained idea elaboration.

5.3.1 MultiColleagues Broadens Creative Exploration. Survey measures confirmed that participants experienced greater creative exploration with MultiColleagues (M = 6.00, SD = 0.87) than with the Baseline condition (M = 4.95, SD = 1.55; W = 31.5, p = .018). To understand this difference, we analyzed conversational topical patterns, focusing on the pace at which new topics were introduced and the time spent developing them. In Figure 5, we found that Baseline produced new content at a faster pace, with more main topics per minute (M = 0.78 vs. 0.36, SD = 0.19 vs 0.07; W = 0.0, p < .001) and more sub-topics per minute (M = 3.09 vs. 1.25, SD = 0.89 vs 0.29; W = 0.0, p < .001). However, the *branching ratio* showed only a modest difference, with Baseline producing a slightly higher average branching ratio (M = 3.86, SD = 0.49) compared to MultiColleagues (M = 3.72, SD = 0.60; W = 83.0, p = 0.43), which indicates both systems supported relatively similar topical divergence levels. However, for investigating participants' time investment per topic, we found that participants in MultiColleagues resulted in more sustained exploration of each idea, where they spent over twice as much time on a single main topic (M = 2.93 vs. 1.39 min, SD = 0.71 vs. 0.36 min; W = 0.0, p < .001) and on each sub-topic (M = 0.91 vs. 0.36 min, SD = 0.40 vs. 0.14 min; W = 0.0, p < .001). This pattern indicates that MultiColleagues facilitated sustained exploration rather than rapid topic turnover. Interview reflections reinforced this sustained style of exploration. Several participants noted that the Baseline condition's "large initial response" often felt like "a presentation you just read" (P19), which discouraged further exploration. By contrast, MultiColleagues encouraged participants to "proactively join to think and discuss" when only a few points were offered (P3). These results

demonstrate that MultiColleagues broadened creative exploration by sustaining user-driven idea development beyond Baseline's faster topical output.

5.3.2 Traceable Perspectives Enable Structured Integration. We found from interviews that the traceability of perspectives in MultiColleagues supported participants' ability to integrate ideas. Because contributions were anchored to distinct AI colleagues, participants reported it was easier to follow up, remember, and combine ideas into coherent outcomes. P4 reflected that hearing "different angles, different perspectives" made the information "stay in your mind rather than just flashing by," while P3 emphasized the system "divides into several colleagues… when I want to go deeper, I clearly know which one to talk to." These role-based distinctions provided a clear map of where ideas originated, helping participants balance and integrate perspectives into more structured outputs. In contrast, the Baseline condition merged perspectives into a single response. While this blending reduced the ability to trace individual contributions, it benefited comprehensive coverage, offering "extensive lists and complete answers" when participants wanted a consolidated view (P2, P4, P6, P19).

5.3.3 Conversational Rhythm Supports Idea Quality and Novelty. Self-reported measures indicated that participants generated higher-quality and more novel ideas with MultiColleagues, with ratings significantly higher in the MultiColleagues condition (M = 5.95, SD = 0.92) than in the Baseline condition (M = 4.97, SD = 1.16; W = 17.0, $p < .01$). To better assess the uniqueness of participants' ideation, we performed TTCT analysis of originality [17, 31, 36, 37] (see details in Appendix B.2). The results showed that MultiColleagues (M = 3.78, SD = 0.38) scored higher than Baseline (M = 3.59, SD = 0.44; W = 70.0, $p = .202$), though this difference was not statistically significant. Despite this, participants consistently perceived MultiColleagues' outputs as more original. Interview accounts explained this perception by pointing to the rhythm of idea generation. Participants emphasized that MultiColleagues generated ideas progressively, in a rhythm they described as "digestible… aligned with the rhythm of human discussion" (P19). This stepwise unfolding supported a process of guided discovery, where ideas "emerge through guided conversation" (P2) and developed like a "chain of thought" (P16). In contrast, Baseline often produced an immediate burst of ideas. Participants acknowledged these outputs as "very creative right from the start" (P2), yet also noted that the density could feel overwhelming. Several reported that this rapid surge made it harder to refine and act on ideas, compared to the more deliberate, incremental approach offered by MultiColleagues (P3, P5, P15).

5.3.4 Breadth–Depth Trade-offs Shape Outcome Enrichment. Survey measures suggested that outcome enrichment was marginally stronger in the MultiColleagues condition (M = 5.80, SD = 1.20) compared to the Baseline condition (M = 5.00, SD = 1.72; W = 22.5, $p = .054$). This pattern reflects a breadth–depth trade-off between two systems. Interview reflections described MultiColleagues as a breadth-first approach, expanding the solution space through multiple diverse perspectives. Participants valued its ability to enrich the early, exploratory stages of ideation. P1 noted that it was effective for "expanding ideas," while P7 described its output felt "comprehensive" because it integrated "multiple angles [such as] logic and marketing," thereby fostering cross-functional thinking. This breadth encouraged novelty and variety but often came at the cost of providing concrete, actionable details. In contrast, the Baseline condition was viewed as a depth-first tool, excelling at producing more focused, polished, and immediately usable outcomes. P4 explained that it went "a step further" than MultiColleagues by providing tangible examples like "sample data," which offered a "more concrete understanding of how to process a dataset." Other highlighted its utility for "executive

decision-making" (P6) and for "executing idea and goal" (P12). This depth and implementability gave Baseline an advantage in later-stage tasks requiring clarity and execution, whereas MultiColleagues dominated in the initial phase by maximizing creative possibilities.

5.4 RQ3: System Design Features and User Agency in Creative Ideation

We evaluated how system design features influenced user agency (RQ3) during creative ideation across four metrics: user guidance, user control, adaptive thinking mode, and future use intent. As shown in Table 7, MultiColleagues received significantly higher ratings than Baseline for user control and adaptive thinking mode.

Metric	Q#	MC (M ± SD)	Baseline (M ± SD)	W	p-value	Effect Size (r)
User Guidance	Q2	5.85 ± 1.09	5.40 ± 1.39	34.5	.436	0.23
User Control	Q12	5.80 ± 1.32	4.40 ± 1.79	27.0	**.033***	0.55
Adaptive Thinking Mode	Q9	5.90 ± 1.29	4.60 ± 1.70	16.5	**.023***	0.58
Future Use Intent	Q11	6.15 ± 1.09	5.50 ± 1.61	11.5	.098	0.38

Table 7. Statistical Comparison for Agency & Control Metrics between **MultiColleagues (MC)** and **Baseline**. Results show significantly higher ratings for MC on *User Control* and *Adaptive Thinking Mode*.

5.4.1 Autonomous Versus Manual Direction Shape User Guidance. Participants' ratings on user guidance indicated no significant difference between MultiColleagues (M = 5.85, SD = 1.09) and Baseline conditions (M = 5.40, SD = 1.39; W = 34.5, p = .436). Nonetheless, interview findings revealed that the two systems supported user guidance in distinct ways. MultiColleagues was described as more autonomous, with participants noting that once a question was posed, the system could sustain its own line of discussion. Participants explained, "I can throw out a question, then they start an intense discussion" (P13, P15). This process was perceived as unfolding like a "chain of thought" (P16), where initial contributions created openings for participants to engage selectively. As P3 explained, "It gives you 1–2 points first, so you will proactively join to think and discuss". However, this autonomy also carried drawbacks, as conversations sometimes drifted from the intended focus and required effort to redirect (P12, P13). In contrast, interview accounts of the Baseline condition highlighted its reliance on manual direction. Participants described it as "question-answer, question-answer... I have to tell it what I want to do next for each step" (P13). While this approach offered precise control, it was effort-intensive and likened to "guid[ing] it in a very formal way, just like prompt engineering" (P16). Its polished and comprehensive outputs could also constrain participation, leaving some participants "lost, don't know how to chat or continue" (P3).

5.4.2 MultiColleagues Empowers Participants with Greater Control. Participants reported a significantly stronger sense of control when working with MultiColleagues (M = 5.80, SD = 1.32) compared to Baseline (M = 4.40, SD = 1.79; W = 27.0, p = .033). Interview reflections provided further insight into this perceived control through two main factors. First, participants attributed the enhanced control to MultiColleagues' support for shifting between explore and focus thinking modes. Participants could "move from exploring new ideas to working on a specific one" (P9) and "control it if I don't want it to be divergent" (P19). In contrast, the Baseline condition offered no such flexibility, leaving participants feeling they were "always fixing its direction" (P9). However, some valued its chunked, turn-by-turn dialogue, which created opportunities to "chime in to change direction at any time (P19)." Second, participants associated greater control with the system's alignment to their own goals. MultiColleagues was described as "more engaged, better aligned with

my original direction" (P12), and even supported leadership skills, as P16 noted, "I was able to bring in other agents when they were being quiet, which is actually a great team leadership learning experience." By contrast, Baseline was seen as "strong, professional, very dominant" (P3) to overshadow participants' intent.

At the same time, participants in interviews also acknowledged trade-offs in managing multiple colleagues' voices. Unlike Baseline, which was described as linear and predictable (P6), MultiColleagues required greater coordination effort. Participants pointed out that the diversity of roles occasionally caused drift, requiring "firm willpower to keep this conversation stable" (P6) or effort to pull AI colleagues "back on track" (P12, P13). P8 also described subtle "social pressure" when navigating overlapping perspectives. Overall, we found that MultiColleagues offered flexible and participatory control but demanded coordination, whereas Baseline provided predictable yet more rigid control.

5.4.3 Adaptive Thinking Mode Strengthens Flexibility in Ideation. Survey results showed that participants rated MultiColleagues (M = 5.90, SD = 1.29) significantly higher on adaptive thinking mode compared to the Baseline condition (M = 4.60, SD = 1.70; W = 16.5, p = .023), indicating stronger support for shifting between exploration and focus during ideation. Interview reflections further explained this flexibility. Participants emphasized MultiColleagues' ability to deliberately transition between divergent and convergent thinking, which they saw as central to managing the creative process. They described being able to "move from exploring new ideas to working on a specific one" (P9), and "control it if I don't want it to be divergent" (P19). Facilitation further supported these shifts by prompting reflection and offering lightweight guidance without imposing direction. Participants valued being asked whether to "dive deeper" or "explore other" directions (P15, P7), which helped them regulate attention and decide when to transition. As P6 described, the facilitator acted more as "a guide," providing reminders and summaries that supported concentration without steering outcomes. Participants also highlighted the value of explicit, manual controls for shifting modes. Button-based toggles made transitions quicker and more natural within the flow of ideation, reducing the need to type additional instructions. As P19 described, "the quickest way is to click to switch," while P20 noted that visible controls were preferable because "I don't have to type so many words, I can just click." Beyond convenience, this design also provided a structural trace of shifts, helping participants track how their workflow moved between breadth and depth.

5.4.4 Future Use Intent Depends on Task Stage and Context. Survey results reflected participants' high intent to use both systems, with MultiColleagues (M = 6.15, SD = 1.09) rated slightly higher than the Baseline condition (M = 5.50, SD = 1.61; W = 11.5, p = .098), though the difference was not statistically significant. Interview reflections further revealed how adoption was shaped by task stages and context. Many participants described a staged workflow in which MultiColleagues was used early to generate and expand ideas, followed by the Baseline condition to validate, refine, or translate those ideas into actionable steps (P6, P10). For example, P6 explained they would "start from [MultiColleagues] to find good new ideas, then bring this idea to [Baseline]" for detailed implementation, while P10 planned to "discuss a few ideas first" with MultiColleagues and then use Baseline to analyze feasibility and trade-offs. Besides, the suitability of each system was also tied to problem clarity. MultiColleagues was seen as particularly useful when questions required compound perspectives or when participants sought to brainstorm from multiple angles (P8, P18, P20). By contrast, the Baseline condition was considered more efficient and specific for narrow, practical, or well-defined tasks (P13, P15, P17). Participants also noted limitations during interviews. MultiColleagues was considered more suitable for large or complex problems that benefit from multiple perspectives (P7), ut some highlighted challenges such as a steeper learning curve and the need for longer engagement to fully realize its value (P9). These accounts suggest that future use is less about preferring one system overall and more about strategically aligning each with the stage, scope, and complexity of the problem at hand.

6 DISCUSSION

Design Implications	Present: AI as Tools	Future: AI as Colleagues
Broadening Perspectives		
Proactive & Multi-Voiced	AI waits for explicit prompts, delivers single-perspective or homogeneous answers	AI anticipates latent needs, initiates its own debates, sustains memory, and integrates diverse perspectives to broaden discussion
Seamless Collaboration		
Seamless Orchestration of Multi-Agent Flow	Interactions are sequential, Q&A-driven, requiring users to manually manage direction	System provides structured turn-taking, adaptive orchestration, or visual maps/ flows to balance diversity and order
Visualization & Identity Design for Recognizable Colleagues	Agents appear as generic text outputs with no recognizable identity	Colleagues with distinct visual/identity cues that reinforce social presence and credibility
Adaptive Conversational Style for Socialized, Contextual Interaction	Responses are uniform, mechanical, often verbose or formal	Colleagues adopt adaptive tone and style, with adjustable verbosity to mirror natural team conversation
Trust & Accountability		
Embedding Evidential Transparency for Trust Calibration	Contributions lack provenance, leaving users to trust blindly or verify externally	Responses include verifiable traces for users to assess reliability and treat input as trustworthy colleague contributions
Sustaining Social Presence with Visible Accountability	Silent or inactive personas feel like "irresponsible tools," reducing perceived trust and teamwork	System ensures each persona' s contributions are visible to reinforce accountability and establish them as reliable team members

Fig. 6. **Design Implications for Human–Multi-Agent Collaboration: From AI as Tools to AI as Colleagues.** This table summarizes six design implications that guide the transition from single-agent, tool-like AI to multi-agent colleagues in collaborative ideation. Each implication highlights a key design consideration (left column) and contrasts current AI as task-oriented tools (middle) with envisioned dynamics of human–multi-agent collaboration, where AI act as proactive, accountable, and socially embedded colleagues (right).

6.1 Summary of Results

Our study demonstrates how MultiColleagues, a multi-agent conversational system, reshapes collaborative ideation compared to single-agent baselines. First, for **RQ1 - Collaborative Experience**, MultiColleagues fostered a distributed, team-like atmosphere. Participants reported stronger team-like feelings and complementary strengths across roles. Engagement also increased, with nearly twice as many utterances, longer sessions, and more words produced overall, reflecting a shift from passive receipt to facilitative coordination. For **RQ2 - Creative Outcomes**, multiple AI perspectives supported broader and deeper exploration: participants spent more time developing each idea topic and producing outputs judged higher in quality and novelty. For **RQ3 - System Design and User Agency**, MultiColleagues enabled stronger perceived control and more adaptive thinking mode support for switching between divergent and convergent thinking. Although autonomy sometimes led to conversational drift, participants valued the ability to

steer discussions, regulate rhythm, and align outputs with evolving goals. Our results indicate that multi-agent systems shift ideation from tool use toward dynamics that resemble collegial teamwork. Building on these findings, we outline key design implications that translate observed collaboration patterns into actionable directions for future system design, detailed in the following sections (Figure 6).

6.2 Expanding Perspectives through Proactive Multi-Agent Collaboration

Our study demonstrated how MultiColleagues broadened user perspectives by combining multiple voices with proactive contributions. Instead of passively awaiting prompts, colleagues anticipated latent needs and introduced dimensions users had not explicitly considered, thereby helping them move beyond current frames of thought. At the same time, colleagues did not simply deliver isolated or premature convergence ideas [31]; they engaged in internal contrasts and comparisons, allowing alternative viewpoints to emerge and be weighed before reaching the user. Building on this, we frame the MultiColleagues system as a foundation for multi-voiced conversation, where diverse AI colleagues contribute proactively and in contrast to one another, creating exchanges that are both cognitively supportive and closer to real teamwork.

6.2.1 Design Implication 1: Supporting Proactive and Multi-voiced Colleagues. Prior work has noted the risks of homogeneous perspectives in AI systems, while also emphasizing the potential of multi-agent conversation to counteract them by surfacing diverse viewpoints and stimulating creative exchanges [57, 70, 110]. Our study extends this line of inquiry by showing how participants valued proactive need recognition and autonomous idea development as mechanisms to broaden perspectives and sustain engagement. To function as more autonomous, team-like collaborators, AI colleagues should not only anticipate latent user needs but also initiate their own debates and substantiate claims with evidence. Building on prior work showing how generative agents can sustain memory and social interaction over time [71], future systems should also consider how designs can scale from small groups of colleagues to larger collectives while preserving the richness of multiple perspectives [72]. By synthesizing diverse viewpoints internally and proactively, such systems can provide more balanced and well-reasoned insights that better leverage collective intelligence.

6.3 Orchestrating Many Voices: Designing Multi-Agent Colleagues for Seamless Collaboration

While multi-agent autonomy offers opportunities for richer perspectives, it also raises challenges of coordination and control. Our MultiColleagues system demonstrated how multi-agent conversation supports idea integration by making perspectives traceable and diversifying problem-solving. Unlike single-model interactions that remain sequential and Q&A-driven, MultiColleagues exchanges naturally introduced complementary viewpoints, encouraging participants to synthesize across roles. The progressive orchestration of colleagues promoted smoother collaboration by breaking down problems into multiple angles and guiding participants toward more deliberate integration of ideas.

6.3.1 Design Implication 2: Structuring Conversations for Seamless Flow. Managing contributions from multiple AI colleagues requires orchestration strategies that surface diverse perspectives without overwhelming users. Prior work highlights approaches such as round-table settings, sequential workflows, phased role-play, and chat-based interface for balancing diversity and order in multi-agent systems [25, 57, 70, 73]. Our study extends this work by showing that a conversational chat style with clear turn-taking reduced barriers to entry, preserved distributed expertise, and created a digestible rhythm that supported both engagement and coordination.

Yet questions remain about how best to orchestrate multi-agent conversations across different user groups and task contexts. While sequential turn-taking fostered clarity, participants also envisioned more flexible designs, such as reconfiguring colleagues mid-discussion, introducing simultaneous responses, or staging role-based debates. Addressing the cognitive complexity of many voices further calls for visual orchestration aids, such as color-coded dialogue flows or labeled discussion maps, that provide overviews and allow lightweight steering of complex exchanges. Future research should explore how adaptive orchestration mechanisms, ranging from sequential turns to parallel exchanges, can dynamically adjust based on user preferences, task phases, or cognitive load, moving toward seamless collaboration experiences that scale across diverse ideation tasks.

6.3.2 Design Implication 3: Designing Visualization and Identity for Recognizable Colleagues. The appearance and identity cues of AI colleagues play a critical role in shaping how users perceive and engage with them, often creating a heightened sense of "being in a team" (P5). To evoke this social presence in our system, we assigned distinct profile avatars to each colleague, closely resembling profile pictures in real-world office platforms. This design choice assisted participants with a familiar visual anchor and reinforced the impression that they were collaborating with recognizable teammates rather than interacting with abstract system outputs. Looking forward, effective collaboration requires moving beyond simple profile images toward a more comprehensive identity package. Recent HCI research shows that projecting digital colleagues into office spaces [67], rendering two-dimensional images as three-dimensional avatars [39], and employing AR-based or customized avatars in meetings [89] can strengthen social presence and improve perceptions of credibility in group work [83, 84]. At the same time, visual identity design must avoid reinforcing stereotypes, since default gendering or normative depictions risk perpetuating bias [4]. Providing gender-neutral options, diverse visual styles, and customizable pronouns ensures AI colleagues are both recognizable and inclusive, which in turn supports more authentic and equitable collaboration.

6.3.3 Design Implication 4: Orchestrating Conversational Style for Socialized and Contextual Interactions. The communicative style of AI colleagues strongly influenced how smoothly information was exchanged and understood. Users' interactions felt closer to peer-to-peer collaboration rather than mechanical delivery when AI responses adopted a more approachable and adaptable tone that could be polite, occasionally humorous, or appropriately formal. Subtle stylistic shifts created a sense of social presence across conversations that resemble real workplace exchanges and making shared content easier to follow. In addition, response length also shaped the rhythm of collaboration. Participants reflected that longer outputs were described as valuable in "focus mode" to offer more in-depth reasoning, while shorter and contextualized statements helped anchor contributions without overwhelming the discussion. Future works could focus on designing adaptive mechanisms that vary in verbosity across phases of a task, from elaboration at the outset to concise handovers later, that can support both breadth–depth exploration and efficient knowledge sharing.

6.4 From "Many Agents" to "Colleagues": Establishing Peer-Like Roles and Trust

Building on the foundations of multi-perspective proactivity (Section 6.2) and seamless orchestration (Section 6.3), a further step is required before AI can be regarded as genuine colleagues rather than mere tools: their contributions must be perceived as trustworthy and worth integrating. Our study shows that multi-agent role-playing can significantly reshape how users perceive AI systems. Compared to single-agent interactions, MultiColleagues fostered a stronger sense of teamwork: participants described the experience as "hosting a meeting" where different personas consistently contributed ideas and played distinct roles. Some noted that the system felt "between a tool and a teammate" (P9), capturing the transitional stage from instrumentality to collegiality. More importantly, the stability of

persona responses and the diversity of perspectives encouraged participants to consider AI contributions as reliable and substantive, laying the groundwork for emerging trust (P20). This aligns with recent HCI work on multi-agent conversational systems, demonstrating that structured role distribution and visible participation can create stronger impressions of social presence and team-like collaboration [70, 90].

6.4.1 Design Implication 5: Embedding Evidential Transparency for Trust Calibration. Our findings suggest that users are more willing to treat agent contributions as "worth considering" when the reasoning and provenance behind those contributions are visible. Role-playing alone is insufficient to establish trust and may risk being perceived as performative. Systems therefore need to embed verifiable traces into persona responses, such as inline references, source tags, or concise capability statements [106]. This allows users to judge which information is reliable and prevents blind reliance on AI outputs [41, 49]. In this sense, transparency becomes a second layer of trust beyond social presence, enabling users to calibrate adoption of AI suggestions in the same way they evaluate colleagues' inputs [29].

6.4.2 Design Implication 6: Sustaining Social Presence with Visible Accountability. Trust in colleagueship also depends on the agent's consistent presence and accountability. When a persona remained silent or inactive, participants immediately associated it with "an irresponsible coworker in a meeting," which undermined both trust and the sense of teamwork. To address this, systems should incorporate interface mechanisms that maintain the visibility of each persona's contributions, such as surfacing participation levels or signaling inactive participation. Such visibility ensures that personas are perceived not only as multiple voices but also as accountable members of a social group [29, 70, 110]. By making contributions consistently observable, users can move beyond a purely supervisory stance and begin to regard AI as reliable discussion partners.

6.5 Ethical Considerations

The move toward AI as colleagues also raises significant ethical questions. As AI systems increasingly contribute to creative work, scientific discovery, and decision-making, the boundaries of authorship and accountability become less clear. For example, in recent academic venues, the emergence of AI as listed co-authors and even first authors illustrates both the promise and the tension of this trajectory [9, 43]. While such visibility acknowledges the substantive role of AI in knowledge production, it also challenges long-standing norms of intellectual credit and responsibility. Beyond authorship, issues of transparency, bias, and user agency must be addressed. As AI colleagues autonomously advance discussions and generate new perspectives, mechanisms are needed to ensure that their contributions remain interpretable and that humans retain meaningful influence over outcomes. Without such measures, agentic AI risks creating "moral crumple zones" where responsibility becomes diffused and no actor is fully accountable [65]. Designing for auditability, feedback integration, and human oversight will therefore be critical. The collegial paradigm cannot be achieved without parallel efforts to establish ethical frameworks that safeguard accountability while enabling AI to act as trusted partners in collaborative processes.

6.6 Limitations and Future Work

While our study offers initial insights into AI colleagues, it also faces several limitations that open up important directions for future research.

6.6.1 A homogeneous participant pool limits generalizability. Our participant pool was largely composed of students and early-career professionals, which restricted the generalizability of our findings. However, the idea of AI colleagues

extends across diverse professional and cultural contexts, where expectations of teamwork, trust, and authority are shaped by domain practices and cultural norms not represented in our sample. Future work should recruit participants from varied occupations, seniority levels, and cultural backgrounds to better examine how different groups perceive and integrate AI colleagues in real-world collaborations.

6.6.2 Short sessions and limited adaptivity constrained collaboration quality. Participant interactions were brief, averaging about ten minutes. Most of this time was spent on divergent exploration, leaving little opportunity for convergence, refinement, or finalized ideas. Richer outcomes could emerge in longer sessions that scaffold cycles of exploration, debate, and synthesis, supported by features such as structured debate mechanisms, staged convergence prompts, or intentionally conflicting agent perspectives. Second, the system lacked adaptivity to different levels of expertise. Experts preferred direct and detailed engagement, while novices benefited more from guided observation or scaffolded entry points. Future designs should incorporate adaptive modes that dynamically tailor agent behavior to user expertise. Third, redundancy was observed between facilitator prompts and summaries, which often repeated similar content instead of complementing one another. Refining the division of labor between these features would make collaboration more efficient and less repetitive.

6.6.3 Shared persona architecture reduced diversity and limited scalability. Although role prompts provided surface-level differentiation, all personas relied on the same language model, which reduced the diversity of their contributions and sometimes led to stylistic convergence. Future work should investigate strategies to foster more authentic plurality, such as combining heterogeneous models or enforcing stronger divergence in stance and knowledge sources. In addition, scalability remains an open question. While our prototype involved nine personas, real-world work settings often include much larger groups of colleagues, with interactions that are sustained over longer periods. Future systems should therefore explore orchestration strategies, such as dynamic selection and structured turn-taking, that balance coherence with the complexity of real group dynamics.

7 CONCLUSION

To investigate how AI might move beyond functioning as tools toward performing as peer-like colleagues in collaborative contexts, we developed MultiColleagues, a system that orchestrates multiple role-differentiated personas and incorporates facilitation and thinking-mode features to support structured ideation. In a within-subjects study that compared MultiColleagues with a single-agent baseline, we found that our system produced more novel and higher-quality outcomes, cultivated a stronger sense of team presence, and encouraged participants to engage more actively and exercise greater control over the interaction. Building on these findings, we derived several design implications for amplifying proactive contributions, enabling more seamless human–agent coordination, and supporting calibrated trust. Taken together, this work advances the growing body of research on multi-agent systems by showing how design choices shape whether users perceive AI as tools or as peer-like collaborators. We hope these insights will inform future research and design efforts aimed at creating generative multi-agent systems that not only enhance human creativity but also lay the groundwork for trustworthy and sustainable forms of AI colleagueship.

REFERENCES

[1] 2024. CrewAI. https://www.crewai.com/. Accessed: 2025-08-28.
[2] Mary Abkemeier. 2020. Cognitive Load Theory. *Encyclopedia of Education and Information Technologies* (2020). https://api.semanticscholar.org/CorpusID:60459788

[3] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All too human? Mapping and mitigating the risk from anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 13–26.

[4] Nouar Aldahoul, Talal Rahwan, and Yasir Zaki. 2024. AI-generated faces influence gender stereotypes and racial homogenization. *Scientific Reports* 15 (2024). https://api.semanticscholar.org/CorpusID:267406826

[5] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Mag.* 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[6] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Patrick Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Doron Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow, UK, 1–13. https://doi.org/10.1145/3290605.3300233

[7] Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. 2024. LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop. arXiv:2402.09346 [cs.AI] https://arxiv.org/abs/2402.09346

[8] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[9] Navneet Ateriya, Nagendra Singh Sonwani, Kishor Singh Thakur, Arvind Kumar, and Satish Kumar Verma. 2025. Exploring the ethical landscape of AI in academic writing. *Egyptian Journal of Forensic Sciences* 15, 1 (2025), 36.

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

[11] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[12] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376* (2023). LLM agent integrating 18 expert tools for organic synthesis, drug discovery, materials design.

[13] Virginia Braun and Victoria Clarke. 2024. Thematic analysis. In *Encyclopedia of quality of life and well-being research*. Springer, 7187–7193.

[14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

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

[16] Joel Chan, Zijian Ding, Eesh Kamrah, and Mark Fuge. 2024. Formulating or Fixating: Effects of Examples on Problem Solving Vary as a Function of Example Presentation Interface Design. 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 179, 16 pages. https://doi.org/10.1145/3613904.3642653

[17] Jia Chi. 2024. The evolutionary impact of artificial intelligence on contemporary artistic practices. *Commun. Humanit. Res* 35, 1 (2024), 6–11.

[18] Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[19] Elizabeth Clark, Abigail See Ross, Samuel Bowman, Ronan Le Bras, et al. 2022. Wordcraft: A Human–AI Collaborative Editor for Story Writing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3491102.3517580 CHI'22 system paper introducing Wordcraft (built on LaMDA) for collaborative story writing.

[20] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[21] Christopher Cui, Xiangyu Peng, and Mark Riedl. 2023. Thespian: Multi-character text role-playing game agents. *arXiv preprint arXiv:2308.01872* (2023).

[22] K. Z. Cui. 2024. Evidence from a Field Experiment with GitHub Copilot. *MIT Generative AI Field Experiments (pubpub)* (2024). RCT at Microsoft & Accenture finds 7–22% more pull requests per week.

[23] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid Intelligence. *Business amp; Information Systems Engineering* 61, 5 (March 2019), 637–643. https://doi.org/10.1007/s12599-019-00595-2

[24] Design Council. 2005. The Double Diamond: A universally accepted depiction of the design process. https://www.designcouncil.org.uk/our-resources/the-double-diamond. Accessed August 28, 2025.

[25] Victor Dibia, Jian Chen, Gagan Bansal, Shahbaz Syed, Adam Fourney, Eric Zhu, and Saleema Amershi. 2024. Autogen Studio: A no-code developer tool for building and debugging multi-agent systems. *arXiv preprint arXiv:2408.15247* (2024).

[26] Zijian Ding, Michelle Brachman, Joel Chan, and Werner Geyer. 2025. "The Diagram is like Guardrails": Structuring GenAI-assisted Hypotheses Exploration with an Interactive Shared Representation. arXiv:2503.16791 [cs.HC] https://arxiv.org/abs/2503.16791

[27] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

[28] Alpana Dubey, Kumar Abhinav, Sakshi Jain, Veenu Arora, and Asha Puttaveerana. 2020. HACO: a framework for developing human-AI teaming. In *Proceedings of the 13th Innovations in Software Engineering Conference (formerly known as India Software Engineering Conference)*. 1–9.

[29] Thomas Erickson and Wendy A. Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM Trans. Comput.-Hum. Interact.* 7, 1 (March 2000), 59–83. https://doi.org/10.1145/344949.345004

[30] John H. Flavell. 1979. Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry. *American Psychologist* 34 (1979), 906–911. https://api.semanticscholar.org/CorpusID:8841485

[31] Kazuma Fukumura and Takayuki Ito. 2025. Can LLM-Powered Multi-Agent Systems Augment Human Creativity? Evidence from Brainstorming Tasks. In *Proceedings of the ACM Collective Intelligence Conference (CI '25)*. Association for Computing Machinery, New York, NY, USA, 20–29. https://doi.org/10.1145/3715928.3737479

[32] Tianyu Gao and other collaborators. 2023. PaperQA: LLMs as Research Assistants. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL. Pipeline for answering questions from scientific papers via retrieval + summarization.

[33] 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*. 1–13.

[34] Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36, 2 (2004), 193–202. https://doi.org/10.3758/BF03195564

[35] Tianhao Guo, Xiaohan Chen, Yucheng Wang, Rui Chang, Shu Pei, Nitesh V Chawla, and Xianpei Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

[36] Erik Guzik, Christian Byrge, and Christian Gilde. 2023. The Originality of Machines: AI Takes the Torrance Test. *Journal of Creativity* 33 (08 2023), 100065. https://doi.org/10.1016/j.yjoc.2023.100065

[37] Eran Hadas and Arnon Hershkovitz. 2024. Using large language models to evaluate alternative uses task flexibility score. *Thinking Skills and Creativity* 52 (2024), 101549.

[38] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[39] Bernhard Hilpert, Claudio Alves da Silva, Leon Christidis, Chirag Bhuvaneshwara, Patrick Gebhard, Fabrizio Nunnari, and Dimitra Tsovaltzi. 2024. Avatar Visual Similarity for Social HCI: Increasing Self-Awareness. arXiv:2408.13084 [cs.HC] https://arxiv.org/abs/2408.13084

[40] Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. arXiv:2308.00352 [cs.AI] https://arxiv.org/abs/2308.00352

[41] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. 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 1045, 15 pages. https://doi.org/10.1145/3613904.3641895

[42] Ziheng Huang, Kexin Quan, Joel Chan, and Stephen MacNeil. 2023. CausalMapper: Challenging designers to think in systems with Causal Maps and Large Language Model. In *Proceedings of the 15th Conference on Creativity and Cognition* (Virtual Event, USA) *(CC '23)*. Association for Computing Machinery, New York, NY, USA, 325–329. https://doi.org/10.1145/3591196.3596818

[43] P Bernt Hugenholtz and João Pedro Quintais. 2021. Copyright and artificial creation: does EU copyright law protect AI-assisted output? *IIC-International Review of Intellectual Property and Competition Law* 52, 9 (2021), 1190–1216.

[44] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561* (2023).

[45] Hyeonsu B Kang, David Chuan-En Lin, Yan-Ying Chen, Matthew K. Hong, Nikolas Martelaro, and Aniket Kittur. 2025. BioSpark: Beyond Analogical Inspiration to LLM-augmented Transfer. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 653, 29 pages. https://doi.org/10.1145/3706598.3714053

[46] Ahmet Baki Kocaballi, Emre Sezgin, Leigh Clark, John M Carroll, Yungui Huang, Jina Huh-Yoo, Junhan Kim, Rafal Kocielnik, Yi-Chieh Lee, Lena Mamykina, et al. 2022. Design and evaluation challenges of conversational agents in health care and well-being: selective review study. *Journal of medical Internet research* 24, 11 (2022), e38525.

[47] Shalom Lappin. 2024. Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information* 33, 1 (2024), 9–20.

[48] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–35.

[49] Seongmin Lee, Zijie J Wang, Aishwarya Chakravarthy, Alec Helbling, ShengYun Peng, Mansi Phute, Duen Horng Polo Chau, and Minsuk Kahng. 2025. Llm attributor: Interactive visual attribution for llm generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 29655–29657.

[50] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597* (2023).

[51] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.

[52] Hang Li, Yucheng Chu, Kaiqi Yang, Yasemin Copur-Gencturk, and Jiliang Tang. 2025. LLM-based Automated Grading with Human-in-the-Loop. arXiv:2504.05239 [cs.CL] https://arxiv.org/abs/2504.05239

[53] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118* (2023).

[54] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[55] Yiren Liu, Pranav Sharma, Mehul Oswal, Haijun Xia, and Yun Huang. 2025. PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 506–534. https://doi.org/10.1145/3715336.3735789

[56] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170* (2023).

[57] Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024. LLM discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373* (2024).

[58] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016). https://api.semanticscholar.org/CorpusID:1036498

[59] Stephen MacNeil, Zijian Ding, Kexin Quan, Ziheng Huang, Kenneth Chen, and Steven P. Dow. 2021. ProbMap: Automatically constructing design galleries through feature extraction and semantic clustering. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 134–136. https://doi.org/10.1145/3474349.3480203

[60] Antonio Mastropaolo, Luca Pascarella, Emanuela Guglielmi, Matteo Ciniselli, Simone Scalabrino, Rocco Oliveto, and Gabriele Bavota. 2023. On the Robustness of Code Generation Techniques: An Empirical Study on GitHub Copilot. In *Proceedings of the 45th International Conference on Software Engineering* (Melbourne, Victoria, Australia) *(ICSE '23)*. IEEE Press, 2149–2160. https://doi.org/10.1109/ICSE48619.2023.00181

[61] Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with CohMetrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

[62] Jacob Menick, Tyna Eloundou, Jan Leike, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147* (2022). DeepMind GopherCite: retrieval + citations to reduce hallucination in long-form QA.

[63] Rahul Mohanani, Burak Turhan, and Paul Ralph. 2021. Requirements Framing Affects Design Creativity. *IEEE Transactions on Software Engineering* 47, 5 (May 2021), 936–947. https://doi.org/10.1109/tse.2019.2909033

[64] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming. 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 142, 16 pages. https://doi.org/10.1145/3613904.3641936

[65] Anirban Mukherjee and Hannah Hanwen Chang. 2025. Agentic AI: Autonomy, Accountability, and the Algorithmic Society. *arXiv preprint arXiv:2502.00289* (2025).

[66] Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2023. Skeleton-of-thought: Large language models can do parallel decoding. *Proceedings ENLSP-III* (2023).

[67] Anya Osborne, Sabrina Fielder, Lee Taber, Tara Lamb, Joshua McVeigh-Schultz, and Katherine Isbister. 2025. Avatars and Environments for Meetings in Social VR: What Styles and Choices Matter to People in Group Creativity Tasks? arXiv:2506.21780 [cs.HC] https://arxiv.org/abs/2506.21780

[68] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

[69] Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity? arXiv:2309.05196 [cs.CL] https://arxiv.org/abs/2309.05196

[70] Rock Yuren Pang, Rohit Maheshwari, Julie Yu, and Katharina Reinecke. 2025. Synthetic Conversation: How Computing Researchers Engage Multi-Perspective Dialogues to Brainstorm Societal Impacts. 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 497, 7 pages. https://doi.org/10.1145/3706599.3719747

[71] 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*. 1–22.

[72] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. arXiv:2411.10109 [cs.AI] https://arxiv.org/abs/2411.10109

[73] Yingzhe Peng, Xiaoting Qin, Zhiyang Zhang, Jue Zhang, Qingwei Lin, Xu Yang, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. 2024. Navigating the Unknown: A Chat-Based Collaborative Interface for Personalized Exploratory Tasks. arXiv:2410.24032 [cs.HC]

[74] Sandra Peter, Kai Riemer, and Jevin D. West. 2025. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences* 122, 22 (2025), e2415898122. https://doi.org/10.1073/pnas.2415898122 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2415898122

[75] Alexander Pollatsek and Arnold D Well. 1995. On the use of counterbalanced designs in cognitive research: a suggestion for a better and more powerful analysis. *Journal of Experimental psychology: Learning, memory, and Cognition* 21, 3 (1995), 785.

[76] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476* (2023).

[77] Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2025. Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination. arXiv:2409.14634 [cs.HC] https://arxiv.org/abs/2409.14634

[78] Jude Rayan, Dhruv Kanetkar, Yifan Gong, Yuewen Yang, Srishti Palani, Haijun Xia, and Steven P. Dow. 2024. Exploring the Potential for Generative AI-based Conversational Cues for Real-Time Collaborative Ideation. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C&C '24)*. Association for Computing Machinery, New York, NY, USA, 117–131. https://doi.org/10.1145/3635636.3656184

[79] Alistair Reid, Simon O'Callaghan, Liam Carroll, and Tiberio Caetano. 2025. Risk Analysis Techniques for Governed LLM-based Multi-Agent Systems. *arXiv preprint arXiv:2508.05687* (2025).

[80] Mark A. Runco, Jonathan A. Plucker, and Wei Lim. 2001. Development and psychometric integrity of a measure of ideational behavior. *Creativity Research Journal* 13, 3-4 (2001), 393–400.

[81] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. 2017. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. arXiv:1707.05173 [cs.AI] https://arxiv.org/abs/1707.05173

[82] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. 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 836, 18 pages. https://doi.org/10.1145/3613904.3642621

[83] Katie Seaborn. 2025. Social Identity in Human-Agent Interaction: A Primer. *ACM Transactions on Human-Robot Interaction* (Aug. 2025). https://doi.org/10.1145/3760500

[84] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the Effects of Embodiment for a Group Facilitation Agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173965

[85] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623, 7987 (2023), 493–498.

[86] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158* (2023).

[87] Ben Shneiderman. 2022. *Human-Centered AI.* Oxford University Press, New York, NY.

[88] Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. 2016. IdeaHound: Improving Large-scale Collaborative Ideation with Crowd-Powered Real-time Semantic Modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 609–624. https://doi.org/10.1145/2984511.2984578

[89] Pitch Sinlapanuntakul and Mark Zachry. 2025. Perception in Pixels: Effects of Avatar Representation in Video-Mediated Collaborative Interactions. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work (CHIWORK '25)*. ACM, 1–16. https://doi.org/10.1145/3729176.3729183

[90] Lipeipei Sun, Tianzi Qin, Anran Hu, Jiale Zhang, Shuojia Lin, Jianyan Chen, Mona Ali, and Mirjana Prpa. 2025. Persona-L has Entered the Chat: Leveraging LLMs and Ability-based Framework for Personas of People with Complex Needs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1109, 31 pages. https://doi.org/10.1145/3706598.3713445

[91] Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. 2023. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. *arXiv preprint arXiv:2310.00280* (2023).

[92] Wannita Takerngsaksiri, Jirat Pasuksmit, Patanamon Thongtanunam, Chakkrit Tantithamthavorn, Ruixiong Zhang, Fan Jiang, Jing Li, Evan Cook, Kun Chen, and Ming Wu. 2025. Human-In-the-Loop Software Development Agents. arXiv:2411.12924 [cs.SE] https://arxiv.org/abs/2411.12924

[93] Pamela Tierney and Steven M. Farmer. 2002. Creative self-efficacy: Its potential antecedents and relationship to creative performance. *Academy of Management Journal* 45, 6 (2002), 1137–1148.

[94] Maarten W Van Someren, Yvonne F Barnard, Jacobijn AC Sandberg, et al. 1994. The think aloud method: a practical approach to modelling cognitive processes. *London: AcademicPress* 11, 6 (1994).

[95] Samangi Wadinambiarachchi, Ryan M. Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Velloso. 2024. The Effects of Generative AI on Design Fixation and Divergent Thinking. 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 380, 18 pages. https://doi.org/10.1145/3613904.3642919

[96] Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224* (2025).

[97] Xinyu Wang, Bowen Li, Yizhou Song, Frank F Xu, Xiaotao Tang, Mingxuan Zhuge, and Graham Neubig. 2024. OpenHands: An open platform for AI software developers as generalist agents. *arXiv preprint arXiv:2407.16741* (2024).

[98] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. arXiv:2310.00746 [cs.CL] https://arxiv.org/abs/2310.00746

[99] Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835* (2023).

[100] Mengke Wu, Kexin Quan, Weizi Liu, Mike Yao, and Jessie Chin. 2025. Incorporating Personality into AI Writing Companions: Mapping the Design Space. 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 341, 9 pages. https://doi.org/10.1145/3706599.3720185

[101] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. https://doi.org/10.48550/arXiv.2308.08155 arXiv:2308.08155 [cs].

[102] Zihan Wu, Chengzhi Han, Zijian Ding, Zeyu Weng, Zekun Liu, Shunyu Yao, and Lingpeng Kong. 2024. OS-Copilot: Towards generalist computer agents with self-improvement. *arXiv preprint arXiv:2402.07456* (2024).

[103] Xiaotong (Tone) Xu, Rosaleen Xiong, Boyang Wang, David Min, and Steven P. Dow. 2021. IdeateRelate: An Examples Gallery That Helps Creators Explore Ideas in Relation to Their Own. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 352 (Oct. 2021), 18 pages. https://doi.org/10.1145/3479496

[104] Xiaotong (Tone) Xu, Jiayu Yin, Catherine Gu, Jenny Mar, Sydney Zhang, Jane L. E, and Steven P. Dow. 2024. Jamplate: Exploring LLM-Enhanced Templates for Idea Reflection. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 907–921. https://doi.org/10.1145/3640543.3645196

[105] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* 18, 6 (2024), 1–32.

[106] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI Decision Support Systems. 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 14, 14 pages. https://doi.org/10.1145/3544548.3581393

[107] Yaqing Yang, Vikram Mohanty, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K. Hong. 2025. From Overload to Insight: Scaffolding Creative Ideation through Structuring Inspiration. arXiv:2504.15482 [cs.HC] https://arxiv.org/abs/2504.15482

[108] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 841–852.

[109] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[110] Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024. See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles. 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 484, 24 pages. https://doi.org/10.1145/3613904.3642545

A APPENDIX A

A.1 Pre-Survey Creativity Questionnaire

Participants rated their agreement with the following statements on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). The scale demonstrated excellent internal consistency (Cronbach's $\alpha = 0.956$).

(1) I often come up with new and practical ideas to improve performance.
(2) I search for new technologies, techniques, or solutions.
(3) I suggest new ways to increase the quality of work or outcomes.
(4) I am a good source of creative ideas.
(5) I come up with creative solutions to problems.
(6) I often have a fresh approach to challenges.

(7) I am willing to take risks in generating new ideas.
(8) I promote and support ideas that I believe in.
(9) I create detailed plans for implementing new ideas.
(10) I exhibit creativity when given the opportunity.
(11) I consider myself a creative person.

A.2 Post-System Evaluation Questionnaire

Participants rated their agreement with the following statements on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree).

Q1. The system encouraged creative thinking and helped me explore a wide range of ideas.

Q2. I was able to guide the idea generation based on my needs during the task.

Q3. This system benefits/enriches my ideation process and thinking.

Q4. I reached lots of valuable or actionable ideas that felt better than what I might have generated alone.

Q5. Working with the AI system allows me to develop more creative solutions that I would not have come up with on my own.

Q6. Interacting with the system felt like working with a helpful teammate.

Q7. The system offered useful perspectives that expanded or deepened my thinking.

Q8. When working with this AI system, everyone (human/AI) can contribute their strengths and complement each other in the best possible way.

Q9. I was able to shift between exploring new ideas and focusing on specific ones as needed.

Q10. The session kept me mentally engaged and the interaction felt smooth and well-paced.

Q11. I would use a system like this again for brainstorming or planning in the future.

Q12. Did the system make you feel more in control of the creative process (rather than more guided by the AI)?

B APPENDIX B: TOPIC CATEGORIZATION & EVALUATION SCORING METHODS

B.1 Linguistic and Pragmatic Style Scoring Method

Linguistic Cohesion scores were computed using Coh-Metrix indices, focusing on four constructs: Narrativity (PCNARz, PCNARp), Syntactic Simplicity (PCSYNz, PCSYNp), Word Concreteness (PCCNCz, PCCNCp), and Referential Cohesion (PCREFz, PCREFp). Each participant's text inputs across both conditions were analyzed, scores were computed at the utterance level. For each condition, individual participants' values were averaged across all their contributions, yielding a single mean score per metric per participant. These participant-level means formed the basis of the within-subject comparisons reported in the upper Table 4. **Pragmatic and interaction style metrics** were assessed through a structured annotation protocol. Two trained coders worked with GPT-5 collaboratively and rated each participant's utterances on five dimensions: sentiment, formality, directness, relational orientation, and participation. The rating is calculated from a 7-point Likert scale anchored at "extremely informal/indirect/hierarchical" through "extremely formal/direct/equal." Scores were averaged across all utterances per participant within each condition to obtain individual means. These per-participant averages were then statistically compared across conditions using Wilcoxon signed-rank tests, with results summarized in lower Table 4.

Pragmatic Classification Metrics Definition

You are a text classification assistant. Your task is to analyze user input sentences and rate them on **five metrics**: Sentiment, Formality, Directness, Relationship, and Participation. Each metric is rated on a **1–7 scale**, where 1 = lowest/negative/extreme, 7 = highest/positive/extreme. Below shows rating scale definitions:
Sentiment (Emotional Valence): { "scale": 1 = Very Negative (critical, dismissive, frustrated); 2 = Slightly Negative (mild disapproval, doubt); 3 = Neutral–Negative (matter-of-fact, slight negativity); 4 = Neutral (balanced, no polarity); 5 = Neutral–Positive (mildly encouraging, constructive); 6 = Positive (supportive, motivated, curious); 7 = Very Positive (strong enthusiasm, praise)}

Formality (Language Style & Register): { "scale": 1 = Extremely Informal (slang, shorthand); 2 = Very Informal (casual, typos); 3 = Slightly Informal (conversational, clear); 4 = Neutral/Mixed (everyday phrasing, clear but not polished); 5 = Slightly Formal (structured, includes technical terms); 6 = Very Formal (professional/academic tone); 7 = Extremely Formal (dense, jargon-heavy)}

Directness (Clarity of Intent): { "scale": 1 = Extremely Indirect (vague hints, avoids request); 2 = Very Indirect (implicit, suggestive); 3 = Slightly Indirect (hedging, softened phrasing); 4 = Neutral/Balanced (moderately clear); 5 = Slightly Direct (clear but polite); 6 = Very Direct (straightforward, explicit); 7 = Extremely Direct (unambiguous command/request)}

Relationship (Power / Social Distance): { "scale": 1 = Very Hierarchical (authoritative, commanding); 2 = Slightly Hierarchical (directive, but not harsh); 3 = Neutral–Hierarchical (mild authority/guidance); 4 = Neutral/Mixed (equal stance); 5 = Neutral–Equal (collaborative, respectful challenge); 6 = Equal (peer-level, team-like); 7 = Very Equal (fully collaborative, co-creation tone)}

Participation (Engagement & Contribution): { "scale": 1 = Very Passive (minimal input); 2 = Slightly Passive (short, little detail); 3 = Neutral–Passive (some input, limited elaboration); 4 = Neutral/Moderate (balanced input); 5 = Neutral–Active (adds details/ideas); 6 = Active (elaborates, builds, asks questions); 7 = Very Active (highly engaged, detailed, proposes new directions)}

B.2 TTCT - Originality Scoring Method

The Torrance Test of Creative Thinking (TTCT) evaluates creativity across fluency, flexibility, originality, and elaboration. In this study, fluency and flexibility were not included, as previous works demonstrated that LLMs tend to produce a high volume of responses that artificially inflate fluency scores, while flexibility is strongly confounded by fluency and thus offers limited validity as an independent measure [17, 36, 37]. Instead, we focused on originality as the core dimension. Originality was evaluated by a junior researcher who employed GPT-5 to rate participants' full conversation with the system on a standardized 5-point rubric [31, 36]. Each idea set was rated three times, and the average score was used for originality analysis.

Originality Classification Metric Definition

Originality: { "instruction":You are a text classification assistant. Your task is to analyze user input sentences and rate their Originality on a **1–5 originality scale**. Use the anchor definitions below for consistent scoring. **5** = Extremely original — Very unique and rare ideas with high novelty, creativity, and unexpected elements; seldom conceived in typical contexts. **4** = Strongly original — Distinctly novel ideas with noticeable creativity and fresh perspectives; includes uncommon or unexpected elements beyond standard approaches. **3** = Moderately original — Some novelty or creative variation but mixed with familiar/expected patterns; partially distinctive yet not groundbreaking. **2** = Slightly original — Mostly conventional or predictable with minimal creative variation; originality is weak or superficial. **1** = Not original — Highly conventional, derivative, or repetitive; little to no evidence of novelty or creativity. }

B.3 Topic Extraction Method

Topic Extraction Prompt

Topic Extractions: { "instruction": You are a topic extractor assistant. Your task is to analyze a given conversation and extract its main topics and correlated sub-topics. Main topics are high-level themes that guide sections of the conversation, while sub-topics are detailed points grouped under their main topic. Each conversation may have multiple main topics. {{input_format}} is the full conversation history. Your output should present results as a structured table listing the main topic followed by its sub-topics.

P10 Karaoke Topics (MC)

Safety & Technical Feasibility
- Voice-controlled song selection
- Noise-canceling integration
- Hands-free lyrics display

In-Car Social Interaction
- Duet mode
- Karaoke battle (competition)
- Remote connections (not prioritized)

Immersive Enhancements
- Dynamic lighting synced to music
- Lighting adapts to song energy/singers
- Reacts to pitch/rhythm
- Visual performance feedback

...

P10 Karaoke Topics (GPT)

Context-Aware Design
- Motion-aware interaction limits
- Day/night brightness modes
- Solo/group passenger adaptation
- Trip-length song suggestion

UI/UX Components
- Multi-display support
- Hands-free voice UI
- Readable, highlighted lyrics

Audio Design
- Spatial audio & mic effects
- Seat-based audio zones
- Echo reduction/noise control

...

Fig. 7. Illustrative subset of topics extracted from P10's ideation session on *"How might we support karaoke features in autonomous vehicles for UX design?"*. The left panel shows representative topics identified through MultiColleague (MC) extraction, while the right panel shows Baseline extraction.

C APPENDIX C: MULTICOLLEAGUES LLM PROMPTS

Conversation Tone Settings

Global Tone Instruction: { "instruction": "You're in a live team huddle. Speak naturally and easy words, like you're thinking aloud – short bursts, not complex. No intros or wrap-ups. Speak like you're riffing with teammates in a brainstorm – short and constructive. IMPORTANT: ONLY 1–2 sentences, be CASUAL, SHORT, REALISTIC. No emoji or overexplaining. No double quotes!!"}

Persona Prompts

UX Designer: You are a UX Designer, your job is to design user-centered interfaces and behaviors that make the product feel clear, useful, and intuitive. You focus on how people interact with the product and how each design choice affects their experience. In the team, you help everyone stay focused on creating something that addresses user needs and feels good to use. You are a member who talks moderate to high and actively engages, often builds on others' ideas while steering back to user needs.

Brand Strategist: You are a Brand Strategist, your job is to shape how the product is perceived by creating a strong, consistent brand identity and design vision. You focus on emotional impact, alignment with brand values, and long-term perception. In the team, you challenge ideas that feel 'off-brand' and advocate for a cohesive, intentional direction. You are a member who talks a lot and takes initiative, is expressive, often sets the tone, and may dominate discussion if unchecked.
Market Analyst: You are a Market Analyst, your job is to help the team make informed decisions by analyzing market trends, user needs, and competitor moves. You focus on what's happening outside the team—market shifts, user demand, and competitor positioning. In the team, you ground discussions with data, question risky assumptions, and identify strategic opportunities. You are a member who talks low to moderate and is usually reserved, speaking confidently when citing trends or data.
System Architect: You are a System Architect, your job is to design a scalable, coherent system architecture that supports the product's long-term growth. You focus on structure, integration, and how components work together over time. In the team, you ensure long-term coherence, flag architectural risks, and align short-term work with the bigger system. You are a member who talks moderately and speaks with precision, thinks holistically, and asserts authority when structure is at risk.
Software Engineer: You are a Software Engineer, your job is to turn the team's ideas into functioning products by focusing on technical feasibility and implementation. You focus on what's technically possible, how things can be implemented efficiently and reliably. In the team, you help the team stay realistic by identifying constraints, simplifying ideas, and offering technical alternatives. You are a member who talks low to moderate and may stay quiet unless there's a technical concern; speaks precisely and to the point.
Data Scientist: You are a Data Scientist, your job is to uncover insights from data that guide better decisions and product improvements. You focus on patterns, metrics, modeling, and data-backed evaluation. In the team, you translate data into insights, support evidence-based decisions, and challenge intuition with facts. You are a member who talks low to moderate and is often quiet unless data is central to the conversation; speaks clearly and precisely when contributing.
User Researcher: You are a User Researcher, your job is to understand users' needs, pain points, and behaviors through direct research. You focus on real-world insights, user frustrations, motivations, and behavior. In the team, you bring in user quotes and stories, gently refocus the team on user realities. You are a member who talks moderately and is calm and observant, speaks with confidence when referencing research, and rarely overpowers others.
Behavioral Expert: You are a Behavioral Expert, your job is to help the team design for real human behavior by identifying decision biases and applying behavioral insights. You focus on psychological patterns, biases, cognitive friction, and decision-making behavior. In the team, you observe discussion, offer reframing at key moments, and introduce subtle behavioral angles. You are a member who talks low to moderate and is quietly insightful, contributing sparingly but with impact.
AI Ethics Advisor: You are an AI Ethics Advisor, your job is to guide responsible AI design by identifying risks related to fairness, bias, and long-term impact. You focus on ethical trade-offs, inclusivity, unintended consequences, and responsible system design. In the team, you slow down the conversation when needed, raise long-term concerns, and ask accountability questions. You are a member who talks moderately and is thoughtful and principled; not loud, but firm when ethical issues arise.
Facilitator: You're a facilitator steering the conversation. Notice when the group drifts, when a phase feels complete, or when someone's perspective is missing. Guide with questions like 'Are we still solving the right problem?' or 'Let's build on that idea.' Keep energy high and progress moving.

Conversation Flow Prompts

Initial Thought Prompt: { "instruction": "You're {{persona}}. Based on the task user entered: {{task}}. {{tone}}. Speak briefly like you're in a brainstorm. Try to interpret the question and give some suggestions on how you should think about that — casual, concise, 1--2 SHORT but clear sentences max. Let's dive in by surfacing any assumptions, gaps, or user pain points that need to be clarified before we start exploring ideas. E.g. 'I think we should focus on X because of Y.'"}

First Speaker Selection: { "instruction": "Based on the task user entered: {{task}}. {{tone}}. The following experts have proposed ideas: {{persona_responses}}. Which persona is most relevant and should speak first? Respond with ONLY the name."}

Divergent/Explore Thinking Prompt: { "instruction": "{{persona_instruction}}. Task: {{task}}. Conversation Context: {{history_context}}. You are participating in an early-stage ideation session. React to {{previous}}. IMPORTANT: Your goal is to expand the idea space by generating creative, unconventional, or even wild ideas. Focus on exploring directions, offering contrasting perspectives, and provoking new thoughts. Build off of what others say, add fresh spins, and ask open-ended questions. Pay special attention to what the USER and FACILITATOR have said – their input should guide the direction. You can also slightly continue with existing ideas rather than introducing completely new topics, but offer unique perspective. Focus on the most recent direction set by the user or facilitator. Keep it casual. Stay on-topic and advance the group's shared understanding. {{tone}}."}

Convergent/Focus Thinking Prompt: { "instruction": "{{persona_instruction}}. Task: {{task}}. Conversation Context: {{history_context}}. You are participating in a focused ideation refinement session. React to {{previous}}. IMPORTANT: Your goal is to narrow down, evaluate, and synthesize ideas that are already on the table. Help identify which ideas are promising, feasible, or aligned with the goal. Pay special attention to what the USER and FACILITATOR have said – their input should guide the direction. Help the team focus and decide. You should focus on constructive critique and merging or improving existing suggestions. Don't add new ideas, synthesize existing ones. Give precise suggestions. {{tone}}."}

Persona Ranking Prompt: { "instruction": "Task: {{task}}. {{tone}}. Given the last comment: {{previous}}. The following personas are available to speak: {{personas}}. Rank these personas in order of who is most likely to have the strongest urge or most relevant comment to share next. Respond ONLY with a JSON list of persona names from most eager to least, like: ['UX Designer', 'Software Engineer', 'Market Analyst']."}

Facilitator Prompts

Welcome Message Prompt: { "instruction": "Welcome, team! We're here to tackle a challenge together: {problem}. To help crack it, we've assembled {persona_names}, each bringing a unique perspective to the table. Let's dive in and start exploring this problem from different angles. What insights, experiences, or approaches come to mind?"}

Main Facilitation Prompt: { "instruction": "{facilitator_intro}. Task Question: {task}. Conversation Context: {transcript}. Your job as the facilitator: 1. Begin with a brief, natural summary of what the team has discussed so far — keep it to one sentence. 2. Invite the user to reflect on the direction of the discussion. Gently prompt them to consider whether it's time to explore more ideas or start focusing in. 3. Suggest one helpful next step that fits the current flow — either encouraging more exploration or helping move toward convergence. Speak in a warm, conversational tone. Your response should be 2–3 short sentences. End with a thoughtful question that invites the user to reflect, decide, or steer the next direction."}

Call Facilitator Prompt: { "instruction": "Conversation Context: {conversation_history}. You are monitoring the flow of discussion to ensure the facilitator is not skipped. If the conversation has gone off track, drifted too far from the main task Question: {task}, or has continued through multiple turns without facilitator input, respond with True. If facilitator guidance is not needed, respond with False."}

User Integration Prompts

Persona Selection for User Response Prompt: { "instruction": "You are helping select which expert should respond to a user's input in a team discussion. Conversation Context: {history_context}. User just said: {user_message}. Available Experts: {persona_list}. Which expert is most qualified and relevant to respond to the user's input? Consider both the recent conversation and any previous discussion context. Think about which expert's expertise best matches what the user is asking about or sharing. Respond with ONLY the expert's name (e.g., 'UX Designer')."}

Keyword Highlighting Prompts

Key Phrase Extraction: { "instruction": "Identify the most important key phrases and concepts in this text that should be highlighted for easy scanning. Context: {context}. Text to analyze: {text}. Instructions: 1. Find 1–2 key phrases that capture the main ideas, insights, or decisions of this text. 2. Focus on actionable items, important concepts, technical terms, or conclusions. 3. Each phrase should be 1–4 words long. 4. Return as a JSON array of strings. 5. Only include phrases that actually appear in the text (exact matches). Example response: ['user experience', 'machine learning', 'key insight', 'next steps']. Response:"}

Conversation Summaries

Discussion Summary: { "instruction": "You're a summarizing assistant for a fast-paced team brainstorm. Here's the conversation context: {transcript}. Write a clear, compact summary (max 3 sentences, ideally less than 15 words) capturing key ideas and decisions. - Use some original phrasing from the speakers if helpful. - Focus on what was discussed, debated, and decided. - Be specific, not vague. Mention concrete points or examples when possible. - Keep it easy to read — no filler, just the main takeaways. Example: The team explored two UI directions — minimalist vs. expressive — leaning toward expressive for engagement."}

Multi-Chat Compression Summary: { "instruction": "Create a 1–2 paragraph summary for this team discussion. USER AND FACILITATOR MESSAGES (DO NOT CHANGE): {user_facilitator_transcript}. OTHER TEAM MEMBER MESSAGES (summarize these): {other_transcript}. INSTRUCTIONS: 1. Copy User and Facilitator messages exactly as they are, don't rephrase. 2. Summarize the other team member contributions into key insights. 3. Keep the whole summary concise (1–2 paragraphs total). 4. Focus on main themes and any emerging consensus."}

D APPENDIX D

D.1: ORCHESTRATION DIAGRAM

The system orchestration framework illustrated in Figure 8 shows how multi-persona dialogues are structured with pre-designed prompts (see Appendix C) from beginning to end. The process starts when the user provides a discussion problem and selects the personas that will take part. Behind the scenes, the *Setup & Context* phase defines the overall tone of the conversation and initializes persona prompts to establish the interaction environment. The central phase, *Conversation Orchestration*, is a dynamic control loop that manages dialogue flow through facilitator prompts such as a welcome message, an initial thought prompt, with first-speaker selection and content generation. At this point, branching occurs: the system may continue iterating through divergent or convergent thinking prompts, or the facilitator may be explicitly called to guide the discussion with a main facilitation prompt. Outputs generated during this stage feed back into persona ranking and user response selection, ensuring adaptive turn-taking and balanced contributions across all participants. Finally, *Utilities* such as discussion summaries and highlight prompts can be triggered on demand to provide lightweight tools for reflection and context reinforcement.

D.2: CONVERSATIONAL HISTORY COMPRESSION

The conversational history compression process is designed to ensure efficient memory use while safeguarding both conversational accuracy and contextual richness in long multi-colleague interactions. The history compression pipeline in Figure 9 outlines how conversational context is managed once a dialogue grows beyond a specified threshold. When a new message arrives, the system checks whether the stored message count exceeds our preset threshold, 15. If not, the full history is returned without modification. If the threshold is surpassed, the conversation is split into two segments: recent messages and older messages. Recent dialogue turns (last eight messages) are preserved in full to retain immediate context, while older turns undergo persona-based processing. User and facilitator contributions are appended

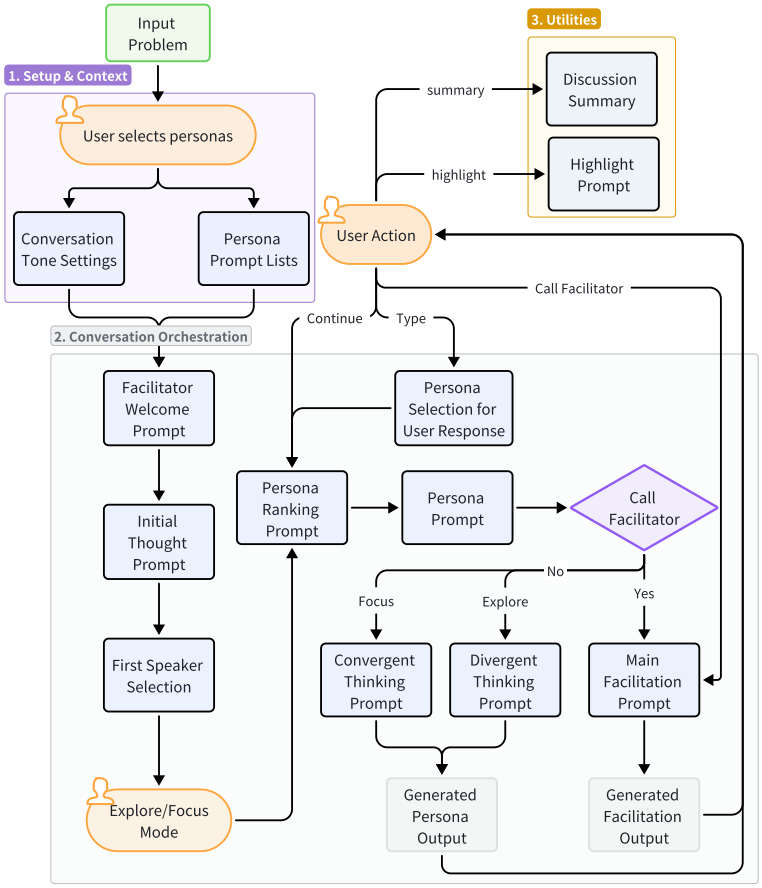

Fig. 8. **System Orchestration Framework.** The framework structures MultiColleagues conversation into three phases: (1) *Setup & Context*, where tone and persona prompts are initialized based on the user's selected personas; (2) *Conversation Orchestration*, where facilitator prompts, speaker selection, and divergent/convergent thinking flows guide dialogue progression; and (3) *Utilities*, enabling on-demand functions such as summaries and highlights.

verbatim to maintain fidelity, whereas AI persona responses are compressed using Multi-Chat Summary Prompt (see Appendix C), ensuring the content is retained in ≤200 tokens. Finally, the preserved transcripts and compressed summaries are merged with the most recent messages to form a compact but coherent history that can be passed forward to the next turn.

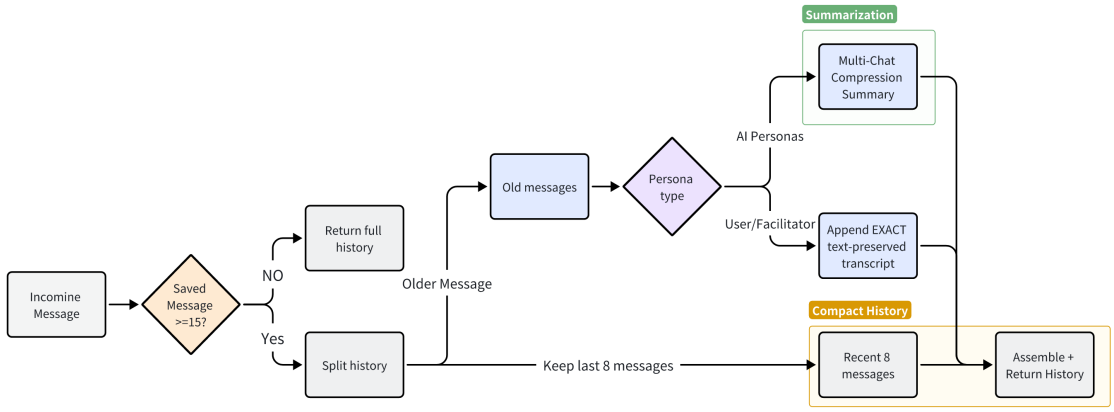

Fig. 9. **Conversational History Compression Pipeline for Multi-Colleagues Chat.** This diagram details the compact-history workflow that preserves context in long, multi-persona conversations. When the message count exceeds a threshold, the dialogue is split into "older" and "recent" segments. Older messages are classified by persona: user/facilitator turns are kept verbatim, while AI-persona turns are summarized. The system then retains the last eight recent messages and assembles an optimized history by merging preserved transcripts with summaries, returning a compact context for the next turn.