Perspectra: Choosing Your Experts Enhances Critical Thinking in Multi-Agent Research Ideation

Yiren Liu
Informatics
University of Illinois Urbana-Champaign
Champaign, Illinois, USA
yirenl2@illinois.edu

Viraj Nischal Shah
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
Champaign, Illinois, USA
virajns2@illinois.edu

Sangho Suh
Allen Institute for AI
Seattle, Washington, USA
sanghos@allenai.org

Pao Siangliulue
Allen Institute for AI
Seattle, Washington, USA
paos@allenai.org

Tal August
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
taugust@illinois.edu

Yun Huang
School of Information Sciences
University of Illinois Urbana-Champaign
Champaign, Illinois, USA
yunhuang@illinois.edu

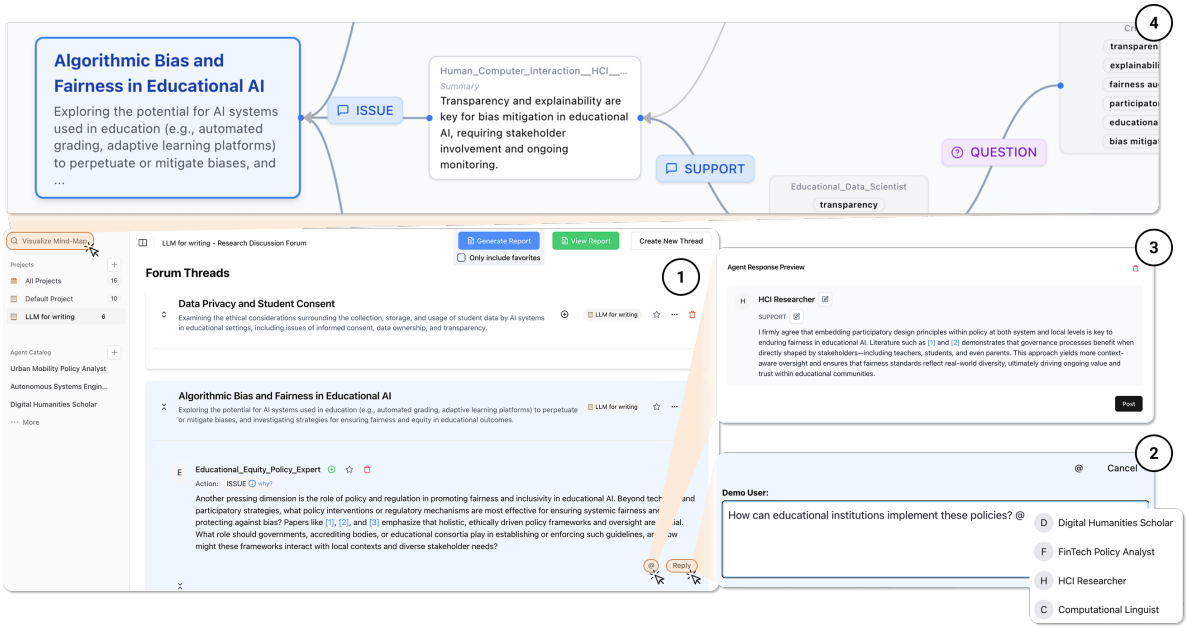


Figure 1: Perspectra interface. The system supports the following features: ① Upon login, the main dashboard presents a forum-style interface that enables reply threading and structured multi-agent deliberation. ② Clicking the reply button allows users to respond to a post in the forum and tag additional agents into a single thread discussion. ③ Clicking the button next to "Reply" activates the what-if feature, allowing users to explore hypothetical responses from agents. ④ A mindmap feature provides a visualization of agents' posts, replies, and deliberation actions with rationales, including semantic zooming to support sensemaking.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791560

Abstract

Recent advances in multi-agent systems (MAS) enable tools for information search and ideation by assigning personas to agents. However, how users can effectively control, steer, and critically evaluate collaboration among multiple domain-expert agents remains underexplored. We present Perspectra, an interactive MAS

that visualizes and structures deliberation among LLM agents via a forum-style interface, supporting @-mention to invite targeted agents, threading for parallel exploration, with a real-time mind map for visualizing arguments and rationales. In a within-subjects study with 18 participants, we compared Perspectra to a group-chat baseline as they developed research proposals. Our findings show that Perspectra significantly increased the frequency and depth of critical-thinking behaviors, elicited more interdisciplinary replies, and led to more frequent proposal revisions than the group chat condition. We discuss implications for designing multi-agent tools that scaffold critical thinking by supporting user control over multi-agent adversarial discourse.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; *Collaborative and social computing systems and tools*; **Visualization**; **Empirical studies in interaction design**; • **Computing methodologies** → **Multi-agent systems**.

Keywords

Multi-agent systems, Human–AI collaboration, User control, Sensemaking, Visualization, Critical thinking, Argumentation

ACM Reference Format:

Yiren Liu, Viraj Nischal Shah, Sangho Suh, Pao Siangliulue, Tal August, and Yun Huang. 2026. Perspectra: Choosing Your Experts Enhances Critical Thinking in Multi-Agent Research Ideation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 27 pages. https://doi.org/10.1145/3772318.3791560

1 Introduction

LLM-based Multi-Agent Systems (MAS) [22] are increasingly being adopted across application domains because of their ability to perform more complex tasks than single-agent pipelines through collaboration among multiple agents [2, 6, 110]. Such dynamics in MAS have the potential to benefit both observers and overall task completion. This is similar to human collaboration, where dialogue among human experts promotes critical thinking in interdisciplinary learners [28, 48, 56, 65, 99] and early-stage researchers [28, 99]. Despite this potential benefit, prior research has largely focused on designing effective communication with a single AI agent [26, 111] for complex knowledge work, leaving underexplored the opportunities and challenges of how users perceive and control multiple agents in collaborative settings.

Challenges in designing interactive MAS generally involve three issues: 1) the cognitive burden required for users to select and coordinate agents [96, 112] (which agents to call); 2) information overload over parallel generation [107]; and 3) difficulty interpreting agent actions and rationales [80, 108]. In this paper, we ask: *how to better support users to control multi-agent collaborations?* Specifically, we choose to explore this question in the context of how enhanced control can improve ideation quality and critical thinking when conducting interdisciplinary literature reviews. This is because interdisciplinary research fundamentally requires researchers to reconcile differing assumptions, methods, and disciplinary norms [46]. Multi-agent systems with agents representing unique perspectives from different expert domains have been shown to be effective at enabling cross-field knowledge integration, increasing the richness of the interaction process, and stimulating critical thinking [45, 106]. However, these systems did not allow users to control multi-agent deliberation on different topics simultaneously.

To better understand the support users desire in such collaborations, we first conducted two rounds of pilot studies that uncovered key challenges perceived by users during ideation in interdisciplinary research contexts: (a) a lack of user control to steer the discourse among agents and probe emergent sub-topics, and (b) difficulty comprehending the evolving structure of multi-agent discussions, which increases cognitive load and impedes sensemaking.

We then designed and implemented Perspectra, which provides two major design components to support user control and sensemaking of multi-agent deliberation: 1) @-mention and *reply* that allow users to participate in and invite user-chosen agents into an ad-hoc conversation, meanwhile revealing the reasoning processes of domain-expert agents during a group discussion through interactive visualizations; and 2) thread branching that facilitates parallel exploration of multiple topics while allowing users to easily blend and remix outputs from agents with different backgrounds. These designs are inspired by traditional forum interfaces, which are commonly used to gather feedback from individuals with interdisciplinary expertise and perspectives (e.g., ResearchGate [89, 90, 105] and Academia.edu [67]). We extend this approach by allowing users to dynamically control which personas to engage with on the fly. In order to reduce cognitive load and improve sensemaking across threads, Perspectra also provides visualization of agent deliberative actions and rationales through a dynamic mind map feature that assists navigation. To evaluate the effectiveness of Perspectra design, we also implemented a group chat interface as the baseline condition. The group-chat design allows the user to chat with agents playing multiple personas, in one session, without being able to create different threads or (de-)select personas for sub-topics. This baseline approximates mainstream usage of systems such as ChatGPT, Claude, Grok, and Gemini, in which users can request persona shifts within a single chat session.

We conducted a within-subjects study with 18 participants, who were asked to use both Perspectra and the baseline group-chat design to develop a brief proposal on an interdisciplinary research topic. We analyzed system log, participants' survey scores, think-aloud data, and interview feedback. We compared the two designs in terms of proposal quality, required revision effort, observed interaction patterns, and the prevalence of critical-thinking activities.

Our findings make novel and timely contributions to the HCI community:

- **Perspectra, a new system design.** Support users to select personas and steer multi-agent deliberation. Features such as @-mention and thread branching allow users to create ad-hoc panels for tackling unfamiliar topics, while visualization of discussion structure, such as ISSUE, CLAIM, SUPPORT, REBUT, and QUESTION, assists their sensemaking.
- **Empirical evidence of enhanced critical thinking.** The experimental results show that Perspectra significantly promotes critical-thinking activities, e.g., *Application*, *Analysis*,

Inference, and *Evaluation*, compared to a group-chat baseline. Participants initiated panel-like discussions to engage cross-disciplinary perspectives, examine assumptions, and refine interpretations, activities essential for higher-order reasoning.

- **Measurable improvements in proposal revisions and quality.** Participants using Perspectra revised their proposals more frequently and achieved greater improvements in proposal quality compared to when using the group-chat condition. This finding suggests Perspectra's practical impact, potentially going beyond interaction patterns to influence real work outcomes.
- **Discovery of designed and emergent affordances.** We observed both expected and emergent user behaviors. Designed affordances, such as @-mention driven deep dives with both a single agent and multi-agent sensemaking via the mind map, were used as intended. Beyond these design goals, participants created emergent practices such as leaving TODO-anchors and performing verification checks, showing the system's flexibility, and new design opportunities for user-driven adoption.
- **Design implications for future systems.** Our findings offer concrete design implications for knowledge-intensive ideation tools. We argue for enabling *adversarial or dissenting agent responses* to foster critical reflection and for developing *hybrid interaction models* that balance user control with agent autonomy, preventing cognitive overload while still supporting rich deliberation.

2 Related Work

2.1 Multi-Agent Systems for Research and Ideation Feedback Solicitation

Many recent studies have considered the use of LLM-based agents to be an effective method for research ideation [2, 18, 106]. Recent advancements in LLM-related research have explored how multi-agent systems can be applied to facilitate ideation in various domains and gather feedback from diverse perspectives [62, 109]. The architectural foundation of multi-LLM ideation systems centers on role specialization and coordinated collaboration between distinct agents, which decomposes complex tasks into subtasks handled by specialized agents (e.g., "Scientist," "Critic") to improve accuracy, completeness, and idea diversity [20, 45, 47, 49, 92]. Coordination strategies in these systems range from user-orchestrated frameworks that prioritize user control [72] to automated approaches where agents self-organize to solve problems [20]. To manage the complexity of these interactions and reduce cognitive load, recent work has focused on visual coordination tools and structured interfaces. These tools help users design and explore collaboration strategies visually [70] and shift from reactive dialogues to more proactive, structured interactions with the multi-agent system [34, 52, 74, 102]. There also has been work on using conversational agents to engage in community discussions [81]. Li et al. [41] proposed a multi-agent approach to simulate a society of LLM agents by allowing them to communicate with each other, where the dialogue and interactions can later be used for understanding agents' behavior and reasoning processes.

While prior work has focused on the architecture and coordination of multi-agent systems, less discussion has centered on how to offer fine-grained user control to selectively compose subsets of agents for emergent sub-topics, as interaction is often broadcast to all agents or fully automated in current MAS implementations [20, 72]. We address these gaps with an interaction design that combines a forum-style interface with user control over agent selections through @-mentions for ad-hoc panel formation and a visualization of deliberative moves (e.g., claim, support, question) to surface stance relations and reasoning. Our design also reframes multi-agent output from a flat message stream into a navigable argument structure for scaffolding users' ideation processes across parallel topics.

2.2 Balancing Learning and Cognitive Load: Collective Discourse and Distributed Cognition

Recent studies leverage adversarial stances in conversational agents to provoke counter-arguments and promote critical thinking in groups. For design ideation, such agents help reduce design fixation by actively challenging dominant proposals [39]. In group decision-making contexts, devil's-advocate agents amplify minority voices to counter conformity pressure [40] and mitigate social influence to improve deliberative quality [38]. Additionally, research has shown that structured deliberation among experts can be valuable for interdisciplinary learners [28, 99]. Similar discourse can also be found in research related to inquiry-based learning [48, 56] and in the context of argumentation-centered collaborative peer learning [65]. A complementary perspective from cognitive neuroscience suggests *Cognitive Synergy* as a driver of complex human cognition [50]. This suggests that systems should foster explicit combination of diverse and complementary perspectives. Recent research has also heavily discussed applications of LLM agents in performing knowledge-extensive information retrieval tasks, with a well-established application example of deep research agents [26, 111]. This methodological paradigm emphasizes the use of LLM agents, often multiple and parallelized, to assist users in exploring and synthesizing large volumes of information into report. However, a challenge remains in how to 1) effectively grant users agency and control over the search process, and 2) how to present the information in a way that is easy for users to understand and evaluate. Without a clear view of how agents reason, debate, and build upon each other's ideas, users may struggle to critically evaluate the generated feedback.

Distributed Cognition theory [27] has been applied in a wide range of system designs to support sensemaking [1, 73] through the breakdown of complex cognitive tasks into subtasks. We build on these insights by applying guidance of explicit argument structures in multi-agent deliberation to scaffold users' critical evaluation during ideation, by providing visualization of agents' deliberation acts in a LLM-based ideation system to reduce cognitive load and nudge active reasoning instead of passive consumption [70, 74].

2.3 Critical thinking activities in interdisciplinary learning

Critical thinking skills and activities have been widely studied in educational psychology and pedagogy, particularly in the context

of interdisciplinary learning [14]. We consider these categories as higher-order critical thinking activities because they require learners to move beyond recall and literal comprehension to generate, transfer, and judge knowledge. Bloom's taxonomy [3] has been widely used to classify different types of cognitive activities, including *knowledge*, *comprehension*, *application*, *analysis*, *synthesis*, and *evaluation*. In the cognitive domain, *Application* entails using concepts and procedures in novel contexts, and *Evaluation* requires making warranted judgments against criteria [3, 100]. Contemporary critical-thinking accounts likewise identify *inference* and *evaluation* as core operations of expert judgment, which draws justified conclusions from evidence and assessing the credibility and quality of claims [13]. Although the taxonomy lists "inferring" under understanding, the operation involves integrating prior knowledge with incomplete information to construct meaning not explicitly stated, which aligns with higher-order reasoning in practice. Empirical assessment work also groups tasks requiring *application/analysis/evaluation* as higher-order because they demand integration and transfer to unfamiliar problems [29].

Critical thinking is also essential for developing scientific literacy and fostering learning [14]. Recent research has revealed the risk of the use of GenAI technology in knowledge work reducing critical thinking skills [36]. Research has begun investigating how to design GenAI systems that can support critical thinking activities [45]. In this work, we explore designs that nudge participants to engage in critical thinking activities. More specifically, we design agents' deliberation actions and the overall interaction model to promote active reasoning from users, targeting the forms of reasoning closely associated with critical thinking, rather than mere recall and comprehension.

3 Methods

3.1 Iterative Design through Two Rounds of Pilot Studies

To address the gap of exploration across multiple perspectives in a deliberation setting, we introduce Perspectra. The design of Perspectra is informed by an iterative design process that involved two rounds of prototyping and user feedback, as shown in fig. 2. We conducted two rounds of pilot studies with a total of 8 participants, including researchers from various disciplines. We first introduce an initial design of Perspectra, drawing inspiration from online forum discussion layouts and interaction designs, which have been shown to facilitate collaborative ideation [90] by supporting context tracking and sensemaking in forum-based online discussions [105].

During the first round of pilot, we presented a low-fidelity prototype to users and gathered their feedback and thoughts on the design. The prototype featured a forum-style interaction design where users could engage in threaded discussions with multiple expert agents by replying to a post or reply. We found that users generally expressed interest in the online forum-based interaction. However, they identified several pain points, mostly related to information overload, difficulty in tracking discourse context, and challenges in accurately understanding agents' rationales behind responses. Participants suggested several improvements, including: 1) a feature to allow users to more easily explore and verify the literature sources and rationales each agents referenced during discussions, for both purposes of transparency and sensemaking of agents' background (*"You can also add the list of papers mentioned... For me, finding good relevant papers is really hard."* — I1); 2) a mechanism to support easier navigation of discussion context and structure (*"I kind of feel like the interaction between several people in the conversation might be a little bit... complex."* — I3).

Based on the feedback, we further refined the prototype and conducted another round of pilot study with a different group of participants (N=4). The second version of the prototype implemented improvements based on the feedback from participants during the first round, including 1) a mind map visualization to help users track the discussion context and structure; and 2) a panel that displays detailed agent profiles. Our second round of pilot studies revealed several key areas for improvement. First, participants (I5, I6, I8) noted that the agents often lacked awareness of user intent during interactions, making it difficult to steer conversations in desired directions. Second, participants (specifically I5) expressed the need for clearer definitions and improved visibility of deliberation mechanics, as they sometimes struggled to understand why agents were making certain argumentative moves. Finally, multiple participants (I7, I8) requested mechanisms to more easily follow up with individual agents during parallel and multi-threaded discussions, suggesting that more direct engagement options with specific personas would enhance the deliberation experience. These insights guided our final system refinements before the formal user study. The two rounds of pilot studies informed two major design goals for the system:

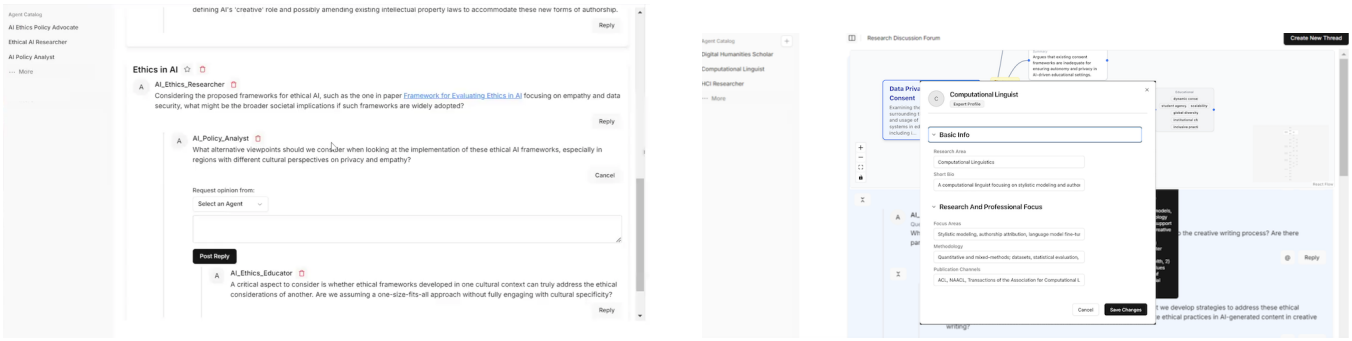
- **DG1 (Choosing Expert Personas to Steer Forum Discourses)**: Enable users to dynamically involve and steer the discourse between multiple agents to deepen dialogue around emerging topics.
- **DG2 (Structuring Dialogue for Effective Comprehension)**: When there are multiple agents involved in discuss, the interface should facilitate users' sensemaking of evolving discussion and its dynamics by reducing cognitive load, e.g., through formalization and visualization of discourse actions.

3.2 System Design

We propose Perspectra, an interactive system designed to enhance critical thinking through structured multi-agent deliberation and visualization of discussion dynamics. In the section, we describe the design of Perspectra's key components that address the design goals outlined in section 3.1. We open-source the code of the Perspectra prototype[1].

*3.2.1 Agent Interaction and Perspective Exploration (**DG1**).* Perspectra allows users to interact with multiple expert agents through a threaded interface as if in a panel discussion setting involving multiple researchers with their own domain background knowledge. The interaction follows a two-phase design: first, the system creates the initial threads based on user-chosen topics, and populates a small number of initial agent responses so that the user can quickly browse through the initial points of discussions; then, the user can manually mediate the discussion by invoking agents through reply and @-mention. During the second phase, users can

[1] https://github.com/yiren-liu/perspectra-multi-agent-research-ideation



(a) Initial low-fidelity prototype (Pilot Round 1) with basic threaded forum interaction.

(b) Refined prototype (Pilot Round 2) with mind map and agent profiles

Figure 2: Iterative prototype evolution of PERSPECTRA: (a) the initial forum-style design used to elicit early feedback; (b) the refined design incorporating participant suggestions to reduce information overload, surface agent rationale sources, and improve navigation of discussion context.

engage in interaction with agents, each with a specific background profile, through several mechanisms designed to facilitate discourse, including:

(1) **Free-text replies**: As shown in fig. 3, users can reply to any agent's post, with @-mentions to involve specific agents in the conversation. Once a user's reply is posted, all mentioned agents will respond to the user's reply. If no additional agents are tagged, the agent that made the target post will respond. The design of this feature requires users to explicitly select which agents to interact with, specifically encouraging them to evaluate and critically reflect on agent responses.
(2) **Action requests**: The system also provides a quick action for users to select a particular agent and take a specific stance when generating the response (i.e., agree, disagree, question). Similar to the free-text reply feature described above, this design requires users to manually invoke specific agents, which is designed for fostering critical thinking by requiring users to actively decide which perspectives or counter-arguments to explore, rather than passively reading agent responses in bulk. We chose to adopt a simplified list of stances instead of the full deliberation actions (as in section 3.3.2) to reduce cognitive load for users. Once the user clicks the action, the system displays a "What-if" perspective panel to help users explore alternative viewpoints on the same topic.
(3) **Thread branching**: Users can also choose to create new discussion threads based on specific responses for deeper exploration of topics that emerge during the discussion. This allows users to steer the conversation into new directions without being limited to the pre-generated topic. Note that this feature allows users to create a new thread and populate the initial discussions similar to the thread initialization process, without having to invoke individual agents using the reply and @-mention features.

This interaction design facilitates users' exposure to diverse perspectives from different disciplinary backgrounds. PERSPECTRA further provides features to enable transparency and customization of the agent personas. For example, users can access and edit detailed profiles for each persona, which are initially derived from the Persona Hub dataset [19]. This feature allows users to tailor agent backgrounds to better match their own research contexts or explore specific types of disciplinary perspectives. More details about the implementation of agents and personas can be found in section 3.3.1. Each agent simulates a researcher with a particular domain background, and the agents have access to their own collections of literature and unique ways of thinking. More specifically, each agent persona is presented by:

(1) **Profile information**, which details expertise, background, familiar methodology of research, etc., accessible by clicking on an agent's name on the left panel on the interface.
(2) **Literature collection**, which is accessible by clicking on an agent's avatar, revealing the sources of knowledge informing the agent's contributions.
(3) **Agent memory**, which is presented through an alternative tab within the same panel as the literature collection that displays the agent's internal memory state, enhancing transparency about how the agent forms and maintains its perspective. The memory viewer provides two forms of visualization: 1) a chronological stream, and 2) a collapsible lineage tree to support user sensemaking of why an agent now argues a certain way. This design aims to increase transparency (users can inspect "why this shift occurred"), which supports **DG1** by externalizing intermediate reasoning states, while avoiding overwhelming users with sensemaking of agent intentions through responses.

Screenshots of the agent profile editor and memory viewer are provided in Appendix 15. These transparency features are designed to help users understand why agents offer particular perspectives and how their backgrounds led to their responses, better supporting users' sensemaking of agents' messages.

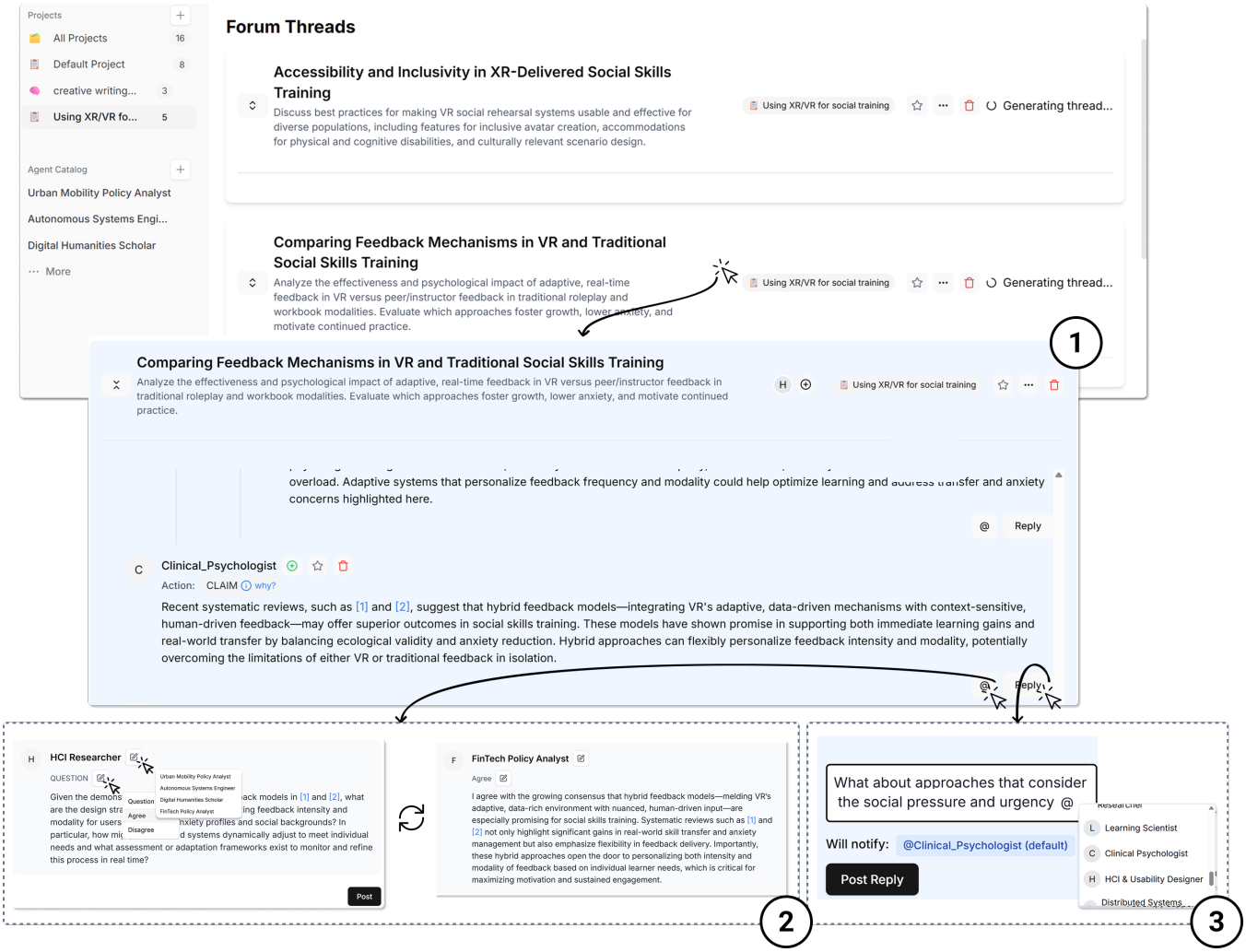


Figure 3: Perspectra system interface utilizes forum-style layout to support unique turn-taking deliberation interactions between agents, and visualization of threaded parallel topics. The interface supports interactions including: ① clicking on a discussion thread to collapse/expand the thread to hide/show detailed replies; ② clicking on the @ button under each reply allows generation of a "what-if" panel showing hypothetical replies from a chosen agent and stance combination, the user can also change the agent or stance with the panel to re-generate the reply; ③ clicking on the reply button spawns a text input box in which the user can type their reply, and the user can also type @ to select from a list of available agents to be included in follow-up discussion.

*3.2.2 Threaded Forum and Mind Map for Multi-Expert Deliberation (**DG2**).* The core component of Perspectra is an interface that facilitates deliberation between LLM-based agents through a design that emulates an interactive online discussion forum. This design allows users to engage in multi-agent discussions, where each agent represents a different disciplinary perspective, as shown in fig. 3. The system is structured around a threaded forum format, which organizes discussions around different topics into threads. Each thread is dedicated to a specific sub-topic, allowing users to focus on one aspect of the discussion at a time while providing the flexibility to switch between parallel threads. The system also provides a project tab that allows users to create and manage threads around different individual research projects or ideas. Each project page can host multiple discussion threads. The system generates initial suggestions of thread titles and descriptions based on the user's initial research proposal input, which can be edited and confirmed by the user before starting the forum discussion. Grounded in Distributed Cognition theory [27], our threaded forum design aims to assist users in decomposing complex topics into focused discussions to reduce cognitive load and enhance ideation, as supported by prior work [32, 51, 88]. This structure aims to help users explore specific sub-topics in depth while maintaining awareness of the broader discussion context.

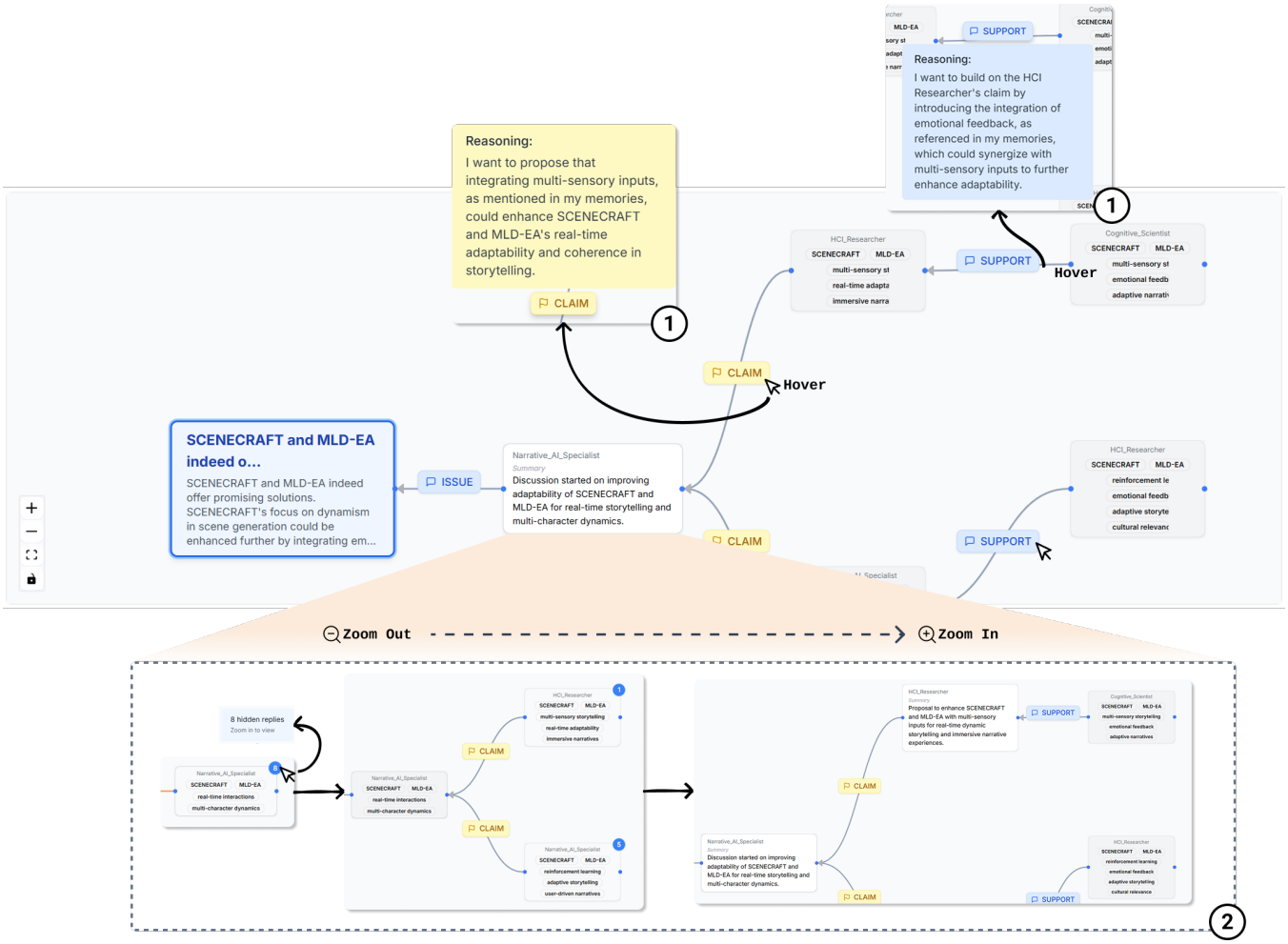


Figure 4: An interface of the mind map feature. Nodes represent posts/threads/replies with different levels of detail with semantic zooming, and edges encode agents' deliberation acts (details can be found in 3.3.2). ① Hovering an action chip or its connecting edge surfaces an inline rationale card with the posting agent's reasoning. ② Semantic zoom shifts between the keyword view and the summary view for each node.

In order to offer a more organized and coherent visualization for deliberation dynamics, Perspectra also includes an interactive mind map visualization that represents the discussion in a graph-based layout, as shown in fig. 4. The nodes display thread/post-level information, while edges convey actions and rationale from each participating agent. This visualization provides an alternative, spatial representation of the deliberation structure.

The mind map implements semantic zooming capabilities as used by previous sensemaking support research [86], allowing users to adjust the level of detail displayed as they zoom in or out of the visualization. At a high level, users can see main argument structures and key points; zooming in reveals more detailed summaries of the original posts/replies. This feature is designed to directly support different levels of ideation and navigation needs. While the mind map offers a good overview of the deliberation's structure, we designed it as a supplementary view to the main forum interface. This is because we discovered during the second pilot study that a dense mind map can become hard to navigate for reading and authoring detailed responses compared to a nested, collapsible, threaded interface. The mind map also incorporates context tracking, helping users to track back to the original content (post/thread) whenever a node is clicked on. The mind map visualization is designed to support users in more effectively processing complex information across disciplines and topics, while still maintaining awareness of interconnections through the system's visualization components.

3.3 Backend Implementation

3.3.1 Multi-Agent System for Research Ideation. We implement a multi-agent system that enables users to interact with multiple LLM-driven research personas implemented using AutoGen [101]

framework. Each agent operates over shared tools and a graph-based retrieval database to support cross-perspective ideation while maintaining independent goals and memory.

Agent components. Each agent consists of three major components that guide its behavior: 1) **Persona profile**: Persona profiles are populated from a structured taxonomy derived from Ge et al. [19]'s work (e.g., discipline, methodological stance, research role, epistemic orientation, focus areas, methodology, publication channels, skills, and communication style) using agglomerative clustering; The full taxonomy can be found in appendix A.6. Persona prompts also include a brief background narrative to encourage coherent and diverse reasoning from each agent's own perspective; 2) **Agent memory**: Beyond short-term memory (i.e., conversational history maintained by AutoGen), agents also have long-term cross-turn memory that uses a persistent store (implemented with a modified version of LangMem [2]) store; Each agent periodically distills interaction context (its own prior posts, user prompts, retrieved literature summaries, and other agents' challenges) into compact research idea snippets. Each snippet captures a single evolving hypothesis, question, rationale shift, or methodological consideration, which can also reference earlier snippets it refines or extends. The memories are later referenced during each time of inference; 3) **Literature database**: A GraphRAG database (implemented using a modified version of LightRAG [23] to support citation tracing) constructs and queries a knowledge graph over both entities and text snippets extracted from papers. The system also implements a paper search tool that queries the Semantic Scholar (S2) API [35] and OpenAlex [77]. Retrieved papers are incrementally inserted into the GraphRAG database.

Agent reasoning and tool use. Agents follow a ReAct-style reasoning loop [104] integrated with Autogen's tool-calling capabilities. In each turn, an agent first generates a *plan* based on the current dialogue context and its persona, deciding whether to respond directly or use a tool. If a tool is needed, it executes an *action*, such as querying the GraphRAG database for existing knowledge, searching external literature via the Semantic Scholar and OpenAlex APIs, or adding a newly found paper to its knowledge base. The raw output from the tool is then summarized and used to augment the agent's context. The agent then *reflects* on this new information to revise its initial plan if necessary. For instance, if a literature search fails to support a planned argument, the agent might pivot to a different claim and conduct another round of search. Finally, the agent generates its *response* based on the augmented context. To maintain a clean and interpretable deliberation history for the user, only the final agent rationale and a summary of the tool call are persisted, rather than the entire chain-of-thought process.

3.3.2 Designing Agent Deliberation Framework. We design the agents' action space by adopting several existing argumentation frameworks. Specifically, Prakken [76] on argumentation speech acts, Toulmin's model of argument structure [91], Walton & Krabbe's dialogue types and commitment rules [93], and the Agent Dialogue Framework (ADF) [55]. The goal of the design is to maintain expressive power for multi-agent deliberation while keeping the interaction readable to human readers' interpretation of agents' rationales.

[2]https://github.com/langchain-ai/langmem

Action	Definition	Theoretical Adaptation
ISSUE	Introduce a new question, sub-topic, or decision point.	Agenda setting in deliberation dialogue types [93]; sub-dialogue launch schemas in ADF [55]
CLAIM	State a position that the speaker commits to defend.	Assert, claim locutions and commitment updates [76, 93]; *Claim* from Toulmin [91]
SUPPORT	Provide explicit support with argumentative content.	Toulmin *Grounds*, *Backing*, and *Warrant* [91]; supplying reasons to "why?" challenges [76]
REBUT	Provide a counter-argument that attacks a prior claim or support.	Toulmin *Rebuttal* [91]; attack and defeat moves [76] Premise/inference attacks collapsed for legibility
QUESTION	Ask for justification or clarification ("Why?") about a claim.	Challenge, "why" locutions and burden transfer [76, 93]

Table 1: The design of action space for the agent deliberation driven by existing argumentation frameworks.

We adopt a deliberation framework as agents' action space, following the typology of Walton and Krabbe [93] and the composition principles in ADF [55]. Individual turns are guided using a small set of *locutions* adapted from Prakken [76] (e.g., assert, challenge, concede, retract) and Toulmin [91] (Claim, supporting Grounds/Backing, potential Rebuttals). In the UI, these backend *locutions* are displayed as labels (e.g., Claim, Support), the root threads always start with Issue posts (which open sub-threads/replies), and the mind map panel displays nodes (posts/replies) with edges labeled according to the deliberation acts (i.e., *locutions*) that lead to each node, as a visualization of the structure of the deliberation. The action set and theoretical origin are shown as in table 1. We note that this is an alternative approach for triggering agent responses compared to prior multi-agent systems that rely on numerical measures such as confidence score [21]. We deliberately chose to represent agent positions through these structured argumentation acts in order to prioritize the content of the deliberation (why each agent posts the response based on their rationale) over manually defined metric(s). This design encourages users to actively evaluate agent arguments and decide which experts to involve and interact with.

4 User Study

We designed a within-subject user study to evaluate the effectiveness and user experience using the Perspectra. The user study focuses on validating the effectivess of Perspectra and understanding how users engage with Perspectra during ideation.

4.1 Baseline

Past research has explored the use of multiple LLM-based agents for supporting information discovery by integrating insights from multiple perspectives and domains [30]. We implemented a baseline condition that enables similar interactions to mainstream chat-based

systems such as ChatGPT, Claude, Grok, and Gemini in order to evaluating and understanding our proposed interaction designs. We chose to implement the baseline instead of using existing systems (e.g., [30]) in order to separate the confounds from the additional interface and interaction design differences. The baseline offers a conversational interface with a single-textbox input. The interface allows users to chat with multiple agents in a vanilla group-based setting. The agents are designed to respond to user queries in a similar manner as the agents used in Perspectra with access to the literature search tool and GraphRAG database. However, the agents do not implement the deliberation action space as in Perspectrain order to simulate the common implementations of MAS supporting information search.

4.2 Study Procedures

We conducted a within-subjects study with 18 participants to compare Perspectra with a baseline multi-agent group-chat interface. The study was designed to evaluate how the proposed different interaction designs affected participants' critical thinking, perceived domain clarity, and research ideation outcomes.

4.2.1 Participants and Recruitment. We recruited 18 participants spanning undergraduate, master's, Ph.D., and post-doctoral researchers with interdisciplinary research experience. Participants have diverse domain background, and were recruited through university mailing lists and social media platforms. The study was conducted online through Zoom, and participants were compensated with $20 per hour for their time. This study was approved by the university's Institutional Review Board (IRB). Details of participants' demographics are shown in table 2.

4.2.2 Study Design and Protocol. We employed a counterbalanced within-subjects design where each participant completed two research ideation tasks, one using our Perspectra system and one using the baseline group-chat interface. To control for learning and ordering effects, we counterbalanced the order of system presentation across participants.

Each participant's study session lasted approximately 90 minutes. Each session began with a 10-minute introduction where we explained the study goals and completed consent procedures. During this initial phase, we conducted a brief interview to gather information about participants' research backgrounds. Before interacting with each system, participants completed a pre-session survey (7-point Likert scale) to establish baseline measures, including familiarity with the chosen topic, trust in GenAI, and self-assessed initial proposal quality and domain clarity. The core part of the study consisted of two 30-minute system interaction sessions. In one session, participants used the baseline system featuring the group-chat interface for interacting with multiple AI agents. In the other session, they used Perspectra with its threaded forum interface, mind map visualization, and other specialized components. Throughout both sessions, participants were instructed to think aloud, verbalizing their thoughts, reactions, and decision-making processes. These think-aloud sessions were recorded for later analysis.

For each system, participants followed a structured task sequence: they first entered their research idea in sections, including motivation, description of past research, methodology, and hypothetical findings, then engaged with the AI agents to explore their research topic. They used the respective system features to navigate different perspectives and disciplines, collecting insights to develop their research proposal. We demonstrate a typical workflow of how a user interacts with the system in fig. 6. We use the written proposal as a proxy of ideation outcome in research ideation, as proposal writing entails the use of complex reasoning activities (e.g., problem framing, articulation of claims and rationales, synthesis of prior work, and methodological justification) to create an artifact aligned with established critical-thinking constructs [48, 61, 65]. Before asking participants to engage with the systems, we first provided a detailed tutorial of each system's features and functionalities in the form of a walk-through using a sample research idea. During the study, users were encouraged to explore the system freely, but were also instructed to use each feature of the system at least once and ensure they engaged with key features of each system. After each system interaction, participants completed post-session surveys measuring cognitive load, usability, and critical thinking self-assessment (7-point Likert scale). The survey also collects the same set of measurements as in the pre-session survey, including self-perceived interdisciplinary clarity and proposal quality. Detailed survey items for both pre- and post-session surveys are provided in Appendix A.1. Each session concludes with a 15-minute semi-structured exit interview exploring participants' experiences with both systems. The interview focused on identifying "aha moments," useful information gained, perceptions of agent personas, and the utility of different system features. We asked specific questions about which features participants found most helpful for critical thinking to explore different perspectives and compare between the two conditions. The exit interview script can be found in appendix A.5.

4.3 Data Analysis

To evaluate our research questions, we collected and analyzed a combination of qualitative and quantitative data, including system logs, think-aloud protocols, surveys, and exit interviews.

4.3.1 Analysis of System Logs. We collected and analyzed all users' interactions with the system. Specifically for users' textual interactions with agents (replies in Perspectra and chat messages in group-chat condition), two researchers performed a qualitative analysis of users' inputs to the systems during these interactions. We analyzed users' input when replying to agents' posts during the forum condition, and the free-text input when interacting with agents in the baseline chat condition. We first coded these inputs to identify patterns of interaction, focusing on users' intents. Two researchers independently coded the data, and then met to discuss and resolve any discrepancies. the final code includes 13 categories, including `design`, `method`, `critique`, `expand`, `data-seek`, `reflect`, `risk`, `alternative`, `apply`, `compare`, `clarify`, `ethics/impact`, and `sum-marize`. Detailed descriptions of these categories are provided in Appendix A.7. Two researchers independently coded all user inputs using the codebook, achieving a Cohen's kappa of 0.88, which indicates strong inter-rater reliability. Discrepancies were later resolved through discussion with both researchers to reach consensus on the final coding results.

Table 2: Participant demographics and research experience.

PID	Background	Education Level	Research Experience
T1	Computer Science and Artificial Intelligence; Education and Learning Sciences; Human Computer Interaction	PhD Student	3–4 years
T2	Computer Science and Artificial Intelligence; Education and Learning Sciences; Human Computer Interaction	PhD Student	3–4 years
T3	Physics and Astronomy; Engineering and Technology; Computer Science and Artificial Intelligence; Social Sciences (e.g., Sociology, Anthropology); Psychology and Cognitive Science; Mathematics and Statistics; Humanities (e.g., History, Philosophy, Literature); Data Science and Information Technology	Undergraduate Student	1–2 years
T4	Physics and Astronomy; Computer Science and Artificial Intelligence; Mathematics and Statistics	Undergraduate Student	3–4 years
T5	Engineering and Technology; Psychology and Cognitive Science; Neuroscience and Behavioral Sciences; Data Science and Information Technology	Master's Student	1–2 years
T6	Computer Science and Artificial Intelligence; Education and Learning Sciences	PhD Student	3–4 years
T7	Engineering and Technology; Computer Science and Artificial Intelligence; Environmental Science and Sustainability	Master's Student	3–4 years
T8	Biology and Life Sciences; Chemistry and Materials Science	Postdoctoral Researcher	5+ years
T9	Computer Science and Artificial Intelligence; Law, Political Science, and Public Policy; Data Science and Information Technology	PhD Student	5+ years
T10	Engineering and Technology; Environmental Science and Sustainability	Master's Student	1–2 years
T11	Biology and Life Sciences; Medical and Health Sciences; Social Sciences (e.g., Sociology, Anthropology); Law, Political Science, and Public Policy	Undergraduate Student	3–4 years
T12	Agricultural, Food, and Nutritional Sciences	PhD Student	3–4 years
T13	Social Sciences (e.g., Sociology, Anthropology); Psychology and Cognitive Science; Economics, Business, and Management; Neuroscience and Behavioral Sciences	PhD Student	3–4 years
T14	Computer Science and Artificial Intelligence; Education and Learning Sciences; Arts, Design, and Creative Studies	PhD Student	1–2 years
T15	Computer Science and Artificial Intelligence; Education and Learning Sciences; Human Computer Interaction	PhD Student	3–4 years
T16	Computer Science and Artificial Intelligence; Education and Learning Sciences	PhD Student	3–4 years
T17	Computer Science and Artificial Intelligence; Human Computer Interaction	Undergraduate Student	1–2 years
T18	Medical and Health Sciences; Computer Science and Artificial Intelligence; Arts, Design, and Creative Studies; Data Science and Information Technology	PhD Student	3–4 years

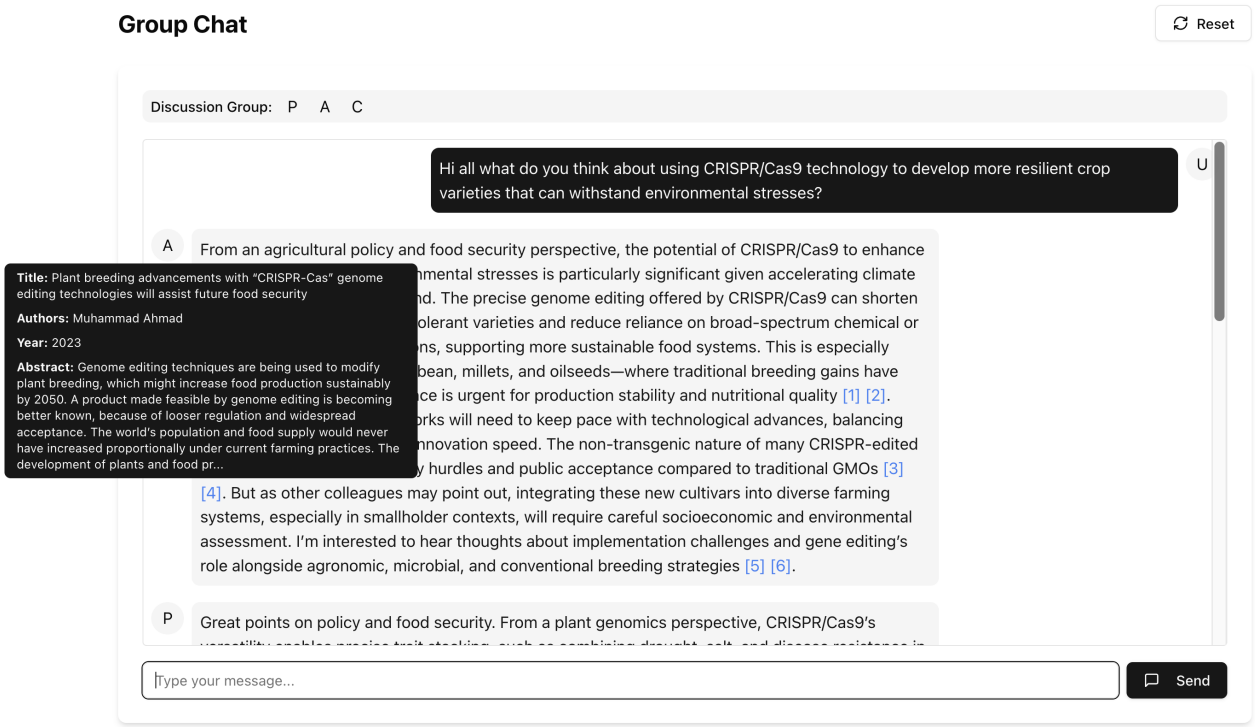


Figure 5: The interface of the baseline group-chat system used for the study. The baseline system implements a single-stream chatting interface without parallel-threading or persona (de)selection support, which approximates mainstream chatting applications. The system does have support for embedding and displaying citations.

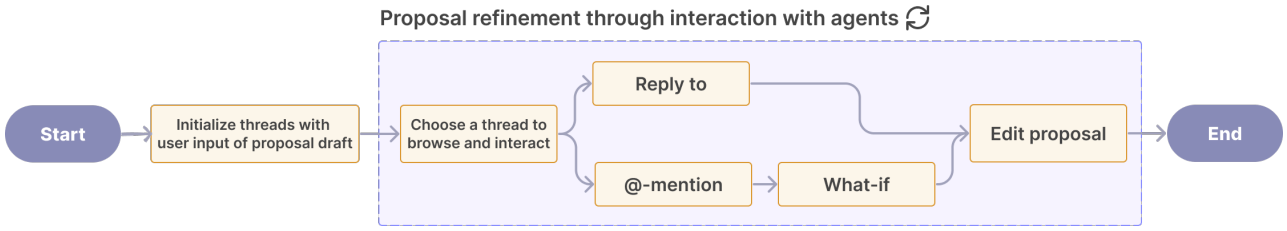


Figure 6: The flow chart of a workflow with Perspectra. Note that this is an example ideal usage scenario rather than a fixed workflow, as the system supports a flexible, non-linear process. Users can iteratively interact with agents and refine their proposal using the integrated notepad at any stage of exploration.

4.3.2 Analysis of Proposal Data. We also perform an LLM-as-a-judge evaluation of the quality of participants' research proposals. We use OpenAI's GPT-5 model to assess the proposal quality based on the same criteria as the self-assessment survey (i.e., coverage, significance, depth, feasibility, and clarity). We excluded relevance as a dimension since the relevance of the proposal and a user's actual intended research topic can only be self-assessed. We prompt the model to rate each proposal on a 1-7 Likert scale across all six dimensions in a pairwise manner, by providing both the initial proposal and the revised final proposal as the context of the prompt. Then we calculate the difference in scores between the two proposals to measure the degree of improvement in proposal quality, where the delta for each metric m and pair i is $\Delta_{i,m} = \mu_{i,m}^{\text{final}} - \mu_{i,m}^{\text{init}}$. To ensure the consistency of the evaluation, we ran 10 independent judgments and averaged the numeric scores by each dimension (for each user and condition). To ensure the reliability of LLM-as-a-judge ratings, we conducted human validation by sampling 10 proposal pairs (5 with the highest LLM ratings and 5 with the lowest LLM ratings) and had two researchers rate them independently using the same rubric.

We obtained a good inter-rater reliability across the five dimensions of an average Krippendorff's alpha of 0.75 using GPT-5, indicating substantial agreement. In order to account for the potential bias of using LLM for evaluate their own outputs [71, 98], we conducted additional experiments using different LLMs from several providers, including LlaMa-4 Maverick[3] (400B MoE model with 17B active parameters) yielding a Krippendorff's alpha of 0.77, Gemini-2.5 Pro [9] yielding a Krippendorff's alpha of 0.73, and Claude Sonnet 4.5[4] yielding a Krippendorff's alpha of 0.77.

4.3.3 Analysis of Think-Aloud Data. During the study sessions, we collected think-aloud data to gain insight into participants' cognitive processes. The analysis of this data focused on identifying moments of critical thinking, such as behaviors that align with Bloom's taxonomy, instances where participants identified contradictions or gaps in arguments, and "aha" moments indicating a change in perspective.

4.3.4 Survey Data. Post-session surveys were used to collect participants' subjective feedback on their experience. The surveys included established instruments such as the NASA-TLX for cognitive load and the System Usability Scale (SUS) for usability. We also included custom scales to measure perceived interdisciplinary clarity, adapted from prior work [68, 85], which covered conceptual, methodological, role, and communication clarity. Participants also completed a self-assessment of their critical thinking and reflection, covering a range of cognitive activities from knowledge recall to self-regulation. Finally, a self-assessment of the quality of their research proposal was collected, evaluating aspects such as coverage, significance, relevance, depth, feasibility, and clarity.

4.3.5 Exit Interview. We conducted 15-minute semi-structured exit interviews to gather qualitative feedback on the user experience. Questions were designed to elicit reflections on moments that were particularly helpful or challenging for their critical thinking process and to identify which specific features supported the exploration of different perspectives. The exit interview script can be found in appendix A.5.

5 Findings

To understand how Perspectra supports users' interdisciplinary deliberation, we conducted a mixed-methods analysis of user interactions, survey responses, and think-aloud data. Our findings aim to address the following research questions:

- **RQ1:** *How does Perspectra impact users' proposal revisions and quality?*
- **RQ2:** *How do users leverage Perspectra's unique features?*
- **RQ3:** *How does Perspectra's interaction design influence users' critical thinking activities?*

5.1 Perspectra Enhances Proposal Quality without Increasing Cognitive Load (RQ1)

In this section, we first present improvement of users' proposal edits across each condition; we then present two case studies to demonstrate typical usage patterns of the two systems; finally, we present the user-perceived usability of Perspectra and each feature.

5.1.1 Improved Clarity and Feasibility of Written Proposals. The Perspectra condition led to better improvements in proposal quality compared to the baseline condition, as evaluated by both LLM-based assessment and user self-assessment. The forum condition showed larger mean gains than chat on the dimensions of *Clarity* (M=0.87 vs. 0.39; $t = -2.15$, $p = .039^{*}$) and *Feasibility* (M=0.56 vs. 0.23; $t = -2.37$, $p = .024^{*}$). The *Overall* rating also favored forum condition but not significantly (M=0.90 vs. 0.66; $p = .235$).

[3] https://ai.meta.com/blog/llama-4-multimodal-intelligence/
[4] https://www.anthropic.com/news/claude-sonnet-4-5

T18's proposal edits (Perspectra):
Family Conversational Agents

Motivation:

I want to learn more about how people work with or receive service from a robot team instead of just one robot
[+] clarify aspects to explore - (human perception): efficiency, usability, and emotional comfort.
[+] clarify robot team attributes to explore - (robot team attributes): specialized and clearly differentiated roles versus homogeneous teams, and whether and how they show those attributes to the users (visibility of specialization, or social cues)
[+] Other factors that can affect the perception of the interaction: task complexity can vary the perception between the same user and the same robot team, also task nature

Related Work:

much prior exploration is about one human interacting with one robot, but not much about one human with multiple ones and how they feel about the team dynamics
[+] transparency related work is limited in multi-robot setting
[+] some current techniques can help increase transparency
[+] too much competency can lead to different perception

Methods:

literature review, select a domain
[+] potentially healthcare service
, select a scenario for detailed experiments, user study design

Potential Outcomes:

robot teammates' interaction quality will affect people's trust; in different cases, when robots make mistakes, people may have different tolerance to the mistakes
[+] now focus on add transparency during the process and control other variables like task complexity, but mention in discussion

T18's proposal edits (Group Chat):
Human–Robot Interaction

Motivation:

how to improve transparency of human-robot-interaction specifically in a human - multiple robots settings

Related Work:

human robot one on one interaction but not much about robot team
[COPIED] definition: transparency in robot-team collaboration is fundamentally ...
[+] social perspective related work: SAT
[COPIED] effective explanation interfaces must support "drill-down" capability ...
[COPIED] Another critical challenge as teams scale is maintaining mutual awareness ...
[COPIED] adaptive transparency At a sociotechnical level ...

Methods:

literature rievew, tool design, propotype development, evaluation
[COPIED] system design: expose their internal decision-making processes ...
[+] open question: strategies to balance detail vs overload?
[COPIED] From the qualitative side, I'd suggest deploying adaptive transparency in a hospital logistics context ...
[COPIED] That's a fantastic take—I'd echo the critical value of "soft" outcome measures ...

Potential Outcomes:

more efforts should be put on how to show the coordination between the robot teammates, and level of this transparency should be carefully considered

Figure 7: Side-by-side comparison of T18's proposal edits across two conditions. Purple text with a leading [COPIED] refers to text chunk copied directly from agents' quotes and pasted in proposal (we omitted the full texts of these quotes except the leading sentences due to their often verbosity, full text can be found in appendix B). As shown in this figure, T18's edits using Perspectra reflect more active thinking, whereas the edits using the group-chat interface are mostly copy-and-pastes.

Users' self-perceptions mirrored these results, where participants gave higher rating improvement (comparing pre- and post-session ratings) of proposal quality under Perspectra compared to the group-chat condition, with the largest gains in Coverage (M=0.41 vs. M=0.19) and Significance (M=0.50 vs. M=0.25).

5.1.2 Significantly More Revisions and Better Motivated Proposals. The use of Perspectra also led to significantly more revisions overall (M=5.35 vs. M=2.19). A chi-squared test revealed a strong association between the condition and which fields of the notepad users edited ($\chi^2 = 180.33, p < .001^{***}$). Post-hoc tests [5] showed that forum users edited the *Motivation* section significantly more than chat users (33.1% of revisions vs. 4.3%; $z = -3.98, p < 0.001^{***}$). While other per-field edit counts and magnitudes showed similar trends favoring the forum condition, they were not significant after correction. It is worth noting that we found no significant differences in the amount of revisions over the *Notes*, where most of the additions were made using the quick note-taking feature (as described in section 3.2). This observation suggests that although users find a similar amount of relevant content in each condition, they are more likely to translate this content into structured proposal changes during the use of Perspectra. One example of this was T18, whose proposal edits using Perspectra showcased how they were able to utilize diverse agents' inputs to strengthen different aspects of their initial research idea, as shown in fig. 7.

5.1.3 No Major Differences of Cognitive Load between Perspectra and Control. While overall UX and usability ratings (7-point Likert scale) were high and did not differ significantly between the Perspectra and group-based conditions, the forum interface was perceived as slightly less demanding. Regarding cognitive load, Perspectra was rated as requiring slightly less "Mental Demand" (M=3.89 vs. 3.94), "Effort" (M=3.72 vs. 4.13), and inducing less "Stress" (M=2.94 vs. 3.50) compared to the group-chat interface, though these differences were not statistically significant. More detailed results can be found in appendix A.2.3. Participants found the persona-based agents more helpful in the forum condition (M=5.67, SD=1.08) than in the group-chat condition (M=5.00, SD=1.37). For the forum interface, the most valued features were the *Forum-based Reply Interaction* (M=5.76, SD=1.48), *Expert Persona Customization* (M=5.67, SD=1.08), and the *Forum-based Layout* (M=5.39, SD=1.29). Across participants' think-aloud data, we found that participants valued reading multiple expert viewpoints in dialogue, rather than a single stream of answers from a single persona (N=9). As noted by T10, "*... this interface makes more sense than down here (the group-chat*

[5] Two-proportion z-tests with Benjamini-Hochberg correction.

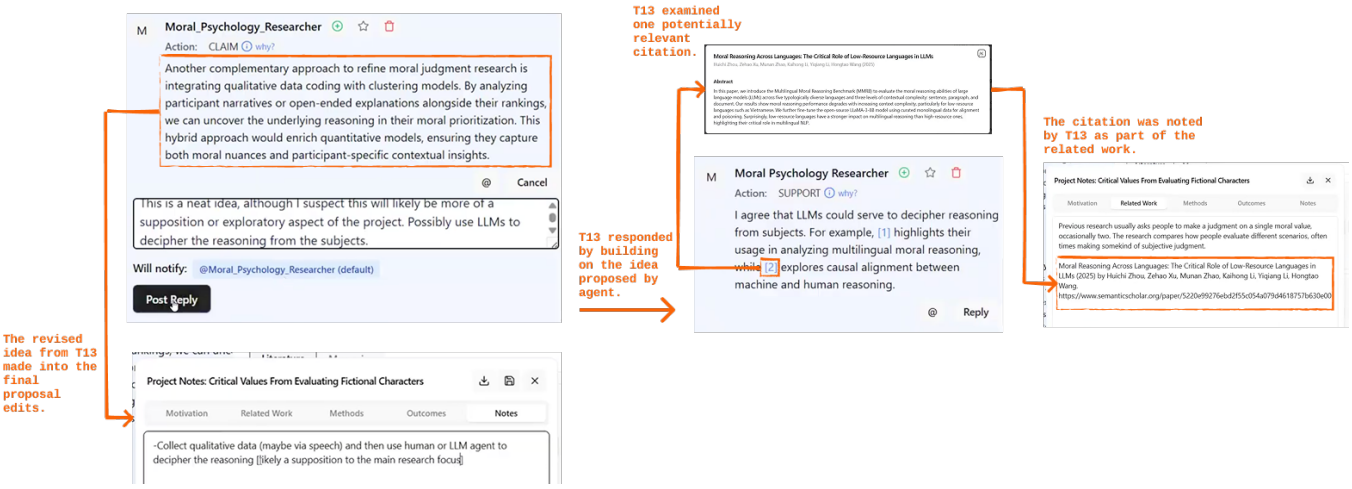


Figure 8: An illustration of how T13 interacted with "Moral Psychology Researcher" agent and collaboratively developed an idea through user reply; the screenshot of the notepad shows the immediate edit made by the user after reflecting on the agent's response.

interface), just because ... I think it's just like easier to understand what everyone is adding to the conversation." T17 mentioned preference for the broader coverage from different perspectives, *"I like the first one, because there is discussion between different roles"*, and mentioned being *"... inspired by some perspectives from several agents ..."*. Participants also noted that agents *"respond[ing] to each other"* made threads clearer (T17) and, when combined with the mind-map, helped them *"mentally group what was going on"* across questions (T6).

5.2 Perspectra Achieves Design Goals and Catalyzes New Affordances (RQ2)

In this section, we analyzed users' interaction system logs and their think-aloud data to uncover common affordances, how they interpreted the system and agents, and new observations about user behaviors and feature use case scenarios emerged beyond the initial design goals.

5.2.1 Designed Affordances of @-mention and reply: Panels, Threads, and Expert Targeting (DG1). In general, we found that users actively sought to synthesize information from multiple disciplines, often using @-mentions to bring different expert agents into a single conversation thread. Users were able to utilize the @-mentions effectively (4.67 times per user on average). We also found that forum replies revealed that 45.1% (23 out of 51) of user replies were interdisciplinary. The proportion of interdisciplinary replies increased when users explicitly used @-mentions, where 58.3% of such replies were cross-disciplinary, compared to 41.0% for replies without mentions.

For example, a case study of T13 is shown in fig. 8, a behavioral economy researcher on discussing the impact of AI on human decision-making. During exploration, T13 came across a "Moral Psychology Researcher" agent who proposed the idea of combining qualitative data coding with clustering models. As the agent explored the methodological extension, T13 explicitly endorsed the idea through reply and reflection, and then further tightened the scope to include deciphering of subject reasoning using LLM. The idea and its retrieved citation were later recorded in the proposal.

Other participants also demonstrated similar use of the reply feature to interact with single agents. More specifically, users would employ single @-mentions to request deep, specialized knowledge from a specific expert. One common use by participants was to seek concrete answers to questions that required domain-specific knowledge, such as asking a "@HCI Researcher" about user evaluation study design (T1) or an "@Healthcare Policy Analyst" about examples of infrastructure gaps and cultural barriers in engaging caregivers in rural areas (T11).

5.2.2 Eliciting Feedback from Multiple Agents via Sensemaking Structured Dialogue Map (DG2). In general, participants appreciated that agents under the forum condition provided more critiques and disagreement compared to the group-chat condition and common off-the-shelf chatting applications and perceived them as beneficial to their ideation process (N=6). This is largely due to Perspectra's feature that enables visualization of agents' deliberation action and rationales.

In the case of T13, they switched to explore a different thread by using the mind map feature. They skimmed the sentence-level summaries of agents' responses, and decided to dig deeper into a branch related to socio-political identity shaping moral judgments. The participant utilized the rationales of agent actions (displayed when the cursor hovers over an action label) to help them understand the thinking process behind agents' responses. As shown in fig. 9, T13 took note of one of the agents' rationale of a "REBUT" action, by combining it with their own ideas. Another example is T5, who was prompted to ask follow-up questions after seeing a reply with action type "QUESTION" from "Cybersecurity Specialist." This reply was asked in context to another response from "HCI Research" giving an opinion on designing for privacy preservation during real-time feedback loops. "Cybersecurity Specialist" asked a

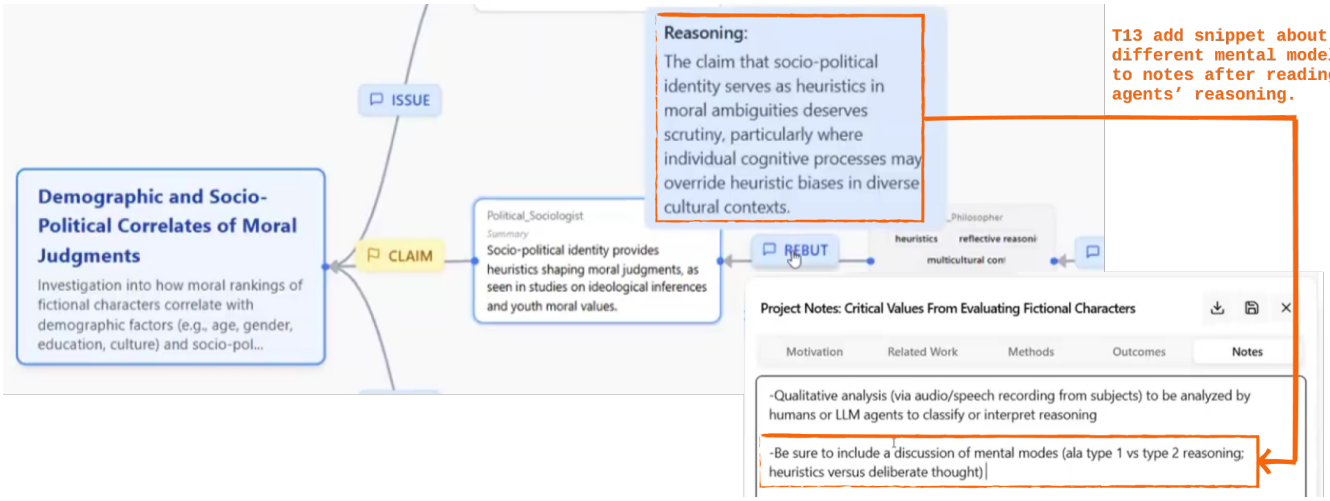


Figure 9: An illustration of how T13 used rationale between agent action for sensemaking of agent responses; the reasoning of the "Rebut" action by the "Moral Philosopher" inspired users' consideration about different cognitive mental models.

two-part question to this response: 1) how to ensure user consent, and 2) what are practical techniques that can be applied for this use case. T5 then commented on his rationale for involving multiple agents, noting that since it was both a *"privacy question... [and] an engineering question"*, they needed to tag both *"the policy ones... and the engineers."* They then included five agents ("@AI Researcher," "@Data Science Ethicist," "@Privacy Advocate," "@Machine Learning Engineer," and "@Ethics and Policy Researcher") to form a discussion over the topic of blockchain-based encrypted consent workflow for an ML pipeline.

T13 also adopted agents' suggestions within a self-chosen scope, with the goal of protecting the core contribution being original and validated. This discussion was later incorporated in their final proposal:

> *" Collect qualitative data (maybe via speech) and then use human or LLM agent to decipher the reasoning*
> *[+] likely a supposition to the main research focus.*
> *[+] Qualitative analysis (via audio/speech recording from subjects) to be analyzed by humans or LLM agents to classify or interpret reasoning.*
> *[+] Key Terms to Check — Chain of Thought (CoT)*
> *[+] Be sure to include a discussion of mental modes (ala type 1 vs type 2 reasoning; heuristics versus deliberate thought). "* — addition over the initial method section of the proposal by T13

This preference for critical discussions is also reflected upon by many other participants, as mentioned by T9 *"if you're just going to all keep agreeing... I didn't really learn anything new"*, noting that LLMs should *"give us different view[s]... another way of seeing it"* beyond what papers already provide (T3). The visible critique across personas normalized constructive disagreement, as mentioned by T17 *"I like the debating a lot ... criticism make projects better"*, and created opportunities for new perspectives, similar to lab meetings where colleagues articulate why they hold a view and propose ways forward. T7 noted that Perspectra drove them to engage in more active thinking, *"I need to read them and ... and use my own analysis to keep the discussion going ... ,"* while in the group-chat condition, the participant expressed a lack of trust as most content is solely generated by agents without much human input. Several participants also sought to actively steer divergence, *"choose [an agent] to take an agree path versus a disagree path versus bringing up a new question"* (T7), and valued getting different opinions beyond own research domains, as mentioned in *"... [as biology researchers] we never think about ethics, laws, or anything like economics... "* (T8). This preference reflects the need to *"hear what other people are thinking"* (T6) and bring a *"fresh set of eyes"* to their work (T13). T9 also noted the convenience of being able to *"talk to anyone"* when needed. As T6 noted, *"this simulates like how I interact with my lab mates ... we're not just pulling up papers ... we're just thinking about the problem itself, like, how would we go about tackling it. What methods would we use?"* and *"I just want to hear what other people are thinking."*

5.2.3 Emergent Affordances: Proactive Engagement and Verification with User-generated TODOs. While T13 actively involved multiple agents using the @-mention feature to gather input from different perspectives about a methodological question, as shown in fig. 10, we observed an unexpected workflow. Specifically, in contrast with interacting with one agent to gather details about related work (example of T13 in fig. 8), the user planned TODO lists using a notepad, and called in multiple users to tackle one TODO at a time, gathering critical feedback about method design. T13 decided to engage with multiple additional agents with different backgrounds, as they found the initial response generated by "Cultural Psychologist" lacked the desired details about the proposed method. As T13 noted *"Well, how would you (differentiate the demographics based on cultural elements)? ... let's ask one of the quantitative people ... ,"* they selected "Ethics-Focused Cognitive Scientist," "Data Science Ethist" and "Cultural Psychologist" agents to respond to the question, which later converted into anchors in the outcome section for later follow-up. This case also surfaced an emergent affordance: T13

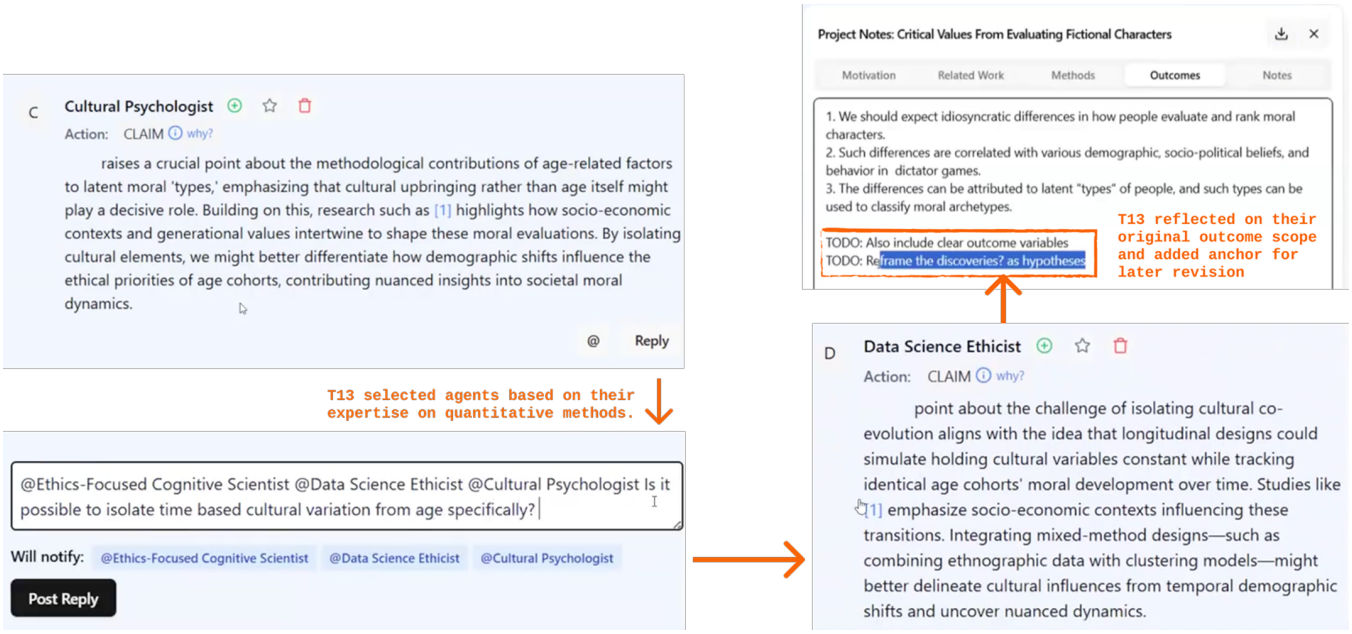


Figure 10: An illustration of how T13 involves multiple agents using the @-mention feature to gather input from different perspectives about a methodological question; the participant rationalized the selection of agents based on their expertise in quantitative methods.

left "TODO" anchors as self-assigned checkpoints for later verification. These anchors flagged agent-suggested claims for subsequent verification (e.g., double-checking on citations). This reflects the user's proactive engagement in thinking and an intention to keep the core contribution original.

When agents conversed with one another, participants found it helpful to observe the exchange of agents acknowledging concerns and offering potential solutions (*"acknowledging the fact that that is a concern, and then also providing solutions to mitigate or overcome that"*) helped them evaluate trade-offs and refine questions. As T6 noted, one agent may surface an inferred stance, *"users might also expect an assistant ... to infer ... the system ... would ... convey some sort of interpretation or stance"*, and that *"the next agent is then acknowledging the fact that that is a concern, and then also providing ... solutions"*, which enabled them to *"pick up their brains and come up with my own reasoning ... think about it in this new light."* These interpretations suggest the system's feature that allows users to participate while the agents build on each other's thoughts facilitates users' active thinking and reflection, and helps them build upon agents' reasoning and thinking process, not through mere recall of information but critical reflections.

5.3 Perspectra Facilitates Higher-Order Critical-Thinking Activities (RQ3)

5.3.1 Commonly Applied Workflows: Charting and Synthesizing Problem Spaces via Branching Multi-Agent Deliberation. We observed that participants often utilized the persona-based agents not only to contextualize their research ideas in unfamiliar domains but also to understand the thinking processes behind different expert roles (N=7). More specifically, participants found the agents helpful for exploring areas beyond their own expertise.

An example of this is T17, who developed a strategy of decomposing complex questions into specific queries targeting chosen experts. In T17's words *"For the 1st tool I ... mentioned the specific agent to answer my question ... break the big questions into small questions."* This strategy was supported by the Perspectra when compared to the baseline condition, as noted by T17: *"But for the second tool (the group-based chat interface) I cannot do that. I can just talking with several agents at the same time ... I cannot mention, specify this agent I want to talk to, and there are too much information."*

The participant then interacted with agents to scaffold understanding. T17 reflected: *"In real world research discussion, like, different people have different backgrounds, knowledge. And we can gain a lot of like visions knowledge from other people's. And in this tool I kind of feel similar because I feel like I... was inspired by some... perspectives from several agents."* The reply-to threading did reinforce the participant's perceived clarity of the collaborative discourse: *"Agents respond to each other makes the ... thread very clear ... it is very clear to review those comments."*

Perspectra's agent deliberation then helped shape T17's proposal from a broad idea into a more feasible study. After agents mentioned the need for "more metrics" and T17 observed that "quantitative measurement would be better," they added concrete measures like "number of turns to concede" and "time to concede," reflecting *Evaluation* and *Inference*. Discussion about personas' "different strategies of negotiation" helped T17 highlight negotiation as a multi-turn process where the process itself matters. Mentions of "consistency" further motivated an added experiment on personality

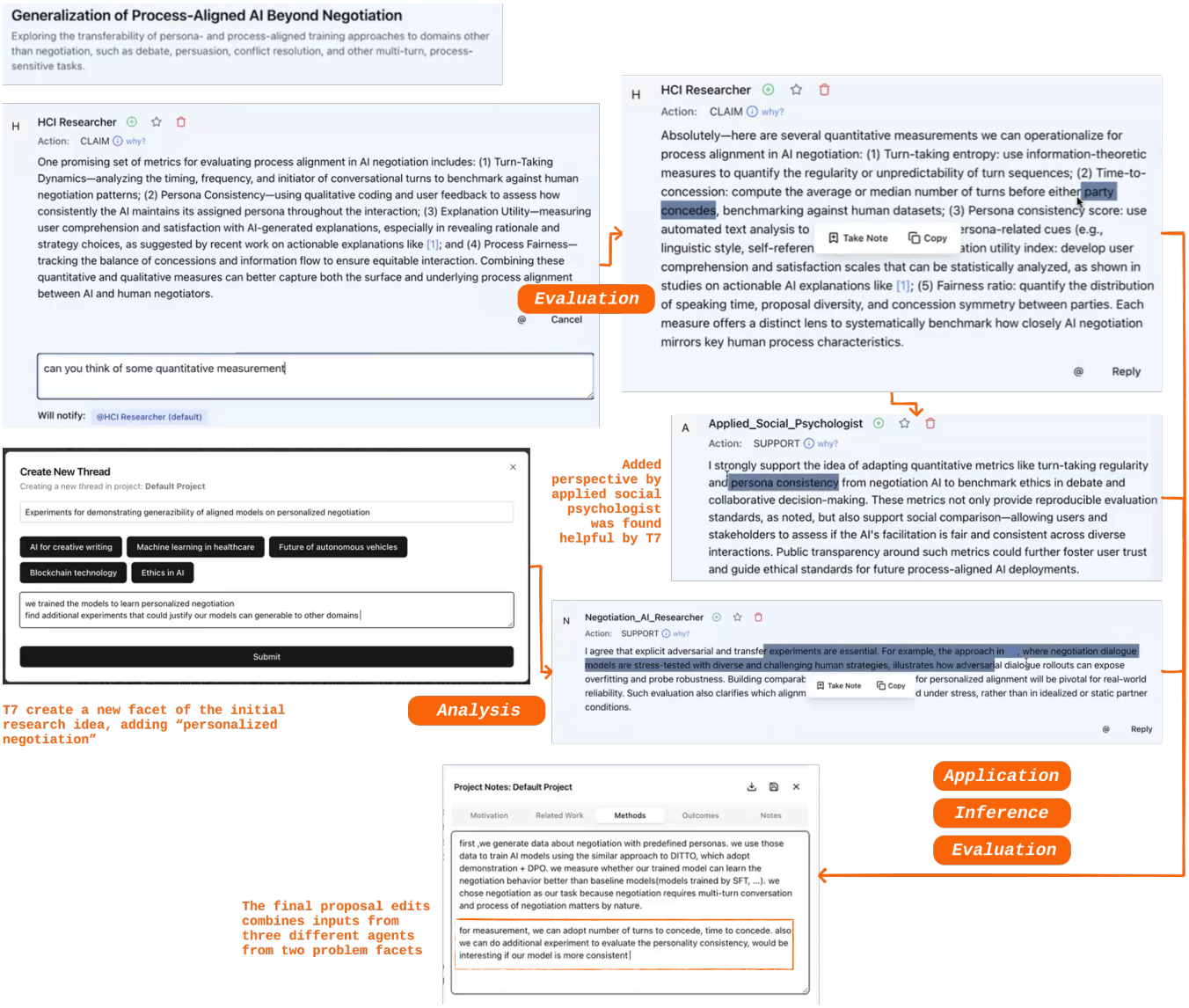


Figure 11: An illustration of T17's workflow, showcasing how the participant decomposes the initial research idea into two different facets during exploration: 1) process-aligned evaluation and 2) personalized negotiation. The user utilized the "create new thread" feature to initiate a new discussion about the second facet; the final proposal edit combines inputs from three different agents throughout the two facets.

consistency, reflecting *Application* and *Inference*. As shown in T17's note edits from fig. 11, it is evident how agents' suggestions made into the final edited proposal.

Other participants also found agents useful for contextualizing their ideas in unfamiliar domains. One noted the system helped them *"think about applications beyond the usual [research topic] scenarios"* by providing concrete examples (T18). For discussions involving multiple agents, another participant found it helpful to *"very quickly see how different disciplines would go about solving the problem"* (T16). Agents also provided context-specific reasoning, for instance, by explaining why a solution applicable in one region might fail in the US due to *"infrastructure gaps"* (T11), thus helping users ground their ideas in realistic settings. Some participants went deeper and reflected on the reasoning process of the personas. T16, for example, noted that through the dialogue, they could see how an engineer persona thinks about success in terms of *"what kinds of metrics you can use to measure your success... how do you concretely turn that into measurable indicators"*. T16 found this insightful, stating, *"I could understand more about... as that kind of persona, how their way of thinking changes... Through multi-round interactions between different agents, I can know how they approach thinking about a problem."* T9 wanted to take this further, as noted in *"create*

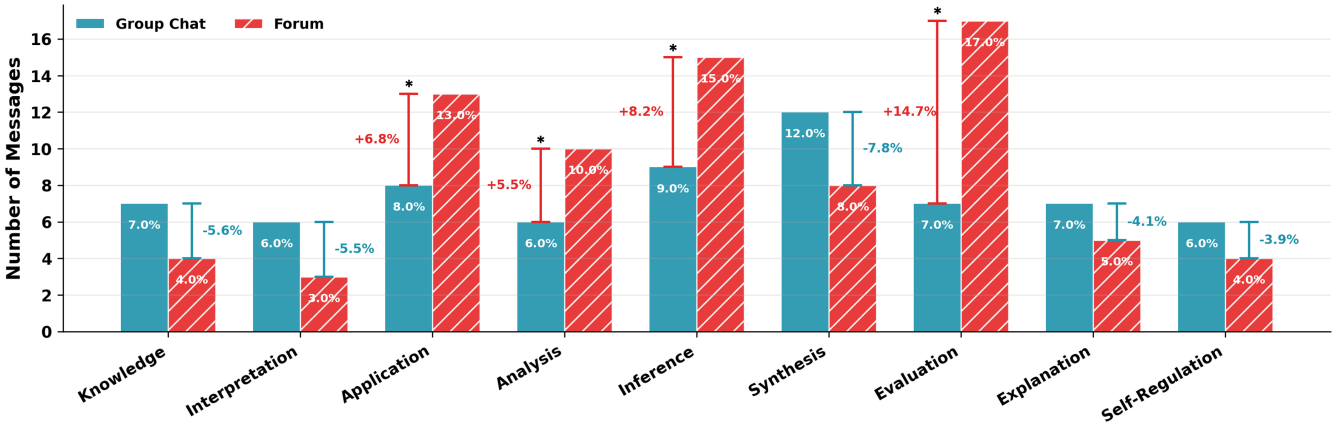


Figure 12: Comparison of user-initiated messages coded with Facione's critical thinking skills across conditions (Positive = PerspectraHigher, Negative = Group Chat Higher).

my own pipeline... force each one of the agent to ask questions... And once they have the questions and answer, then I force them to rebute...", suggesting the need to directly control agents' thinking process beyond mere interpretation.

5.3.2 Varied Critical Thinking Activities with Multi-Agents. In order to obtain a more systematic understanding of how participants engage in critical thinking activities during system use, We analyzed users' reply messages in Perspectra and compared them with their chat messages in the group-based condition, by annotating the messages with the critical thinking codebook described in section 4.3.1. Results show that Perspectra elicited significantly more *Method* (15.6% vs. 4.8%; $t = 2.04$, $p = .044^{*}$) and *Alternative* exploration (12.5% vs. 1.6%; $t = 2.45$, $p = .017^{*}$) activities. Other categories (e.g., *Risk* assessment: 10.9% vs. 3.2%, $p = .089$; *Critique*: 10.9% vs. 9.5%, n.s.) trended in the same direction but did not reach statistical significance given the smaller message sample. As shown in fig. 12, the results reveal a significant difference in the distribution of critical thinking skill codes between the two conditions ($\chi^2 = 5.68, p < .05^{**}$), with users using Perspectra demonstrated more *Inference* (+8.2%), *Analysis* (+5.5%), *Application* (+6.8%), and *Evaluation* (+14.7%) activities during interactions with agents.

5.3.3 Inference, Application, and Evaluation Fostered by Perspectra. While the relative frequency of *Inference*, *Application*, and *Evaluation* codes was lower, skills related to communication and reflection were more prominent under the group-chat condition. Specifically, chat interactions contained more instances of *Knowledge*, *Interpretation*, *Self-Regulation*, and *Synthesis*. Users appeared to use the group-chat interaction for seeking clarification and expanding on ideas.

During the use of Perspectra, as an example of *Evaluation*, T8 asked, *"but CRISPR based imaging has not been validated with high significance thus, would lead to nonsensical analysis"* in Perspectra, actively questioning the validity of the information provided by the agents, while most of T8's messages in the group-based condition are information-seeking questions. T16 also demonstrated *Application* by reflecting on how to implement an idea in practice through requesting demonstration given a user-defined example: *"In addition to interactive repair, are there any other design patterns that can support parent-child joint engagement? For example, how can the CAs initiate a conversation or an activity that involves both parties? @HCI Research do you know any implementations?"* This question was directed towards "Developmental Psychologist," who previously raised the idea of "interactive co-repair" in the discussion. The user grew intrigued by this point and proceed to inquire about other design patterns that could be applied in this context, also by tagging another agent to increase coverage. *Inference* behaviors are associated with envisioning potential risks and ethical concerns relevant to the proposed ideas. For instance, T5 reviewed the idea of integrating circuit theory into transcriptomics using graph-based methods, which proposed by "Computational Biologist." T5 commented that although the idea was interesting, there might be potential issues based on their existing knowledge of this field, thus asking the question *"Since the full gene regulatory network has not been explored, what can be the problems?"*

Participants' think-aloud data also revealed that they were more likely to engage in activities such as questioning validity, examining assumptions, and refining interpretations when using Perspectra. More specifically, participants were more likely to ask questions about the validity of the information provided by the agents (e.g., *"It should be geochemist, not environmental chemistry ... that would be more specific."* — T10) and to reflect on their own understanding of the topic (e.g., *"... this one (Cognitive Science Researcher) is very heavy on like the cognitive stuff. So I'm thinking that they maybe have some sort of background in Psychology or something like that ... "* — T6).

5.3.4 Complementary Engagement in the Group-Chat Condition for More Specialized Inquiries. When using the group-chat condition, the most common observations are information-seeking queries like "*What coding assistants exist that I could test with using user studies?*" (T7) and *"Give me a literature review on algorithmic accountability*

in the past 5 years" (T9), which resemble typical off-the-shelf search engine and chatbot queries. For queries that fall under the *Synthesis* category, participants tend to send out agent-led synthesis queries, which offload most of the reasoning to agents, such as *"How could we frame a research proposal looking to understand the mechanisms by which we can increase research transparency between participants and researchers?"* (T11) or even *"Elaborate on the research gap and draft a research proposal?"* (T9). This is coherent with our observations that they focus on specialized knowledge and gather more information about a given topic.

Users also commented on the complementary roles of Perspectra and the group-chat design in supporting different aspects of critical thinking. Perspectra's panel-like interaction design to support critical thinking and reflections in a controllable manner that allows users to inject more of their own thinking and mitigate anchoring bias. During the exit interview, T9 explicitly compared Perspectra with the STORM system [30] with respect to Perspectra's capability that allows users to perform more fine-grained control over discussion: *"... it's very important to know where things go wrong so we can do counterfactual explanations and follow-up, and I think a pre-generated wall of text just doesn't do that..."*, which helps the user to *"... not be as biased to a preconceived notion ..."* On the other hand, participants commented on the scenarios the group-chat is preferred for quick direct Q&A about simpler information-seeking questions due to lower overhead, as noted by T6 *"... (when using Perspectra) before you get to that point there's a lot of information ... to take in and understand."*

6 Discussion

In this section, we discuss major insights and takeaways from our findings by situating them in and extending discussions from prior research.

6.1 Beyond Dialogue: Enabling User-Steered Multi-Agent Deliberation

The design of Perspectra managed to scaffold discourse in ways a linear group-chat interface did not support. This led to improvement in discussion diversity without extra cognitive load, as observed in how participants used Perspectra to support their own cognitive process: including decomposing complex problems into sub-questions (T17's case from section 5.1.3), ad-hoc "panel-ization" (RQ2), and contextualizing ideas across unfamiliar application settings (RQ3). Perspectra's design contributes to the growing body of research that explores design alternatives to traditional chat interfaces for LLM-based ideation systems that aim to balance between diversity and depth of exploration [30, 33, 46, 78]. Our findings provide further insights into multi-agent ideation support systems, in which case diversity vs. depth largely depends on the balance between user control and agent-driven self-orchestration [16, 111]. In Perspectra's case, we used features such as @-mention and reply to provide users with more control over the exploration process. While it introduced friction for users wanting to quickly gather information and reach conclusions, it also encouraged slower, more deliberate thinking.

Past research has indicated the use of personas not only as a medium of information but also as a mechanism of interaction through persona customization [46, 87] and selection of personas prior to debate [83]. Our results suggest an alternative interaction design that allows users to "panelize" expert personas on the fly and branch into different threads. Participants demonstrated two valuable use case scenarios: 1) selecting and adding agents to better steer the exploration and consolidation of parallel threads (e.g., exploring decomposed sub-topics, and forming an ad-hoc panel of diverse background agents); 2) observing and interpreting the reasoning of agents for unfamiliar topics. This dynamic introduction of additional agents during an ongoing discussion can also be perceived as a new form of social affordances [57], where the ideation process is projected as dialogues, and the inclusion of a new participant (i.e., agent in our scenario) acts as a discursive move. Drawing from the philosophical concept of Platonic dialogues where inquiry is often (almost exclusively in the case of Plato) documented as dialogues among distinct voices [75], this form of interaction allows people to digest and externalize cognitive conflicts that are often less tangible. When situated in the context of collaborative ideation with LLM agents, this affordance is also worthy of discussion because it transforms the direct "inquiry" with LLM agents into a controllable social process. Our system allows users to inspect the reasoning structure, and control the trajectory of thought by choosing specific agents.

6.2 Sensemaking of Multi-Agent Deliberation through Visualization of Argumentation Acts

Perspectra's forum and map design allowing users to build a structure for deliberation among multiple agents. The proposed structure is built upon argumentation actions, including ISSUES, CLAIM, SUPPORT, REBUT, and QUESTION. Users select, combine, and re-engage specific agents to actively synthesize multiple perspectives rather than passively consuming information. Since active synthesis enhances users' intrinsic motivation and engagement [11], this may explain why it leads to better ideation outcomes (RQ1) and drives active synthesis of knowledge and critical thinking activities (RQ3).

Example observations include how T13 became intrigued with the rationale behind a "REBUT" action of an agent, and translated the rationale into a part of their final proposal, as well as T16's comments on that seeing the structure of the debate helped them trace the reasoning process rather than just the final output. These findings shed light on how Perspectra's design approach of agents' response generation drives users to engage in more in-depth sensemaking of the discussion among them and the agents, when compared to recent works using heuristic threshold-based approaches (e.g., confidence score-based [21] and mental model-driven [42] response), and thus leading to more in-depth discussions (RQ3). Our design highlights transparency in agents' reasoning processes that can potentially mitigate biases from using heuristic-based numerical indicators [4]. Also, evaluating explicit deliberation labels and rationales can also foster calibrated trust [37] between the user and agents, potentially mitigating risks of over-reliance from passively accepting agent responses.

However, there are also limitations to this approach in comparison, as agents' behaviors can be prone to instability due to

a lack of quantifiable decision-making criteria when generating responses [54]. Future work could explore formal argumentation framework-informed heuristic metrics and metrics grounded in real human deliberation data as agents' dialogue activation to improve the consistency and reliability of agent behaviors. Another concern is the limited flexibility for users to provide feedback or calibrate the alignment between their intents and agent behaviors. Future iterations could address this by enabling users to define custom deliberation acts or annotation schemas to enable more granular control over the structure of the deliberation.

Theories in argumentation research [94] already posit that confronting opposing viewpoints strengthens reasoning, encourages sensemaking, and improves judgment quality. When people are challenged, they are more likely to scrutinize their claims, consider alternative perspectives, and refine their arguments, leading to deeper understanding and cognitive growth. At the same time, our use of an LLM-as-judge to assess proposal quality may introduce interpretive constraints. Although reflecting a broader trend toward automated evaluation of generative systems, recent work shows that LLM judges can exhibit self-preference, style, and positional biases that systematically deviate from human norms [71, 98]. For this reason, we do not claim that our results as evidence that Perspectra produces objectively better proposals. Instead, we treat the LLM scores as a validation of behavioral observations (e.g., more and deeper revisions) and higher-order critical-thinking behaviors based on the system logs and think-aloud. Related, prior research [15] has also called for preservation of cognitive conflict to promote both divergent and convergent thinking, and eventually creativity during collaborative ideation, through exposure to and reflection on minority dissent [63]. As our findings regarding improvements in ideation quality (RQ1) echo this perspective, future designs could consider mechanisms to further highlight conflicting views such as providing visual highlights of minority opinions or adaptively synthesizing potential agents with contrasting personas.

Finally, recent research on Context Engineering [59] seeks to formalize how information is curated, structured, and communicated in LLM-based systems. Our work contributes to this emerging space by demonstrating how MAS context engineering can be oriented specifically toward communication between users and agents with more tractable knowledge exchange, stronger interpretability, and increased user control over multi-agent deliberation.

6.3 Design Implications for Controlled, Critical, and Active Deliberation with Multi-agents

6.3.1 Fostering Critical Thinking through Visualizing Adversarial Discourse. Participants generally valued the adversarial discourse in Perspectra suggesting further exploration of design strategies that can foster such discourse in LLM-supported ideation systems. Additionally, some participants commented they would like the comments to be even more "critical" (RQ2). This is in stark contrast with existing chat interaction, where LLMs tend to generate responses that agree with the user's given stance (i.e., sycophancy [82]). Our findings suggest the value of adversarial discourse in LLM-supported ideation systems, also drawing from past research related to socio-cognitive conflict [60] and educational psychology [58]. In the context of knowledge-intensive ideation systems, adversarial discourse can be supported by designing agents that are encouraged to take on diverse and potentially conflicting perspectives, yet still grounded in evidence-based reasoning [5, 25, 95]. This also covers the aspect of diversity potentially brought by adversarial deliberation, as underlined by past research suggesting that diverse perspectives can enhance collective problem-solving capabilities [69]. On the other hand, we hypothesize that in a real-world ideation process, there exists a potential trade-off between the cognitive cost of digesting critical discourse and its benefits. Excessive criticism towards one's own ideas could lead to decision fatigue or discouragement and thus hinder creativity [10]. In the design of Perspectra, agents' initial critiques target other agents rather than the user, which can be generalized as a way to mitigate discouragement by directing critical feedback towards an abstract identity, such as agent personas.

Additionally, in order to support better visualization and sensemaking, design considerations can include implementing transparency features, such as temporal visualizations of an agent's reasoning paths and confidence scores for its claims and suggestions. These features could also help users better gauge the certainty of agent responses in terms of their expertise and the evidence gathered (e.g., literature surveyed), thus granting them stronger agency and confidence in the decision-making process [31]. These design considerations extend to broader educational technology contexts, where AI agents may act as facilitators of learning rather than mere information retrievers [8, 17].

Finally, future iterations of LLM-based Multi-Agent Systems should also consider enhancing the design to better facilitate human control and agency over multi-agent collaboration to improve the collaborative output [84]. This aligns with the concept of adjustable autonomy, where users can dynamically manage the level of agent intervention by transferring decision-making control to the human in key situations [79]. Specific design improvements could include interactive mechanisms allowing users to perform direct manipulations of agents' collaboration or discussion history, and directions going forward.

6.3.2 Balancing User Control and Agent Autonomy in Multi-Agent Ideation Systems: a Friction-Guided Approach. Balancing between user control and system autonomy in interactive ideation and information exploration systems has been a long-standing challenge in HCI and information retrieval research [24, 53]. Wider industrial adoption also exists such as the recent OpenAI's branching feature that allows users to create multiple parallel threads of exploration from a single starting point [66], and similar branching design has also been adopted earlier in other LLM-powered systems such as Claude. But still, in a more knowledge-intensive context, it remains a design challenge in terms of how to blend user control with quality of retrieved information [30, 111] — *how can system design grant user control while maintaining the depth of exploration?*

Perspectra piloted one potential design that can effectively support the diversity of exploration while preserving the depth of discussion. This is highly relevant to the design of future knowledge-intensive search systems, such as deep research systems [26, 111], where users need to balance between exploring diverse perspectives and engaging in deep discussions. Perspectra's design can

be seen as an alternative way to support a mixture of both automated and controllable exploration when it comes to ideation and knowledge synthesis, complementing existing methods used in existing research aiming to balance between amount of required user input and the amount of relevant information the system can provide in return, such as works by Liu et al. [46] and Chen et al. [7]. Still, future iterations following this design direction should consider the friction-efficiency trade-off. Practically speaking, designs should consider 1) minimizing the friction when interactions encourage participants to perform active thinking; and 2) providing hybrid options that allow users to switch between low-friction and high-friction modes based on their current needs and goals. Contextualizing this in the design of Perspectra, when designing a mind map, one potential improvement is to keep exploration (interpretation) and synthesis within the same thinking space [1] to reduce the unnecessary cognitive switching cost. To the end of balancing control and efficiency, a hybrid design that incorporates follow-up chatting in a forum-style interface could further reduce cognitive switching costs while preserving the benefits of structured deliberation. Additionally, for existing systems that have already considered the design option of branching [66], offering an integrated or alternative visualization that offers overviews and navigation could further enhance user experience and support deeper exploration.

Our findings also suggest that while both Perspectra and group-chat interface are valuable, supporting complementary facets of critical thinking, the forum-style, panel-like design often included structured reasoning (e.g., organizing the motivation of proposals instead of copy-pasting), internalization of knowledge (e.g., awareness of knowledge gaps when selecting agents), and application of knowledge (e.g., seeking generalizable conclusions), whereas information-seeking was found to be more present under the traditional linear design. For use cases where information gathering is the primary goal, less control and more hand-off to agents could be desirable, but for ideation systems where in-depth reasoning and critical thinking are the main goals, the introduced friction for stronger user control could be beneficial [97]. Additionally, this said productive friction also introduces a potential trade-off between enhanced critical thinking and cognitive load. One is that users can become overwhelmed when threads grow quickly in number and become hard to manage [90]. Future work could explore to improve thread management through features such as grouping and summarization, or prioritization using visual analytics [44] based on user-defined criteria (e.g., relevance to goal). Also, in the context of long-term use, increased cognitive load can lead to reduced usage of the system. The long-term cognitive load likely varies across different scenarios, such as during early exploratory stages where users are more tolerant of friction to break fixation versus later convergence phases where the same friction might mostly be obstructive in the context of information search and gathering. Furthermore, the long-term impact also spans concerns related to over-reliance [43] on such systems for ideation. Users could end up treating systems like Perspectra merely as a tool to evade efforts needed for critical thinking rather than using the system as intended, where they need to engage in and internalize active critical thinking. Further longitudinal research is needed to determine systems similar to Perspectra have over users' independent critical thinking. To mitigate the risk of over-reliance, future designs might consider incorporating adaptive scaffolding strategies [12] that take into account users' behavior, where the system gradually reduces its intervention as the user demonstrates increased proficiency in critical reasoning [103].

7 Limitations and Future Work

While the user study revealed the potential of our proposed system for supporting knowledge-intensive ideation tasks, this study has several limitations that are noteworthy. First, we conducted our user study within the context of interdisciplinary research. We chose this context as an example of knowledge work that requires advanced reasoning, although the generalization of insights from this specific context to broader applications needs to be further validated. Second, we did not strictly control the participants' familiarity with the topics and research experience, as we would in a controlled study, which leaves room for potential confounds from the variance in participants' backgrounds. Future studies may consider conducting more systematic investigations of user behavior and perceptions. Moreover, the use of LLM-as-a-judge during the evaluation of user-written proposals can bring in potential biases due to embedded LLM-generated snippets or content [71, 98], which should be further examined in future work through more comprehensive expert evaluation. Additionally, the design of Perspectra will benefit from further contextualization in users' workflow for gathering understandings through methods such as conducting field deployment studies. Also, future studies should consider further investigating the trade-off between enhancing critical thinking and the potential increase in cognitive load in a longitudinal use setting.

8 Conclusion

In this paper, we conducted an exploratory within-subject study (N=18) to examine how different interaction patterns for user control over LLM-based multi-agent collaborations shape interdisciplinary research ideation. We compared two different designs with varied levels of user control: one offering features that allow users to form ad-hoc discussion panels (i.e., Perspectra) and the other adopting a common single-stream group-chat interface. We found that Perspectra elicited more critical-thinking activities and structured proposal revisions, whereas the group chat condition was used more for tasks with specific information-seeking. The findings suggest treating user control as productive friction, making agent reasoning and deliberation adversarial yet grounded, and adopting hybrid designs that balance between user control and information-seeking efficiency.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2119589. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Additionally, results presented in this paper were obtained using CloudBank [64], which is supported by the National Science Foundation under award No. 1925001.

References

[1] Christopher Andrews, Alex Endert, and Chris North. 2010. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 55–64.

[2] Jinheon Baek, S. Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. *ArXiv* abs/2404.07738 (2024).

[3] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, et al. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain.* Longman New York.

[4] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[5] Edward Y. Chang. 2024. SocraSynth: Multi-LLM Reasoning with Conditional Statistics. *ArXiv* abs/2402.06634 (2024).

[6] Kai Chen, Xinfeng Li, Tianpei Yang, Hewei Wang, Wei Dong, and Yang Gao. 2025. Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation. *arXiv preprint arXiv:2503.13856* (2025).

[7] Pei Chen, Jiayi Yao, Zhuoyi Cheng, Yichen Cai, Jiayang Li, Weitao You, and Lingyun Sun. 2025. CoExploreDS: Framing and Advancing Collaborative Design Space Exploration Between Human and AI. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[8] Akcell Chii-Chung Chiang and Isaac Pak-Wah Fung. 2004. Redesigning chat forum for critical thinking in a problem-based learning environment. *The Internet and Higher Education* 7, 4 (2004), 311–328.

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[10] Alwin de Rooij. 2024. Speaking to your inner muse: how self-regulation by inner speaking influences confidence during idea evaluation. *Creativity Research Journal* 36, 2 (2024), 291–308.

[11] Edward L Deci and Richard M Ryan. 2012. Self-determination theory. *Handbook of theories of social psychology* 1, 20 (2012), 416–436.

[12] Paramveer S Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping human-AI collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[13] Peter Facione. 1990. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). (1990).

[14] Peter A Facione et al. 2011. Critical thinking: What it is and why it counts. *Insight assessment* 1, 1 (2011), 1–23.

[15] Umer Farooq, John M Carroll, and Craig H Canoe. 2008. Designing for creativity in computer-supported cooperative work. *International Journal of e-Collaboration (IJeC)* 4, 4 (2008), 51–75.

[16] Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, et al. 2024. Magentic-one: A generalist multi-agent system for solving complex tasks. *arXiv preprint arXiv:2411.04468* (2024).

[17] Kylea R Garces, Aaron N Sexton, Abigail Hazelwood, Nathan Steffens, Linda Fuselier, and Natalie Christian. 2024. It takes two: online and in-person discussions offer complementary learning opportunities for students. *CBE—Life Sciences Education* 23, 3 (2024), ar34.

[18] Aniketh Garikaparthi, Manasi S. Patwardhan, L. Vig, and Arman Cohan. 2025. IRIS: Interactive Research Ideation System for Accelerating Scientific Discovery. *ArXiv* abs/2504.16728 (2025).

[19] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094* (2024).

[20] Alireza Ghafarollahi and Markus J. Buehler. 2024. SciAgents: Automating Scientific Discovery Through Bioinspired Multi-Agent Intelligent Graph Reasoning. *Advances in Materials* (2024).

[21] 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.

[22] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

[23] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779* (2024).

[24] Joshua Holstein, Moritz Diener, and Philipp Spitzer. 2025. From Consumption to Collaboration: Measuring Interaction Patterns to Augment Human Cognition in Open-Ended Tasks. *ArXiv* abs/2504.02780 (2025).

[25] Xinmeng Hou, Zhouquan Lu, Wenli Chen, Hai Hu, and Qing Guo. 2025. EduThink4AI: Translating Educational Critical Thinking into Multi-Agent LLM Systems. *ArXiv* abs/2507.15015 (2025).

[26] Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, et al. 2025. Deep Research Agents: A Systematic Examination And Roadmap. *arXiv preprint arXiv:2506.18096* (2025).

[27] Edwin Hutchins. 1995. *Cognition in the Wild*. MIT press.

[28] Valerie Imbruce, Vanessa Jaeger, Marisa A Rinkus, Jessica Hua, and Michael O'Rourke. 2024. Raising undergraduate researchers' interdisciplinary consciousness through dialogue. *Journal of Environmental Studies and Sciences* (2024), 1–12.

[29] Jamie L Jensen, Mark A McDaniel, Tyler A Kummer, Patricia DDM Godoy, and Bryn St. Clair. 2020. Testing effect on high-level cognitive skills. *CBE—Life Sciences Education* 19, 3 (2020), ar39.

[30] Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J Semnani, and Monica S Lam. 2024. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. *arXiv preprint arXiv:2408.15232* (2024).

[31] Alexander John Karran, Théophile Demazure, Antoine Hudon, Sylvain Senecal, and Pierre-Majorique Léger. 2022. Designing for confidence: The impact of visualizing artificial intelligence decisions. *Frontiers in neuroscience* 16 (2022), 883385.

[32] Kevin Gonyop Kim, Richard Lee Davis, Alessia Eletta Coppi, Alberto Cattaneo, and Pierre Dillenbourg. 2022. Mixplorer: Scaffolding Design Space Exploration through Genetic Recombination of Multiple Peoples' Designs to Support Novices' Creativity. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.

[33] Taewook Kim, Matthew Kay, Yuqian Sun, Melissa Roemmele, Max Kreminski, and John Joon Young Chung. 2025. Scaffolding Recursive Divergence and Convergence in Story Ideation. *arXiv preprint arXiv:2507.03307* (2025).

[34] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. 2023. Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level. *International Conference on Intelligent User Interfaces* (2023).

[35] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140* (2023).

[36] Hao-Ping Hank Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. (2025).

[37] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[38] Soohwan Lee, Seoyeong Hwang, Dajung Kim, and Kyungho Lee. 2025. Conversational Agents as Catalysts for Critical Thinking: Challenging Social Influence in Group Decision-making. *arXiv preprint arXiv:2503.14263* (2025).

[39] Soohwan Lee, Seoyeong Hwang, and Kyungho Lee. 2024. Conversational Agents as Catalysts for Critical Thinking: Challenging Design Fixation in Group Design. *arXiv preprint arXiv:2406.11125* (2024).

[40] Soohwan Lee, Mingyu Kim, Seoyeong Hwang, Dajung Kim, and Kyungho Lee. 2025. Amplifying Minority Voices: AI-Mediated Devil's Advocate System for Inclusive Group Decision-Making. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*. 17–21.

[41] G. Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. *Neural Information Processing Systems* (2023).

[42] Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang'Anthony' Chen. 2025. Proactive conversational agents with inner thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[43] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.

[44] Yijun Liu, Frederick Choi, and Eshwar Chandrasekharan. 2025. Needling Through the Threads: A Visualization Tool for Navigating Threaded Online Discussions. *arXiv preprint arXiv:2506.11276* (2025).

[45] Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2024. Personaflow: Boosting research ideation with llm-simulated expert personas. *arXiv e-prints* (2024), arXiv–2409.

[46] 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*. 506–534.

[47] Qiuyu Lu, Jiawei Fang, Zhihao Yao, Yue Yang, Shiqing Lyu, Haipeng Mi, and L. Yao. 2024. Enabling Generative Design Tools with LLM Agents for Mechanical Computation Devices: A Case Study. (2024).

[48] Ying-Yan Lu, Huann-shyang Lin, Thomas J Smith, Zuway-R Hong, and Wen-Yi Hsu. 2020. The Effects of Critique-Driven Inquiry Intervention on Students' Critical Thinking and Scientific Inquiry Competency. *Journal of Baltic Science Education* 19, 6 (2020), 954–971.

[49] Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiaoming Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Mengxue Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, Yiqiao Jin, Fan Zhang, Xianhong Wu, Hanqing Zhao, Dacheng Tao, Philip S. Yu, and Ming Zhang. 2025. Large Language Model Agent: A Survey on Methodology, Applications and Challenges. *arXiv.org* (2025).

[50] Andrea I Luppi, Pedro AM Mediano, Fernando E Rosas, Negin Holland, Tim D Fryer, John T O'Brien, James B Rowe, David K Menon, Daniel Bor, and Emmanuel A Stamatakis. 2022. A synergistic core for human brain evolution and cognition. *Nature neuroscience* 25, 6 (2022), 771–782.

[51] Shuai Ma, Junling Wang, Yuanhao Zhang, Xiaojuan Ma, and April Yi Wang. 2025. DBox: Scaffolding Algorithmic Programming Learning through Learner-LLM Co-Decomposition. *arXiv preprint arXiv:2502.19133* (2025).

[52] Xiao Ma, Swaroop Mishra, Ariel Liu, S. Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc V. Le, and E. Chi. 2023. Beyond ChatBots: ExploreLLM for Structured Thoughts and Personalized Model Responses. *CHI Extended Abstracts* (2023).

[53] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[54] Vinayak Mathur and Arpit Singh. 2018. The rapidly changing landscape of conversational agents. *arXiv preprint arXiv:1803.08419* (2018).

[55] Peter McBurney and Simon Parsons. 2002. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of logic, language and information* 11 (2002), 315–334.

[56] Robert McNair, Hai Anh Le Phuong, Levente Cseri, and Gyorgy Szekely. 2019. Peer Review of Manuscripts: A Valuable yet Neglected Educational Tool for Early-Career Researchers. *Education Research International* 2019, 1 (2019), 1359362.

[57] Joshua McVeigh-Schultz and Katherine Isbister. 2021. The case for "weird social" in VR/XR: a vision of social superpowers beyond meatspace. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–10.

[58] Janell M Mead and Lawrence C Scharmann. 1994. Enhancing critical thinking through structured academic controversy. *The American Biology Teacher* (1994), 416–419.

[59] Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. 2025. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334* (2025).

[60] Gabriel Mugny and Willem Doise. 1978. Socio-cognitive conflict and structure of individual and collective performances. *European journal of social psychology* 8, 2 (1978), 181–192.

[61] Mari Murtonen and Kieran Balloo. 2019. Redefining scientific thinking for higher education. *Redefining Scientific Thinking for Higher Education. Palgrave MacMillan. https://doi. org/10.1007/978-3-030-24215-2* (2019).

[62] Suchismita Naik, Amanda Snellinger, Austin L Toombs, Scott Saponas, and Amanda K Hall. 2025. Exploring Early Adopters' Use of AI Driven Multi-Agent Systems to Inform Human-Agent Interaction Design: Insights from Industry Practice. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[63] Charlan Jeanne Nemeth and Brendan Nemeth-Brown. 2003. The potential benefits of dissent and diversity for group creativity. *Group creativity: Innovation through collaboration* (2003), 63–84.

[64] Michael Norman, Vince Kellen, Shava Smallen, Brian DeMeulle, Shawn Strande, Ed Lazowska, Naomi Alterman, Rob Fatland, Sarah Stone, Amanda Tan, et al. 2021. Cloudbank: Managed services to simplify cloud access for computer science research and education. In *Practice and Experience in Advanced Research Computing 2021: Evolution Across All Dimensions*. 1–4.

[65] E Michael Nussbaum. 2008. Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology* 33, 3 (2008), 345–359.

[66] openai. 2025. *About branched conversations.* OpenAI Community forum post.

[67] Steven Ovadia. 2014. ResearchGate and Academia. edu: Academic social networks. *Behavioral & social sciences librarian* 33, 3 (2014), 165–169.

[68] Valentine Joseph Owan, Kingsley Bekom Abang, Cecilia Akpana Beshel, and Roseline Anyiopi Undie. 2024. Development and validation of the perception of interdisciplinary research collaboration (PIRC) scale. In *Technological Tools for Innovative Teaching*. IGI Global Scientific Publishing, 292–321.

[69] Scott Page. 2008. *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition.* Princeton University Press.

[70] Bo Pan, Jiaying Lu, Ke Wang, Li Zheng, Zhen Wen, Yingchaojie Feng, Minfeng Zhu, and Wei Chen. 2024. AgentCoord: Visually Exploring Coordination Strategy for LLM-based Multi-Agent Collaboration. *arXiv.org* (2024).

[71] Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems* 37 (2024), 68772–68802.

[72] Jeongeon Park, Bryan Min, Xiaojuan Ma, and Juho Kim. 2023. ChoiceMates: Supporting Unfamiliar Online Decision-Making with Multi-Agent Conversational Interactions. *arXiv.org* (2023).

[73] Sharoda A Paul and Meredith Ringel Morris. 2009. CoSense: enhancing sensemaking for collaborative web search. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1771–1780.

[74] Yingzhe Peng, Xiaoting Qin, Zhiyang Zhang, Jue Zhang, Qingwei Lin, Xu Yang, Dongmei Zhang, S. Rajmohan, and Qi Zhang. 2024. Navigating the Unknown: A Chat-Based Collaborative Interface for Personalized Exploratory Tasks. *International Conference on Intelligent User Interfaces* (2024).

[75] Plato, George Maximilian Anthony Grube, and J Cooper. 1981. *Five dialogues*. Hackett Publishing Company.

[76] Henry Prakken. 2006. Formal systems for persuasion dialogue. *The knowledge engineering review* 21, 2 (2006), 163–188.

[77] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833* (2022).

[78] Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S Weld. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634* (2024).

[79] Paul Scerri, David Pynadath, and Melind Tambe. 2002. Adjustable autonomy for the real world. In *Agent Autonomy*. Springer, 211–241.

[80] Sarah Schömbs, Yan Zhang, Jorge Gonçalves, and Wafa Johal. 2025. From Conversation to Orchestration: HCI Challenges and Opportunities in Interactive Multi-Agentic Systems. *ArXiv* abs/2506.20091 (2025).

[81] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[82] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548* (2023).

[83] Li Shi, Houjiang Liu, Yian Wong, Utkarsh Mujumdar, Dan Zhang, Jacek Gwizdka, and Matthew Lease. 2024. Argumentative experience: Reducing confirmation bias on controversial issues through llm-generated multi-persona debates. *arXiv preprint arXiv:2412.04629* (2024).

[84] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[85] Daniel Stokols. 2015. The transdisciplinary orientation scale: Factor structure and relation to the integrative quality and scope of scientific publications. *Journal of Translational Medicine and Epidemiology* 3, 2 (2015).

[86] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*. 1–18.

[87] 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*. 1–31.

[88] John Sweller. 2010. Cognitive load theory: Recent theoretical advances. (2010).

[89] Mike Thelwall and Kayvan Kousha. 2017. ResearchGate articles: Age, discipline, audience size, and impact. *Journal of the Association for information Science and technology* 68, 2 (2017), 468–479.

[90] Sunny Tian, Amy X Zhang, and David Karger. 2021. A system for interleaving discussion and summarization in online collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27.

[91] Stephen E Toulmin. 2003. *The uses of argument.* Cambridge university press.

[92] Keisuke Ueda, Wataru Hirota, Takuto Asakura, Takahiro Omi, Kosuke Takahashi, Kosuke Arima, and Tatsuya Ishigaki. 2025. Exploring Design of Multi-Agent LLM Dialogues for Research Ideation. (2025).

[93] Douglas Walton and Erik CW Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning.* State University of New York Press.

[94] Douglas N Walton. 2007. Dialog theory for critical argumentation. (2007).

[95] Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2023. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Neurocomputing* 618 (2023), 129063.

[96] Yuntao Wang, Yanghe Pan, Zhou Su, Yi Deng, Quan Zhao, Linkang Du, Tom H Luan, Jiawen Kang, and Dusit Niyato. 2025. Large model based agents: State-of-the-art, cooperation paradigms, security and privacy, and future trends. *IEEE Communications Surveys & Tutorials* (2025).

[97] Christopher J Ward, Susan B Nolen, and Ilana S Horn. 2011. Productive friction: How conflict in student teaching creates opportunities for learning at the

boundary. *International Journal of Educational Research* 50, 1 (2011), 14–20.
[98] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819* (2024).
[99] Klaus Weber, Annalena Aicher, Wolfang Minker, Stefan Ultes, and Elisabeth André. 2023. Fostering user engagement in the critical reflection of arguments. *arXiv preprint arXiv:2308.09061* (2023).
[100] Leslie Owen Wilson. 2016. Anderson and Krathwohl Bloom's taxonomy revised understanding the new version of Bloom's taxonomy. *The Second Principle* 1, 1 (2016), 1–8.
[101] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.
[102] Tongshuang Sherry Wu, Michael Terry, and Carrie J. Cai. 2021. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *International Conference on Human Factors in Computing Systems* (2021).
[103] Yi Wu. 2024. Critical Thinking Pedagogics Design in an Era of ChatGPT and Other AI Tools—Shifting From Teaching "What" to Teaching "Why" and "How". *Journal of Education and Development* 8, 1 (2024), 1.
[104] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
[105] Amy X Zhang, Michael S Bernstein, David R Karger, and Mark S Ackerman. 2024. Form-from: A design space of social media systems. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–47.
[106] Bo Zhang, Shi Feng, Xiangchao Yan, Jiakang Yuan, Zhiyin Yu, Xiaohan He, Songtao Huang, Shaowei Hou, Zheng Nie, Zhilong Wang, Jinyao Liu, Runmin Ma, Tianshuo Peng, Peng Ye, Dongzhan Zhou, Shufei Zhang, Xiaosong Wang, Yilan Zhang, Meng Li, Zhongying Tu, Xiangyu Yue, Wangli Ouyang, Bowen Zhou, and Lei Bai. 2025. NovelSeek: When Agent Becomes the Scientist - Building Closed-Loop System from Hypothesis to Verification. *ArXiv* abs/2505.16938 (2025).
[107] Gui-Min Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. 2025. G-Memory: Tracing Hierarchical Memory for Multi-Agent Systems. *ArXiv* abs/2506.07398 (2025).
[108] Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, et al. 2025. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *arXiv preprint arXiv:2505.00212* (2025).
[109] 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*. 1–24.
[110] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, et al. 2024. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226* (2024).
[111] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160* (2025).
[112] Yu Zuo, Dalin Qin, and Yi Wang. 2025. Large Language Model-Empowered Interactive Load Forecasting. *ArXiv* abs/2505.16577 (2025).

A Appendices

A.1 Pre-Session Survey

A.1.1 Basic Information.

- **Familiarity:** On a scale of 1 to 7, how familiar are you with the topic written in the initial proposal? (1: Not familiar at all, 7: I consider myself an expert in this domain)
- **Trust in GenAI:** Overall, which statement best describes your level of trust in generative AI (GenAI)?
  - Very low - I rarely or never use GenAI and doubt the accuracy of its information.
  - Low - I occasionally consult GenAI but remain skeptical of its reliability.
  - Moderate - I sometimes use GenAI; it can be helpful, but I still double-check its answers.
  - High - I frequently use GenAI, find it dependable, and believe it helps me solve many problems.
  - Very high - I actively rely on GenAI; I feel confident in its accuracy and usefulness across a wide range of tasks.

A.1.2 Self-Assessment Of Perceived Interdisciplinary Topic Clarity. On a scale of 1-7 (Strongly disagree to Strongly agree):

- **Conceptual Clarity:** I'm well-informed about the core concepts within the topic.
- **Methodological Clarity:** I know clearly the methods/approaches used within this field.
- **Role clarity:** I understand well who (which discipline/which colleague or expert) does what.
- **Communication clarity:** I feel confident explaining this topic to a mixed audience.

A.1.3 Self-Assessment of Proposal Quality. On a scale of 1-7 (Strongly disagree to Strongly agree):

- **Coverage:** The proposal clearly explains an important research gap and demonstrates a comprehensive understanding of prior work.
- **Significance and Novelty:** The proposed study offers an original contribution beyond existing solutions.
- **Relevance:** The content in the proposal is well-aligned with the original/proposed research idea.
- **Depth:** The research questions, design, data-collection, and analysis plan are described in sufficient detail to convincingly answer the stated aims.
- **Feasibility:** The research idea proposed is feasible.

A.2 Post-Survey Usability Questions and Ratings

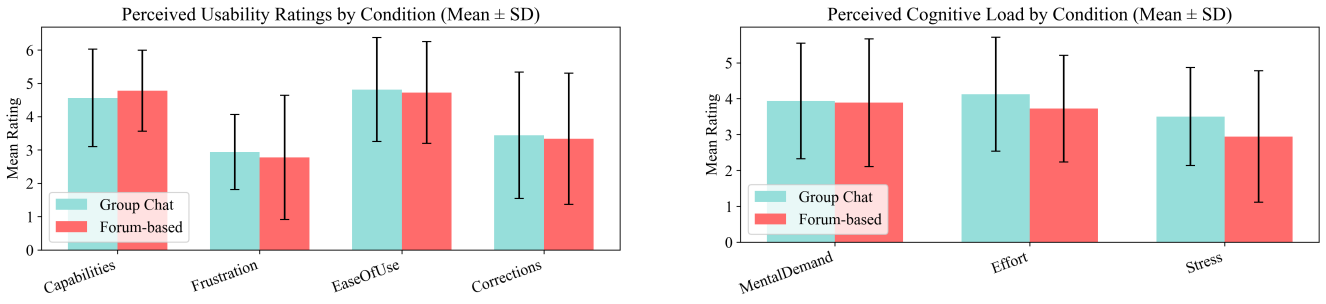
A.2.1 Post-Survey Usability Questions.

- **Capabilities:** The system's capabilities meet my requirements.
- **Frustration:** Using the system is a frustrating experience.
- **Ease of Use:** I thought the system was easy to use.
- **Corrections:** I have to spend too much time correcting things with this system.

A.2.2 Post-Survey Cognitive Load Questions.

- **Mental Demand:** How mentally demanding was the task?
- **Effort:** How hard did you have to work to achieve your goal?
- **Stress:** How irritated, stressed, or annoyed did you feel?

A.2.3 Post-Survey Usability and Cognitive Load Ratings. Detailed plots of post-survey proposal quality evaluation questions results as shown in fig. 13.



(a) Usability Ratings **(b) Cognitive Load Ratings**

Figure 13: Post-survey comparison of (a) usability and (b) cognitive load ratings between conditions.

A.2.4 Post-Survey Proposal Quality Evaluation Questions.

- **Coverage:** The proposal clearly explains an important research gap and demonstrates a comprehensive understanding of prior work.
- **Significance and Novelty:** The proposed study offers an original contribution beyond existing solutions.
- **Relevance:** The content in the proposal is well-aligned with the original/proposed research idea.
- **Depth:** The research questions, design, data-collection, and analysis plan are described in sufficient detail to convincingly answer the stated aims.
- **Feasibility:** The research idea proposed is feasible.
- **Clarity:** The research idea is clearly conveyed and easy to understand.

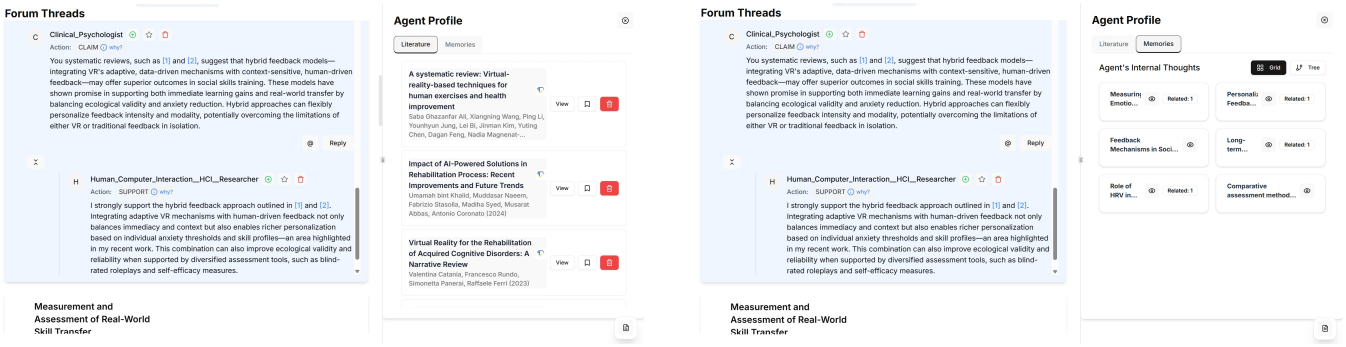
A.3 Figure of comparison of Post-Survey Feature Ratings

Figure 14 shows the comparison of post-survey feature ratings between conditions.

Figure 14: Overview of user ratings (overall helpfulness) for different system features (Surveys)

A.4 Agent Profile and Memory Interface Screenshots

Figure 15 shows the screenshots of agent profile and memory inspection interfaces.



(a) Agent Profile Interface **(b) Agent Memory Interface**

Figure 15: Agent inspection interfaces: (a) profile editor and (b) memory viewer.

A.5 Exit Interview Script (Semi-Structured)

- Can you walk me through your overall experience using the system to revise your proposal?
- Did you have any memorable moments while using the system? What were you doing right before, and what information or interaction triggered this moment?
- What specific information, questions, or suggestions from the agents were most helpful? Can you give an example?
- Did you find the different perspectives from the various personas useful? Why or why not?
- How do you feel about the utility/helpfulness of each feature (enumerate the feature to guide reflection)?

A.6 Persona Taxonomy and Construction

Schema. Persona profiles following the schema below:

```
basic_info:
research_area: ...
short_bio: ...
research_and_professional_focus:
focus_areas: ...
methodology: ...
publication_channels: ...
skills_and_expertise:
technical_skills: ...
analytical_skills: ...
domain_expertise: ...
personalities_and_characteristics:
communication_style: ...
audience_expertise_level: [novice|intermediate|expert]
```

A.7 User Interaction Log Codebook

Table 3: Codebook for User Interaction Logs.

Code	Definition (What it captures)	Facione Critical-Thinking Facet
clarify	Seeks meaning of terms, concepts, or context.	Interpretation
expand	Requests additional detail, depth, or breadth on a point already raised.	Explanation
apply	Seeks concrete application of ideas, theories, or methods.	Application & Inference
compare	Requests similarities, differences, or trade-offs.	Analysis
critique	Asks for judgment of merit, rigor, or limitations.	Evaluation
design	Requests creation or refinement of a study, system, or framework.	Synthesis & Application
method	Seeks specific methodological choices, measurements, or analyses.	Analysis & Application
data-seek	Explicit request for references, datasets, or historic evidence.	Knowledge
summarize	Asks to condense ideas or research into shorter form.	Explanation
alternative	Seeks different angles, methods, or solutions than those proposed.	Analysis & Evaluation
risk	Probes potential problems, risks, or ethical concerns.	Evaluation & Inference
reflect	Comments on or questions the thinking or research process itself.	Self-Regulation
ethics / impact	Considers moral, societal, or policy consequences.	Evaluation & Inference

B T18 Proposal Edits (Full)

It is worth noting that, T18's proposal edits during the forum condition involves stronger thinking and reasoning from themselves, whereas during the group chat the edits are mostly copy pasting appendix B.

T18's proposal edits using the group chat baseline: Human-Robot-Interaction

Motivation:

how to improve transparency of human-robot-interaction specifically in a human - multiple robots settings

Related Work:

human robot one on one interaction but not much about robot team [+] definition: <copied>transparency in robot-team collaboration is fundamentally about making both the processes and intentions of the robots comprehensible and visible to human team members.</copied>
[+] social perspective related work: SAT
[+] <copied>effective explanation interfaces must support "drill-down" capability, allowing a human to start with a high-level summary and then dig into details as needed [2].</copied>
[+] <copied>Another critical challenge as teams scale is maintaining mutual awareness—robots must keep track of not only their own state but the human's knowledge, attention, and goals, adapting their transparency accordingly. In dynamic, high-pressure scenarios like firefighting, traceability and meaningful human control need to be supported by both automated logging and live interactive explanations</copied>
[+] <copied>adaptive transparency At a sociotechnical level, transparency also involves cues like body orientation, signaling "who knows what" in the team, or highlighting moments where the robot's autonomy level is shifting. These are often not just about explanation content, but also timing, modality (verbal, visual, gestural), and invitation for user interaction—core aspects identified in both SAT and DARPA XAI outputs [2] [3]. I'd argue that future research should explore co-adaptive transparency, where both robots and humans shape the level and type of explanatory interaction together, influenced by ongoing context—a dynamic still underexplored in both interface design and team protocols. What do others see as necessary methodological advances to capture and refine this adaptivity in multi-robot teamwork?</copied>

Methods:

literature rievew, tool design, propotype development, evaluation
[+] <copied>system design: expose their internal decision-making processes (added point: transparency hinges not just on exposing internal state or logs, but on providing explanations that are actionable and adaptable to the context and user expertise) adapt their communication style to different users, humans should also be able to probe, question, and receive feedback in ways that make sense in context. </copied>
[+] ideas for overcoming the trade-off between detail and overload?
[+] <copied>senario design: From the qualitative side, I'd suggest deploying adaptive transparency in a hospital logistics context—imagine autonomous robots delivering sensitive supplies, where nurses and staff have varying expertise and urgency levels. Transparency would flex: routine deliveries need minimal prompts, while high-priority or unexpected events (e.g., medication reroute due to emergency) trigger detailed, user-tailored explanation (visual dashboard for seasoned staff; stepwise verbal walk-through for new users). Your user studies could use ethnographic observation and post-task interviews to explore not just trust and error rates, but also deeper aspects of social acceptability and staff perceptions of agency and accountability—key factors in organizational buy-in, as highlighted in sociotechnical and XAI deployments[1]. This offers rich ground for multi-method experiments! Curious what others think about capturing these "soft" social outcomes alongside quantitative measures.</copied>
[+] <copied>evaluation metrics: That's a fantastic take—I'd echo the critical value of "soft" outcome measures, especially in settings like hospitals where team cohesion and perceived accountability heavily influence deployment success. Ethnographic and mixed-methods approaches could reveal nuanced effects of adaptive transparency—for example, whether more granular explanations at critical moments foster not just trust, but also long-term user empowerment and error mitigation, as seen in XAI user studies[1][2]. For experimentation, you could instrument both quantitative metrics (task success, workload, response latency) and qualitative feedback (semi-structured interviews probing perceived clarity, fairness, team fit). This hybrid strategy is well-supported by XAI program guidelines and ensures the adaptivity isn't just technically sound, but genuinely aligns with user values and organizational culture. I'd encourage prototyping transparency "knobs" with multi-modal options (visual + verbal) and field-staging them—adaptive transparency is as much about fit to workflow and norm as explanation content. </copied>

Potential Outcomes:

more efforts should be put on how to show the coordination between the robot teammates, and level of this transparency should be carefully considered