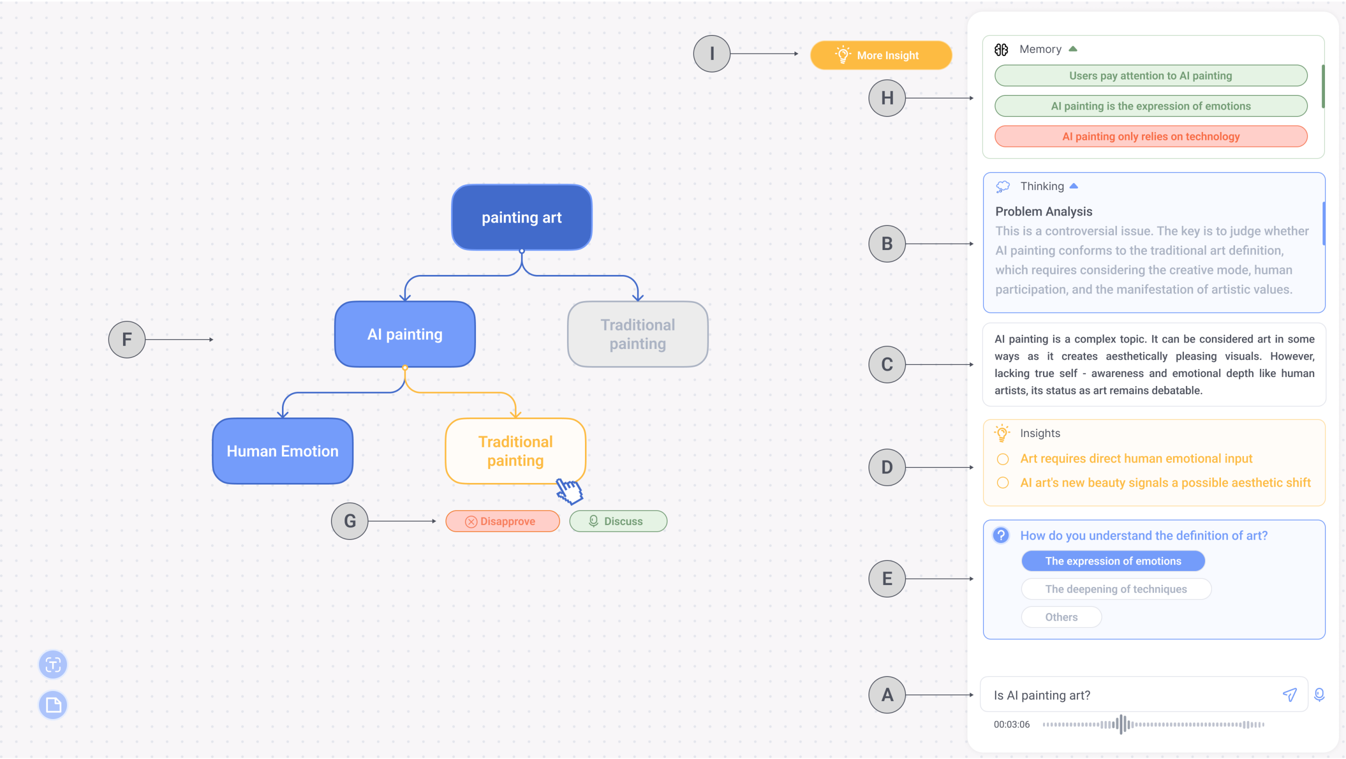
Towards Mutually Illuminative Collaborative Human–AI Deliberate Discussion

Anonymous Author(s)


I
More Insight
H
Memory
Users pay attention to AI painting
AI painting is the expression of emotions
AI painting only relies on technology
Thinking
B
C
D
Insights
E
A
F
G
painting art
AI painting
Traditional painting
Human Emotion
Traditional painting
Disapprove
Discuss
The expression of emotions
The deepening of techniques
Others
00:03:06

Problem Analysis

This is a controversial issue. The key is to judge whether AI painting conforms to the traditional art definition, which requires considering the creative mode, human participation, and the manifestation of artistic values.

AI painting is a complex topic. It can be considered art in some ways as it creates aesthetically pleasing visuals. However, lacking true self - awareness and emotional depth like human artists, its status as art remains debatable.

○ Art requires direct human emotional input

○ AI art's new beauty signals a possible aesthetic shift

How do you understand the definition of art?

Is AI painting art?

Figure 1: Echo's interface. Users input questions/ideas via voice (A). Echo first outputs initial thoughts (B), provides an answer (C), and generates inspiration points (D). In this example, to align on the definition of art, Echo generates a Question Card (E) for quick user interaction and alignment. User interactions with inspiration points and cards are reflected in the Dynamic Thought Tree (F) on the left and stored by the Memory module (H) to learn preferences. Users interact with the Dynamic Thought Tree (F); hovering over a node reveals buttons (G) to disagree or initiate a focused discussion on that node for precise context control. Clicking the inspiration button (I) prompts Echo to find the most valuable node and generate inspirational child nodes for it.

Abstract

LLMs as discussion partners are often felt to be misaligned with human intentions and lack proactive engagement. This limits their ability to support deliberate discussions involving critical thinking, diverse perspectives, and iterative reasoning. To understand the key elements needed for human-AI deliberate discussions, we conducted a formative study (N=8) comparing human-human and human-LLM discussions, identifying user expectations for LLMs: alignment on semantics, viewpoints, and logic, as well as inspiration for thought. Based on these findings, we developed *Echo*, a real-time viewpoint alignment system through two core interaction components. Question Cards support AI questioning and rapid user alignment with the AI's understanding; the Dynamic Thought Tree visualizes and tracks this process. The user study (N=16) demonstrated that *Echo* significantly enhanced users' deep thinking abilities and the generation of creative ideas, while providing a more natural interaction. Furthermore, we found that human-AI deliberate discussion relies on synergy between model capability and interaction design. By systematically defining human-AI collaborative deliberate discussion and its key elements, this research offers design guidance for developing novel systems that foster mutual human-AI inspiration.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Artifact or System; Visualization; Human-AI Collaboration; Interactive AI Alignment; Deliberate Discussion; Large Language Models

ACM Reference Format:

Anonymous Author(s). 2018. Towards Mutually Illuminative Collaborative Human–AI Deliberate Discussion. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Viewpoint discussion is a critical process for promoting critical thinking, deepening understanding, and sparking innovation. Participants must continuously examine, reflect on, and adjust perspectives to reach deeper consensus or discover new possibilities [24]. Such discussions typically rely on two key elements: Alignment of Intent, and Mutual Inspiration. The former requires participants to reach consensus on the core issues of discussion, concept definitions, and logic, ensuring communication shares the same context while avoiding misunderstandings and ineffective dialogue [33]. Secondly, discussions need effective Mutual Inspiration, where participants can introduce different perspectives, question assumptions, or supplement information, thereby challenging existing thinking frameworks and pushing the discussion to deeper, broader levels, avoiding stagnation [24].

In recent years, LLMs as conversation partners have shown enormous potential in information retrieval and content generation. However, current LLMs discussion interaction modes tend toward passive response, meaning AI primarily generates replies that meet user needs based on their input, while rarely proactively raising questions or guiding users to reflect [25, 37, 41]. Without effective interaction mechanisms, LLMs often struggle to accurately capture users' subtle intentions, viewpoint positions, or logical emphases, resulting in responses that, while fluent, may be disconnected from users' true thoughts, or even provide answers that are difficult to understand or overly accommodating [32]. Additionally, the transparency and controllability of LLMs' generation process is poor, making it difficult to guide LLMs to focus on specific discussion points or make local adjustments [9, 35]. More importantly, existing dialogue interfaces limit discussion flexibility, making it challenging for users to conduct global reviews or compare viewpoints during complex, non-linear thinking processes [20, 39]. These limitations make LLMs more like information retrieval tools or draft generators rather than discussion partners capable of in-depth exploration and mutual inspiration.

To deeply explore the types of intelligence needed for current human-AI deliberation and provide a basis for system design, we first conducted a critical formative study (N=13). The core of this study was to directly compare two discussion scenarios: (1) users discussing an open-ended, thought-provoking topic (e.g., "Is cultivating a complete personality more important than practical skills?") with the advanced LLM Deepseek-R1; (2) the same user discussing another similar topic collaboratively with another human participant (in our experiment, this human partner could use various tools to assist thinking and expression). During this process, we encouraged participants to think aloud. Through comparative analysis (human-AI vs. human-human), we found that current LLM interactions fail to meet user expectations for alignment (covering semantic, viewpoint, and logical) and thought inspiration, while also showing significant deficiencies in user controllability and global exploration. These key findings directly established our design goals: to build interaction mechanisms aimed at achieving deep alignment and inspiration, enhancing user control, and supporting exploration from a global and local perspective. Based on these findings, we developed *Echo*, a system that achieves real-time viewpoint alignment through two core interaction components. Among them, "Question Card" support AI in asking proactive questions and help users quickly align with AI's understanding; the "Dynamic Thought Tree" visualizes and tracks this process.

In a user study (N=16), we used the *Echo* system as an experimental probe, comparing it with a baseline that combined standard chat interface and whiteboards. This not only confirmed *Echo*'s advantages in significantly enhancing users' deep thinking abilities, creative idea generation, and interaction experience, but also revealed a deeper principle: achieving high-quality human-AI deliberation depends on the synergistic effect between LLM model capabilities and interaction design. Our findings suggest that only by integrating the two, rather than simply relying on either one, can we truly unlock potential and achieve insightful in-depth discussions.

The main contributions are as follows:

(1) **Clarifying challenges and expects**: By comparing human-AI and human-human discussions, we identified the core challenges faced by current LLMs in deep discussions and clarified users' key needs for deep alignment and thought inspiration.
(2) **Interactive system**: We proposed the *Echo* system and its two core components (Question Cards and Dynamic Thought Tree) as a novel interactive solution to promote alignment and inspiration in discussions.
(3) **User study and revealing principles**: Through user study, we confirmed that the synergy between interaction design and model capability is key to achieving high-quality human-AI deep discussions.

2 Related Work

Our work builds on previous research on Human-AI Deliberation, Human-AI Alignment, and Mutual Inspiration in Human-AI Collaboration.

2.1 Human-AI Deliberation

human-AI deliberation refers to an interactive pattern in which humans and AI systems analyze problems, exchange viewpoints, and reach decisions through structured dialogue and discussion processes [24]. This model aims to address two key challenges in traditional AI-assisted decision-making systems: users' lack of in-depth analytical thinking about AI recommendations and the absence of effective communication and coordination mechanisms when human-machine perspectives are misaligned [33].

Specifically, existing systems often provide static recommendations that users can only passively accept or reject, failing to fully utilize the knowledge and insights of both humans and AI [18]. AI transfers knowledge to humans by providing recommendations and explanations [26], which should promote the synergistic integration of human and AI intelligence, but when viewpoints diverge, interaction interfaces typically lack effective mechanisms to support human-AI knowledge exchange. Furthermore, human dependence on AI recommendations seriously affects decision quality [24, 31]. Successful decision making depends on humans being able to wisely judge when to adopt AI recommendations and when to remain skeptical [24, 40]. Both over-reliance and insufficient trust can lead to adverse consequences, while the ideal state is for humans to decide whether and how to rely on AI recommendations based on specific circumstances [16, 17].

Therefore, Liu et al. proposed human-AI Deliberation, dividing the process into three components: dimension-level opinion guidance, negotiation discussion, and decision updates [24], corresponding to evaluating Deliberation capabilities: (1) Dimensionalized opinion expression: supporting users in questioning or supplementing evidence for specific decision dimensions; (2) Transparency in reasoning processes: AI needs to explain the generation logic and evidence sources; (3) Dynamic adjustment mechanisms: allowing both parties to correct initial positions during discussions [5, 24]. Compared to traditional instructional or question-answer interactions, negotiation-style interaction emphasizes the balance between human subjectivity and AI assistance, and mutual learning and adjustment during the process [21].

The most important aspect of human-AI deliberation is deeply understanding and empowering users, enabling them not only to judge the quality of AI recommendations but also to actively contribute their own knowledge, transforming from passive recipients of information to active co-creators in the decision-making process. However, to ensure that AI can continuously and reliably assist humans in broader scenarios and align with human overall interests and values, we need to address a deeper issue: Human-AI Alignment.

2.2 Human-AI Alignment

In the HCI field, alignment means ensuring that AI systems' behaviors, reasoning, and goals are consistent with human's values, intentions, and expectations, which is crucial for discussions requiring understanding and collaboration, as semantic, viewpoint, or logical deviations can hinder mutual inspiration [12, 13, 30].

Current research has proposed various frameworks. The "bidirectional Human-AI alignment" framework proposed by Shen et al. emphasizes the mutual adaptation process between humans and AI, meaning not only does AI need to align with humans, but humans also need to understand and adapt to AI [33]. Terry et al. mapped alignment to an interaction cycle, proposing three core goals for interactive AI alignment: Specification Alignment, ensuring users can effectively communicate goals; Process Alignment, providing mechanisms to verify and control AI execution processes; and Evaluation Support, for measuring alignment quality [38]. Additionally, identifying and evaluating potential viewpoint conflicts in AI systems is an important aspect of alignment research, with frameworks like ValueCompass providing methods for this purpose [34]. In practice, current Human-AI disagreements mainly stem from different interpretations of information [1], conflicts in values [13], or differences in reasoning paths. For example, Guo et al.'s research shows that when AI assistants' values are not aligned with users, it increases users' frustration in creative collaboration [13]. Meanwhile, "Shared Interest" proposed by Boggust et al. provides a measurement method, identifying decision differences by comparing the focus of attention between humans and AI [1].

Simultaneously, research trends are toward interactive alignment methods, encouraging users' active participation. This includes collaboratively optimizing models through interactive annotation (such as MOCHA [7]), allowing users to directly convey concepts through dialogue [3], leveraging human preference feedback to enhance LLM alignment [36], and aligning explainable AI outputs [19]. These methods empower users to shape AI behavior.

Therefore, effective human-machine alignment transcends static goal setting, trending toward dynamic, reciprocal interactive calibration. In human-AI deliberation scenarios, establishing real-time, bidirectional semantic, viewpoint, and logical alignment mechanisms is essential for achieving deeper mutual inspiration.

2.3 Mutual Inspiration in Human-AI Collaboration

When humans and AI can effectively align semantics, viewpoints, and logic, ideal collaboration should transcend simple task allocation or information retrieval, entering a synergistic state where both parties can inspire each other to generate new insights and ideas [11, 14, 28]. This is particularly important for human-AI deliberation that requires critical thinking and creative generation.

For example, the Supermind Ideator system helps users generate more innovative ideas than when using ChatGPT alone or working independently by providing structured scaffolding for the creative generation process between humans and AI [14]. The COFI framework also emphasizes that to design AI partners that can inspire, interaction dynamics and specific behavioral characteristics are crucial [28].

Achieving this mutual inspiration largely depends on carefully designed interaction mechanisms:

Structured & Guided Interaction: As demonstrated by Supermind Ideator [14], setting collaborative processes and steps (such as iterative questioning, idea combination) can systematically guide users to think deeply. Meanwhile, the AI's initiative level is a delicate balance point: while AI that follows human leadership is generally more acceptable, timely and contextually appropriate AI-initiated questions or different viewpoints may also break thought patterns and inspire new thinking directions [22, 23].

Multi-Perspective Presentation & Exploration: Enabling users to examine problems from different angles is an important way to inspire. Marvista uses natural language processing to support multi-perspective analysis of 3D models, helping human analysts understand complex information from different angles, thereby inspiring new insights [4]. The Scholastic system supports researchers in exploring data in inductive and interpretive research through graphical interfaces and familiar metaphors, inspiring new ideas [15].

Transparency & Communication: Increasing the transparency of AI behavior, for example through clear behavioral descriptions, can help humans understand AI's strengths and limitations, identify its potential failures, and more appropriately adjust their reliance on it, thereby building better collaborative trust and laying the foundation for mutual inspiration [2].

Temporal Dynamics Management: Especially in long-term chat-based collaborations (such as using LLMs), the rhythm, timing, and pace of information exchange may also affect the fluency of creativity and the emergence of inspiration [29].

Therefore, promoting mutual inspiration in human-machine collaboration does not merely rely on enhancing the capabilities of AI models themselves, but more importantly requires building a collaborative environment that promotes insights through thoughtful interaction design based on human-AI alignment. Meanwhile, AI needs stronger understanding, adaptation, and learning capabilities, and relies on real-time dynamic alignment.

3 Formative Study

To deeply understand the key elements and challenges of human-AI deliberate discussions, we designed a comparative study exploring the differences between human-human dialogue and human-AI dialogue in viewpoint discussions. We particularly focused on how participants express, understand, and align complex viewpoints, and how they inspire each other during discussions.

3.1 Study Design

3.1.1 Comparative Study. We recruited 8 participants (S1-S8) through social media and personal networks, aged 23-26 years, including 5 males and 3 females, all graduate students in integrated circuits, artificial intelligence, HCI, and information art design. We collected data on participants' frequency of AI use through questionnaires, ranging from rarely using AI (several times per week) to frequently using AI (more than 3 times daily). The diverse participant backgrounds and AI usage experience helped us identify common challenges and expectations in human-AI discussions. The core of our research compared two collaborative discussion scenarios:

Each participant completed two comparative tasks (Figure 2):

Human-AI Condition: Participants played the role of requesters, collaborating with the Deepseek-R1 model (playing the role of executor) through a standard chat interface (based on Yuanbao platform) to discuss an open-ended topic (e.g., "Which is more important: cultivating a complete personality or practical skills?") and complete an argumentative essay outline.

Wizard-of-Oz Condition: Participants similarly played the role of requesters, discussing another similar topic (e.g., "For major life decisions, should one trust rational judgment or intuitive sixth sense?") with a human agent (human wizard) who played the role of an idealized collaborator, and worked together to complete an outline. This Wizard was played by a researcher, aiming to simulate an ideal discussion partner with strong understanding and active cooperation.

All topics were selected to be highly subjective, requiring deep thinking and having no standard answers. In both conditions, we used think-aloud protocols to capture participants' real-time thinking processes. Additionally, we provided iPad Procreate as a completely free visual note-taking tool. Participants could use it to draw anything (such as mind maps, keywords, sketches), and we aimed to observe their unstructured thinking trajectories and spontaneously adopted visual organization preferences (such as hierarchical or network structures) through screen recording.

We recorded: (1) participants' screen activities, mainly interactions with AI; (2) participants' verbal expressions and think-aloud content; (3) dynamic Procreate notes; (4) participants' reported aha moments (moments of inspiration) and their causes; (5) final output outlines. After the experiment, we conducted 20-minute semi-structured interviews with each participant, exploring their thinking processes, perceptions of mutual understanding, and factors promoting alignment. We identified recurring themes through thematic analysis and analyzed the structural features of visual notes to identify participants' specific needs and challenges in achieving deep alignment and gaining inspirational thinking.

3.1.2 Expert Interviews. To supplement findings from our comparative user study, we conducted semi-structured interviews with five HCI experts. These experts included one professor, one postdoctoral researcher, and three PhD students, whose backgrounds are detailed in Table 1. All were experienced AI users, employing tools like ChatGPT, Claude, Gemini, and Deepseek over 30 times daily for advanced tasks (e.g., writing, coding) in their research and work.

Table 1: Background of the Interviewed HCI Experts.

ID	Research Area	Age	Role / Experience
E1	Interactive Learning	40	Field expert
E2	Cognitive Interaction	34	Postdoctoral Researcher
E3	PBL	26	PhD Student (HCI)
E4	AI+HCI	26	PhD Student (HCI)
E5	CoT Interaction	27	PhD Student (HCI)

The main goal was to understand experts' challenges and pain points when using AI, focusing on three dimensions crucial for deep discussion:

- **Requirement Expression:** How effectively can experts convey complex or evolving intentions to AI? What hinders communicating the desired focus, scope, or perspective?
- **Reasoning Process Experience:** How do experts perceive the AI's reasoning or generation process? Is this process transparent or controllable? Which interaction patterns help or hinder the discussion?
- **Output:** How well does the AI output's structure and format support user understanding, critical assessment, and further ideation?

Besides discussing general experiences, we asked each expert to describe a recent case of using AI for tasks like writing, focusing on specific pain points during the interaction. Each interview lasted about 30-45 minutes. We planned to analyze the transcripts

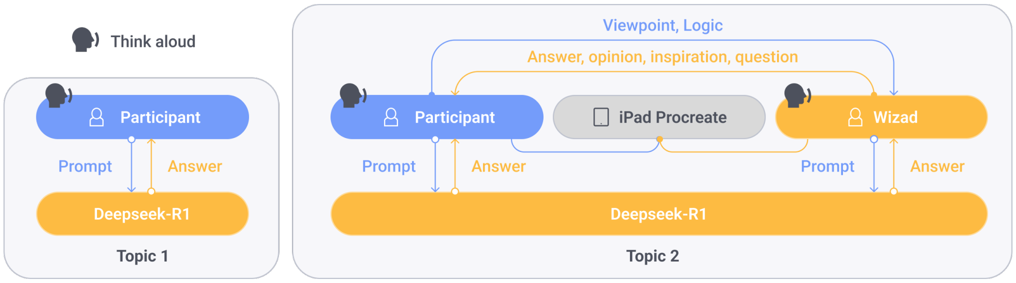


Figure 2: The formative study's comparative experiment prompted participants to think aloud with the Wizard.

using thematic analysis to identify expert insights on current AI limitations.

3.2 Findings

3.2.1 Human-AI Deliberation Challenge. For the 8 participants with different backgrounds, we observed and inquired about the difficulties they encountered during discussions, and summarized the results as follows:

C1: Insufficient deep understanding of user intentions and viewpoints. AI often struggles to understand user needs, especially in the early stages of discussion when there is less conversational context. For example, *"When I first interacted with R1, I just wanted to hear the AI's thoughts, but it directly gave me an article"* (S4, M, 26)), *"I wanted to know how to alleviate loneliness caused by social media, but AI gave me common psychological causes of loneliness without incorporating social media"* (S6, M, 24). Additionally, some AI responses were difficult for users to understand due to their abstractness. When responses included unfamiliar terms or concepts to users (such as quantum mechanics, three-spoon theorem (S3, M, 24)), and AI failed to provide corresponding explanations, the mismatch between the user's discussion context and AI output led to significant understanding difficulties.

C2: Poor controllability of AI output. AI generation is end-to-end, but the discussion process is actually *"a process of users convincing themselves"* (E2, M, 34). Merely generating direct results is difficult for users to accept and understand, and also fails to achieve capability training during the discussion process, unable to reflect the deliberative process. Users indicated that agents like Deepseek sometimes over-interpret user needs, resulting in generated content with many parts not currently needed by users. The system tends to generate redundant information beyond what is required for the current dialogue stage, increasing users' cognitive burden in modifying and correcting AI outputs, making it difficult to control its output. For instance, E1 (M, 40) (M, 40) stated: *"Deepseek-R1 is useful, but difficult to modify locally."*

C3: Lack of constructive questioning and multiple perspectives. AI tends to confirm or simply restate users' viewpoints rather than proactively raising challenging questions, opposing opinions, or introducing different angles of thinking at appropriate times, which limits the possibility of discussions developing to deeper levels. *"Sometimes I don't have ideas either, like whether sixth sense is more important or rationality is more important, because I've experienced both, I want to know what else there is, but AI can't do that, it even follows my previous meaning and starts affirming me, which is useless to me"* (S1, M, 24).

C4: Sequential dialogue interfaces limit global understanding and deep reflection. Current AI dialogues use sequential interfaces where information is presented in a time flow. This conflicts with the thinking processes required for deep discussion (cross-topic backtracking, comparing, merging different viewpoints, etc.). Due to limited human attention and memory [8, 10], users find it difficult to macroscopically combine context and past discussions for reflection in such interfaces. In long conversations, users easily forget key points or find it difficult to backtrack, making higher-level reflection or integration challenging [42]. *"Discussing this matter is globally diffusion, locally a transformer model"* (E3, M, 26) - such interfaces struggle to support the diffusion-style thinking necessary for establishing macroscopic perspectives and connecting viewpoints.

3.2.2 Human-AI Alignment: Semantics, Viewpoints, and Logic. Our formative study revealed that the challenges **C1** and **C2** stem from a lack of shared understanding in current human-AI interaction. Effective deep discussion requires constructing shared meaning, not just transferring information. This means the AI must understand the user's underlying intent, not just surface-level text. We identified three key layers for ideal alignment:

Semantic Alignment: Core concepts discussed must be clearly defined. When a user introduces a term (e.g., social media, sixth sense), misalignment in understanding can derail the discussion or burden the user. In all 8 Wizard-of-Oz (WoZ) sessions, the Wizard clarified key concepts early, asking questions like *"How do you understand social media?"* or *"What does sixth sense mean to you?"*. This shows that defining semantics is the first step towards a shared basis for discussion and helps prevent the mismatch between AI responses and user questions noted in **C1**.

Viewpoint Alignment: Beyond concepts, the user's stance and existing judgments are important context. The AI needs to understand the user's current perspective and avoid incorrect assumptions. S4 (M, 26) noted: *"Before discussing with the AI, I already had my own judgment about... but the AI didn't ask me. The wizard, however, directly asked about my stance."* Actively seeking the user's viewpoint makes AI responses more targeted and helps the user feel understood.

Logical Alignment: The deepest level involves understanding the user's reasoning process and the motivation behind their questions [6]. It's not enough to grasp *what* the user said; the AI needs to understand *why* they said it and *how* they moved from one point to the next. S1 (M, 24) mentioned: *"Discussing with a person is different because a person not only hears what I say but can also think about *why* I'm saying it..."*. In the WoZ experiments, the Wizard frequently asked about the deeper reasons behind a user's question. Exploring this reasoning logic is crucial for grasping the user's true intent and is fundamental to overcoming the issue of irrelevant AI outputs (**C2**).

Therefore, achieving human-AI alignment in deep discussions requires moving beyond simple surface-level matching to delve into these three interconnected layers: semantics, viewpoints, and logic.

3.2.3 The Value of Inspiration: Stimulating Thought. Our formative study also highlighted current AI's limitation in fostering inspiration (**C3**). Users seek not just an information provider but a thought partner that stimulates thinking and broadens perspectives [27]. We found effective inspiration involves more than just providing information or agreeing:

Connecting and Expanding Ideas: An ideal partner helps users see new connections between ideas or suggests overlooked angles. In the WoZ study, the Wizard was praised for inspiring users by directly linking different nodes on their visual notes (iPad Procreate), suggesting new thought paths. S3 (M, 24) also valued the Wizard for *"pointing out omissions"*.

Constructive Questioning and Different Perspectives: True inspiration often arises from moderate challenges and diverse viewpoints. S6 (M, 24) explicitly wanted the AI to*"refute me, or even offer opposing views,"* not just answer questions. S3 (M, 24) also saw *"refutation"* as vital for writing and deep thinking. This indicates users want AI to offer constructive disagreements based on understanding their stance, thus sparking deeper reflection and avoiding cognitive fixation.

Furthermore, alignment and inspiration are not mutually exclusive but synergistic. Alignment does not mean AI merely mimics the user; it requires deep understanding (as discussed in Section 3.2.2). Only when the AI truly grasps the user's semantics, viewpoint, and logic can its inspiration — whether suggesting new links, supplementing omissions, or offering counterarguments—precisely target the user's current line of thought and effectively spark new insights. Inspiration without good alignment risks being superficial, off-topic, or confusing. As S5 (F, 26) experienced, the Wizard's requests for clarification to ensure understanding (an alignment process) sometimes prompted the user to reflect more deeply on their own questions. Therefore, high-quality alignment is the foundation for effective inspiration.

3.3 Design Goals

Based on the challenges in current human-AI discussions identified in our formative study (**C1-C4**), user expectations, and related work, we established the following design goals (DGs) for the *Echo* system:

DG1: Promote Real-time Understanding Alignment. To address the AI's difficulty in deeply understanding user intent and viewpoints (**C1**) and the resulting output deviation or lack of control (**C2**), the system should include real-time, interactive mechanisms for alignment. This needs support across multiple layers: semantic, viewpoint, and logical.

DG2: Enable Fine-grained Control and Structured Exploration. Addressing the difficulty in locally modifying AI output (**C2**) and the user need to switch between discussion granularities, the system must offer more flexible control. Users should be able to focus locally, instructing the AI to respond only to specific parts of the current discussion, avoiding irrelevant context (addressing S3's desire for local adjustments and S6's wish for the AI to know *"where we are discussing now"*). The system should also support flexible context management, allowing users to isolate discussion focus when needed (like S6 (M, 24) wanting a *"new dialogue section"*) or specify the AI's scope interactively (e.g., like S8 (F, 23) circling on the iPad). Ideally, the system could even help guide user attention to key points (as S7 (M, 24) observed the Wizard doing), ensuring both parties stay on the same page.

DG3: More Actively Foster Mutual Inspiration. To overcome the tendency of LLMs to simply agree or provide generic responses (**C3**), the system should enable the AI to act as a more proactive discussion partner. This means encouraging the AI to constructively introduce diverse perspectives, ask relevant questions that prompt deeper thinking, and challenge assumptions when appropriate.

DG4: Support Global Overview and Reflection via Nonlinear Interaction. To tackle the difficulty of reviewing and thinking macroscopically within linear chat interfaces (**C4**), the system should provide non-linear interaction and visualization mechanisms. This aims to visualize the discussion structure, presenting the process and viewpoints clearly in a non-linear way (similar to S2 (F, 24) describing a debate team using a blackboard), helping users grasp the overall picture and understand relationships between ideas. It should also support users in freely backtracking, jumping, comparing, and integrating ideas from different stages, better accommodating divergent and convergent thinking. Ultimately, by offloading some recording and organization burden to the system (as S3 (M, 24) felt the Wizard's note-taking reduced cognitive load), users can focus more on high-quality thinking and creativity.

4 System Design

Based on our design goals, we developed *Echo*, a system designed to facilitate Illuminative Collaborative Human–AI Deliberate Discussion, helping users achieve deeper alignment and mutual inspiration with AI. *Echo* uses two core components—Question Cards and a Dynamic Thought Tree—to enable real-time alignment across semantic, viewpoint, and logical layers, and to stimulate user thinking. This section presents *Echo*'s interaction flow through a user scenario, details the key designs of Question Cards and the Dynamic Thought Tree, and explains the inspiration generation algorithm (Figure 3).

4.1 User Scenario

Imagine a user, Alex, initiates a discussion via voice: *"Is AI painting art?"*. *Echo* first presents initial thoughts in the chat area while building the initial structure of the Dynamic Thought Tree on the canvas, marking key dimensions (e.g., "Tool vs. Expression"). Then, *Echo* uses Question Cards to quickly probe and align Alex's initial viewpoint (e.g., believing AI lacks emotional expression). Alex's

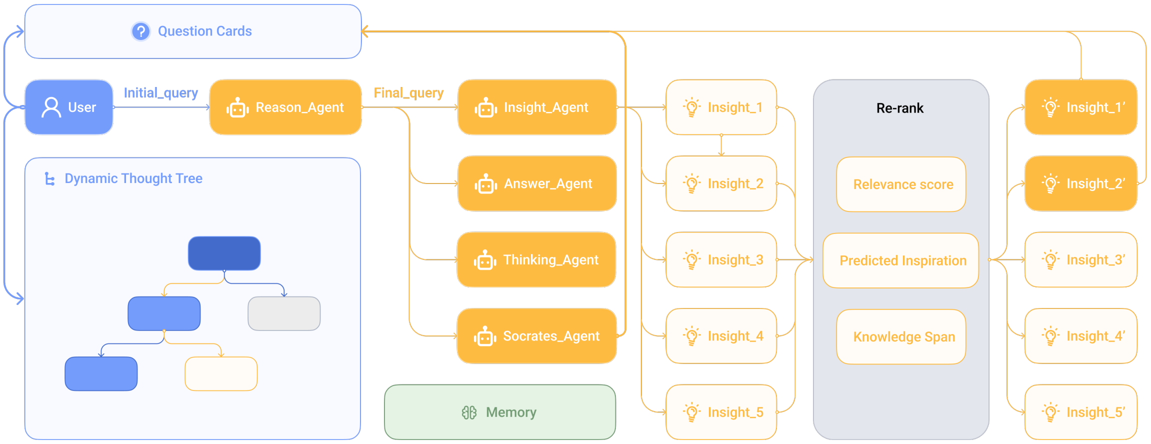


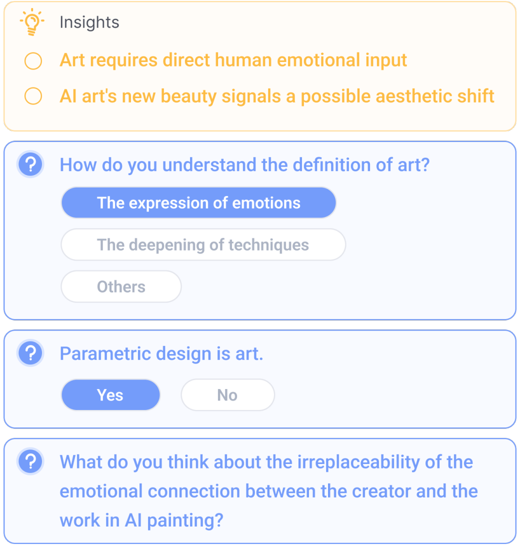
Figure 3: Echo interaction logic diagram. Users primarily interact with Question Cards and the Dynamic Thought Tree. The diagram shows the inspiration algorithm logic: 5 generated candidate inspirations are re-ranked using Relevance Score, Predicted Inspiration, and Knowledge Span.

choice instantly adds a new node to the thought tree, achieving viewpoint alignment. As Alex adds details (e.g., the importance of human involvement), the tree updates accordingly with new nodes and connections. To delve into a specific branch (e.g., "AI painting as a tool"), Alex can directly interact with the corresponding node on the tree via voice. The AI responds focused on that point and updates the tree, enabling structured exploration (**DG2**). During the discussion, Alex can request inspiration (**DG3**). *Echo* analyzes the entire tree, suggests inspirational points connecting different ideas, and stores useful ones in long-term memory. Throughout this process, the Dynamic Thought Tree provides a global overview of the discussion (**DG4**), supporting Alex in reviewing, reflecting, and easily managing the discussion flow (e.g., pruning branches).

4.2 Interface

Echo's interface consists of two main areas (Figure 1): the Canvas Area on the left displaying the Dynamic Thought Tree, and the Chat Area on the right integrating Question Cards. These core components work together to foster human-AI alignment and inspiration.

4.2.1 Question Cards. Appearing in the right Chat Area, Question Cards are a key interaction mechanism in *Echo* for alignment and inspiration. Based on formative study needs, we designed various card types (e.g., inspiration, choice, judgment, rhetorical questions (Figure 4)). AI responses are accompanied by these cards, using quick interactions (e.g., clicking options) to probe user views and clarify semantics, thus efficiently promoting understanding alignment (**DG1**). User interactions provide clear feedback to the AI (the system learns user tendencies, not just copies) and are instantly reflected on the Dynamic Thought Tree (e.g., creating nodes representing consensus). Cards, especially inspiration and rhetorical ones, also serve to foster mutual inspiration (**DG3**), guiding users toward deeper thought or new perspectives.



Insights

○ Art requires direct human emotional input

○ AI art's new beauty signals a possible aesthetic shift

How do you understand the definition of art?

● The expression of emotions

○ The deepening of techniques

○ Others

Parametric design is art.

[✓ Yes] No

What do you think about the irreplaceability of the emotional connection between the creator and the work in AI painting?

Figure 4: Question Card types include: Inspiration, Choice, Judgment, and Probing Questions.

4.2.2 Dynamic Thought Tree. Located in the left Canvas Area, the Dynamic Thought Tree is central to visualizing the discussion process. It automatically tracks and presents viewpoints and their logical connections as nodes and links, offering users a global overview and supporting non-linear review and reflection (**DG4**). The tree's structure dynamically evolves based on AI suggestions, user card interactions, and inputs. Users can achieve fine-grained contextual control (**DG2**) by interacting with specific nodes (e.g., hovering and using voice input) to focus the discussion locally, or input globally. Users can also prune nodes to manage the discussion flow. The tree's automatic layout (using Dagre.js) ensures visual clarity and reduces user cognitive load.

4.3 Implementation Detail

We implemented *Echo* using FastAPI for the backend and React for the frontend. The canvas utilizes React Flow. To ensure responsiveness (avg. 297 tokens/s), we used the qwen-turbo model.

4.3.1 Query Rewriting and Inspiration Algorithm. To generate effective inspiration points, *Echo* employs an algorithm combining query rewriting and multi-dimensional evaluation (Figure 3). Upon receiving a user query (*initial_query*), a Reason Agent analyzes its connection to the previous query and infers the user's deeper intent using preferences stored in the memory module (e.g., inferring interest in human role during an AI painting discussion). Based on this inference, the system rewrites *initial_query* into a *final_query* (e.g., "Copyright?" might become "Discuss copyright issues concerning human creators, AI developers, and the work itself in AI painting, including ethical considerations."). The *final_query* is sent to the LLM to generate 5 candidate inspirations (*candidate_inspirations*). These candidates are then scored and re-ranked by a more powerful evaluation model (Deepseek-v3 in our system). The evaluation uses a weighted scoring function S, considering three key dimensions:

$$S = w_1 \cdot \text{RelScore}(\text{SemGap}) + w_2 \cdot \text{PredInsp} + w_3 \cdot \text{SpanScore} \quad (1)$$

where:

1. **RelScore(SemGap)**: Relevance score. SemGap is the cosine similarity between candidate and *final_query* embeddings. Higher score means more semantically relevant.
2. **PredInsp**: Predicted Inspiration. A score [0, 1] from a heuristic model estimating the likelihood of sparking new thoughts based on candidate specificity and user history.
3. **SpanScore**: Knowledge Span Appropriateness Score. Provided directly by the Deepseek-v3 evaluation model. It assesses if the candidate's knowledge domain is relevant and appropriately scoped relative to the current context (*candidate_inspiration*, *final_query*, context summary). Candidates with large knowledge leaps or irrelevant associations (like the confusing quantum mechanics example from the formative study) receive low scores (near 0), while well-aligned, moderately expansive ones get high scores (near 1).

w_1, w_2, w_3 are fixed weights, set empirically based on formative study findings and preliminary tests, balancing relevance, inspiration potential, and knowledge scope suitability. Finally, the system ranks the 5 candidates by score S and presents the top 2 as inspiration cards to the user.

4.3.2 Socratic Question Cards. The design of Question Cards adopts Socratic questioning, particularly its core idea of elenchus (refutation). This method uses persistent, probing questions to uncover underlying assumptions, vague concepts, and logical inconsistencies in the user's viewpoint, directly addressing the challenges of shallow AI understanding (**C1**) and lack of constructive questioning (**C3**) found in our study. Instead of providing direct answers, Socratic questioning guides users toward self-examination and deeper thinking. Specifically, we use prompts to guide the LLM to generate multiple-choice, true/false, and exploratory question cards in a Socratic style. These cards function by:

Promoting deep alignment (DG1): They actively probe *"What exactly do you mean by X?"* (clarifying semantics), *"What is your core stance?"* (defining viewpoint), *"What is the premise of this argument?"* (examining logic), overcoming biases noted in **C1**.

Stimulating active inspiration (DG3): This method naturally encourages the AI to take a more active, moderately challenging role, moving beyond the passive information delivery observed in **C3**. By questioning assumptions and probing reasons, the AI effectively guides the discussion deeper and sparks new user insights.

4.3.3 Memory Module. To support long-term, coherent discussions, *Echo* includes a Memory Module that stores key user preferences, stances expressed via Question Card interactions, and adopted inspiration points. We observed in the formative study that users might sometimes express or agree with contradictory viewpoints. A dedicated Logic Agent monitors information saved to the Memory Module. When a new user interaction result (e.g., selecting "Yes" on a judgment card) is recorded, the Logic Agent automatically checks it for logical consistency against relevant stored information. If a clear contradiction is detected (e.g., previously agreeing "AI painting is essentially a tool" and later agreeing "AI painting has independent creative intent"), the system actively notifies the user of the potential conflict via a specific cue (e.g., a highlighted notification or a dedicated card).

5 User Study

To evaluate *Echo*'s effectiveness and user experience in promoting human-AI deliberation, we conducted a within-subjects study with 16 participants. Our research questions (RQs) were:

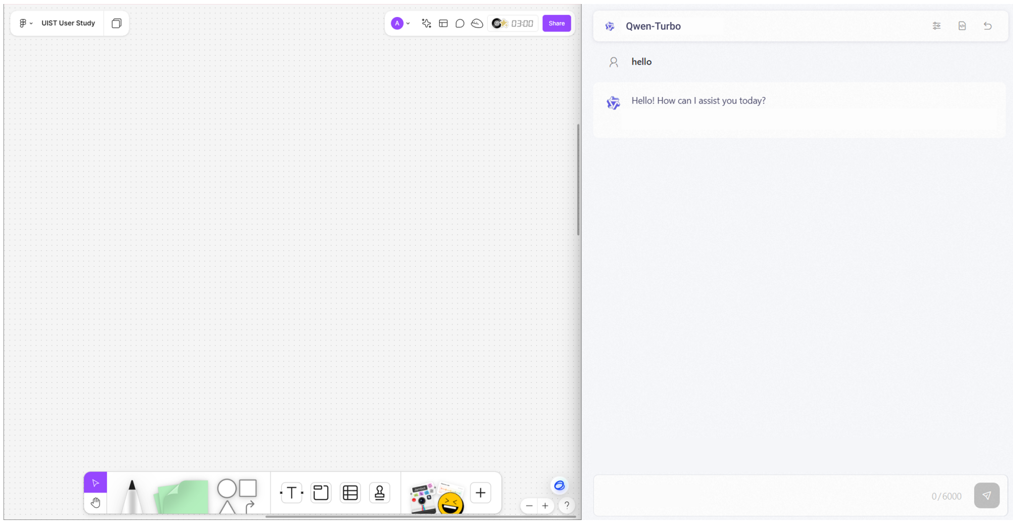
- **RQ1:** How would *Echo* affect the outcome of the human-AI deliberate discussion?
- **RQ2:** How would *Echo* affect user experience during the human-AI deliberate discussion process?
- **RQ3:** How would users interact with and perceive *Echo*?

5.1 Baseline Interface

The baseline interface had the same free canvas area and AI chat area as *Echo*, but lacked *Echo*'s Question Cards and automatic thought tree generation (Figure 5). Both systems used the Qwen-turbo model. Qwen-turbo was chosen for its fast response speed to ensure a good user experience, allowing participants to focus on the human-AI deliberation interaction. The model also supports Chinese well, facilitating participant understanding (as the study was conducted in Chinese). To minimize confounding factors like UI style, we used Figma's FigJam Board for mind mapping in the baseline, similar to *Echo*'s React Flow canvas. Participants were instructed to only use basic node and link functions in FigJam, avoiding other features (like FigJam's AI generation). Although the baseline lacked a dedicated voice input button, we encouraged using the OS's voice input in the tutorial, ensuring similar 'voice + typing' input options as *Echo*.

5.2 Participants and Tasks

We recruited 16 participants (8 female, 8 male; age 22-26; M=23.8, SD=0.83) via social media and personal networks. Recruitment criteria included prior experience using AI for advanced creative tasks



hello

Hello! How can I assist you today?

Figure 5: Baseline interface: Figma's FigJam Board on the left, Qwen-Turbo chat on the right. Users discuss with the LLM and summarize ideas on the left board.

(writing, coding, ideation) and interest in human-AI deliberation. Participants came from 11 majors: Integrated Circuits, Industrial Design, HCI, Information Art Design, Finance, Education, Software Engineering, AI, Mathematics, Chinese Language & Literature, Software Engineering. Questionnaires showed 10 participants used AI multiple times daily (>3 times/day) and 6 used it 1-2 times daily. All had used AI for complex (e.g., data analysis, coding) and creative tasks (e.g., writing, design). This background ensured familiarity with AI discussion capabilities.

We designed two discussion topics, similar to the subjective, open-ended topics from the formative study:

- **Topic A:** Is involution (intense internal competition) a societal problem or a result of personal choices?
- **Topic B:** Is speaking properly/politely an advancement in communication or a suppression of expression?

We used a Latin Square design to counterbalance the order of systems and topics. This resulted in four conditions, each with 4 participants randomly assigned: Topic A (Baseline then *Echo*), Topic A (*Echo* then Baseline), Topic B (Baseline then *Echo*), Topic B (*Echo* then Baseline).

5.3 Measures

RQ1: Discussion Outcome. To measure outcomes, we focused on the number of distinct points participants could articulate afterward. After each task, participants verbally recalled key arguments, supporting reasons, counterarguments, or insights from the discussion. Recalls were audio-recorded and transcribed. Two independent researchers coded the transcripts, counting the number of Meaningful Point Units. A Meaningful Point Unit was defined as a distinct, relevant statement of an idea (e.g., core argument, specific support, counterpoint, novel perspective/solution). Simple agreements ("yes"), repetitions, or procedural statements ("let's discuss...") were excluded.

RQ2: User Experience during the Process. To assess overall user experience, we adapted standard scales from the Technology Acceptance Model (TAM), measuring Ease of Use, Usefulness, and Intention to Use. Each construct had 2-3 items, and internal consistency (Cronbach's α) was calculated. We also used a 7-point Likert scale questionnaire with 6 items to probe other subjective experiences: Focus, Insight, Fluency, Calmness, Flow, and Perceived Reward.

RQ3: Interaction with and Perception of *Echo*. To understand how users interacted with and perceived *Echo*, especially regarding its design goals, we collected multiple data types. We designed 7 additional Likert scale questions directly corresponding to the achievement of the four design goals (**DG1-DG4**) (e.g., perceived alignment, sense of control, inspiration, helpfulness of non-linear interaction). Additionally, we logged key interactions (e.g., card clicks) and collected in-depth qualitative feedback through post-study semi-structured interviews to fully understand the details, experiences, and reasons behind user interactions with *Echo*.

All Likert items for RQ1-3 used a standard 7-point scale (1 = Strongly Disagree, 7 = Strongly Agree).

5.4 Procedure

The study employed a within-subjects design where each participant experienced both *Echo* and the Baseline system, completing one predefined open-ended discussion task per system (Figure 6). Topic assignment and system order were counterbalanced using a Latin Square design. For each system condition, participants first watched a 2-minute tutorial video, followed by 3 minutes of free exploration for familiarization. They then engaged in a 35-40 minute deep discussion with the AI on the assigned topic using the system. Immediately after the discussion, participants verbally recalled key points formed during the session without looking at the screen (for **RQ1** data). Following the recall, they completed an online questionnaire covering user experience (**RQ2** scales) and system perception (**RQ2/RQ3** items). After finishing all steps for the first system, participants took a 5-minute break before repeating the entire procedure (tutorial, exploration, discussion, recall, questionnaire) with the second system and the other topic. Upon completing tasks for both systems, a 10-20 minute semi-structured interview was conducted with each participant to gather qualitative data for **RQ3**, focusing on comparative experiences, opinions on *Echo*'s features, and suggestions. The entire session, including screen activity, verbal recall, and interviews, was recorded via Tencent Meeting for later analysis. The total duration per participant was approximately 1.5 to 2 hours.

5.5 Data Analyses

We used mixed methods (quantitative and qualitative) for data analysis. For RQ1 (discussion outcome - number of meaningful points), we checked normality using Shapiro-Wilk tests. Since the baseline data was non-normal (W(16)=0.856, p=.017) while *Echo* data was normal (W(16)=0.938, p=.321), we used the non-parametric Wilcoxon Signed-Rank test (appropriate for paired, potentially non-normal data) to compare the number of points between the two conditions. For RQ2 (user experience), we first calculated internal consistency (Cronbach's α) for Ease of Use, Usefulness, and Intention to Use, then averaged the item scores for each construct. We used Wilcoxon Signed-Rank tests to compare *Echo* and Baseline on these averaged scores and the other 6 experience ratings (Focus, Insight, etc.). For RQ3 (interaction with and perception of *Echo*), we used Wilcoxon Signed-Rank tests to analyze the 7 Likert items related to *Echo*'s design goals (**DG1-DG4**). We also performed descriptive statistics on system log data and applied thematic analysis to the transcribed interview recordings. The significance level for all statistical tests was set at $p < 0.05$.

6 Result

In analyzing user experience (**RQ2**), we first examined the internal consistency of three constructs: Ease of Use (Q1-3), Usefulness (Q4-6), and Intention to Use (Q7-8). Calculating Cronbach's Alpha revealed good consistency for all constructs under both system conditions (Ease of Use: Baseline α=0.747, *Echo* α=0.843; Usefulness: Baseline α=0.868, *Echo* α=0.808; Intention to Use: Baseline α=0.913, *Echo* α=0.793; all > 0.70). Therefore, we averaged the scores for the items within each construct for subsequent Wilcoxon Signed-Rank test analysis.

The data results for **RQ1-RQ3** are summarized in Table 2.

6.1 Impact on Discussion Outcome (RQ1)

We focused on the number of distinct points participants could articulate post-discussion, measured by the count of Meaningful Point Units. A Mann-Whitney U test revealed a statistically significant difference between the two conditions (U = 196.5, p = 0.008). Specifically, participants using *Echo* generated significantly more meaningful points (M = 5.1, SD = 1.7) than when using the Baseline system (M = 3.5, SD = 0.79). This quantitative increase aligns with our qualitative observations, suggesting *Echo* fostered a more divergent and exploratory discussion process. Participants noted that interacting with *Echo* prompted them to consider aspects they hadn't previously. For instance, when discussing the complex topic of 'involution', several *Echo* users expanded the discussion beyond its usual interpretations to delve into the interplay between personal choice and broader society. This exploration of the relationship between individual agency and societal pressures represented a broadening of perspective that participants felt was less likely or absent in the Baseline condition. Thus, *Echo* appeared not only to increase the quantity of discussion points but also encouraged users to engage with more diverse and potentially deeper conceptual territories.

6.2 Impact on User Experience during the Process (RQ2)

Analysis showed *Echo* significantly enhanced user subjective experience across several key dimensions. Users found *Echo* significantly more Useful (M=6.13, SD=0.71) than Baseline (M=4.27, SD=1.28; Z=5.06, p<0.001). Correspondingly, users expressed a significantly stronger Intention to Use *Echo* (M=5.84, SD=0.75) compared to Baseline (M=3.94, SD=1.64; Z=4.23, p<0.001). This indicates users recognized *Echo*'s value in aiding deep discussion and were willing to use it in the future. Many users reported *Echo* provided high-quality and insightful responses, such as P7 (M, 23) noting *Echo*'s *"responses were high-quality and useful,"* and P11 (M, 25) praising its *"professional and high-level questioning angles...points were valuable."*

Delving deeper, *Echo* significantly enhanced users' Sense of Insight (M=6.12, SD=0.81) far beyond Baseline (M=3.62, SD=1.20; Z=6.90, p<0.001), validating its design goal of fostering mutual inspiration (**DG3**). Several participants described breakthrough moments while using *Echo*. For example, P5 (M, 26) rated *Echo*'s responses as *"high-quality and inspiring,"* and P15 (M, 22) also stated its *"inspiration was higher than expected."* This sense of insight partly stemmed from *Echo* helping users organize thoughts and discover new connections, as P3 (M, 25) stated: *"The visualization (Dynamic Thought Tree) helped us review previous discussions; improvement felt like constant progress."* Furthermore, users felt significantly higher Intrinsic Motivation / Perceived Reward when using *Echo* (M=5.62, SD=1.09) compared to Baseline (M=4.06, SD=1.06; Z=4.11, p<0.001). This suggests the interaction process with *Echo* was inherently interesting and valuable. Users could not only *"obtain valuable information"* (P11, M, 25) but also enjoyed the interaction itself, with P3 (M, 25) stating, *"The visualization feature is really cool."* P9 (M, 24) even mentioned, *"The system's line of thought is very similar to mine,"* indicating this feeling of being understood and aligned contributed to the positive experience. Notably, there was no significant difference in Ease of Use between *Echo* (M=5.40, SD=1.14) and Baseline (M=5.40, SD=0.90; p=1.0000). Considering *Echo* introduced new interaction components (Question Cards, Thought Tree), this suggests its design did not impose additional learning burden or operational difficulty.

6.3 Perceptions towards Echo (RQ3)

6.3.1 Understanding Alignment (DG1). Data showed users significantly felt better understood by the AI with *Echo* (M=5.62 vs Baseline M=4.00; p=0.0005). Although the difference in ease of expressing views was not significant (M=5.62 vs Baseline M=5.00; p=0.1884), combined with the high understanding score, this suggests *Echo*'s interaction mechanisms (like Question Cards and the Thought Tree)

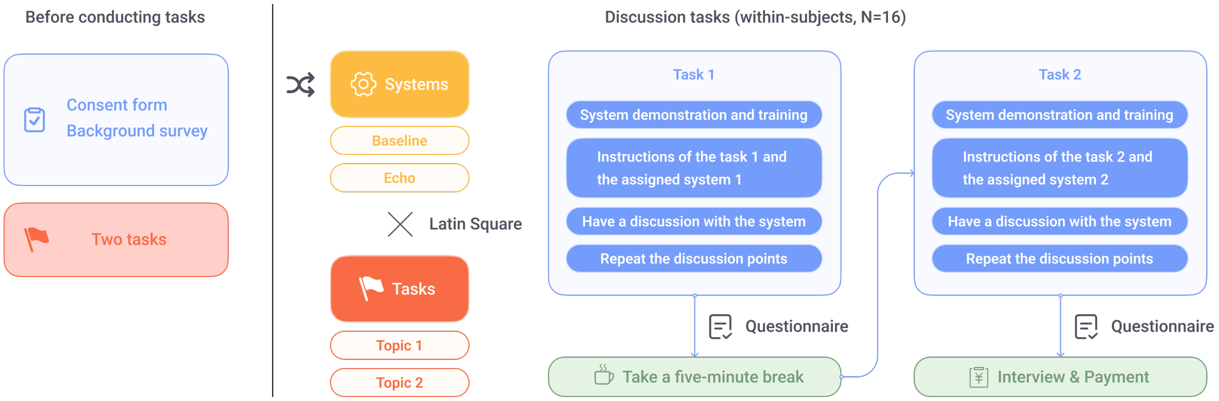

Before conducting tasks
Consent form
Background survey
Two tasks
Systems
Baseline
Echo
Latin Square
Tasks
Topic 1
Topic 2
Discussion tasks (within-subjects, N=16)
Task 1
System demonstration and training
Instructions of the task 1 and the assigned system 1
Have a discussion with the system
Repeat the discussion points
Questionnaire
Take a five-minute break
Task 2
System demonstration and training
Instructions of the task 2 and the assigned system 2
Have a discussion with the system
Repeat the discussion points
Questionnaire
Interview & Payment

Figure 6: User Study Process Diagram.

Table 2: Comparison of *Echo* and Baseline Interfaces on Discussion Outcome (RQ1), User Experience (RQ2), and Perception Metrics (RQ3). Values are Mean (SD). Statistics show Wilcoxon Signed-Rank Z-value (T) or Mann-Whitney U value (U). Significance levels: + p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

Research Questions	Item	*Echo*	Baseline	Statistics		
		Mean(SD)	Mean(SD)	*T/U*	*p*	*Sig.*
(RQ1) Discussion Outcome	Number of ideas	5.1 (1.7)	3.5 (0.79)	196.5 (U)	0.008	**
(RQ2) User Experience during the Process	Ease of Use	5.40 (1.14)	5.40 (0.90)	-0.00 (Z)	1.0000	-
	Usefulness	6.13 (0.71)	4.27 (1.28)	5.06 (Z)	<0.001	***
	Intention to Use	5.84 (0.75)	3.94 (1.64)	4.23 (Z)	<0.001	***
	Concentration	4.94 (1.39)	4.06 (1.34)	1.81 (Z)	0.0798	+
	Sense of Insight	6.12 (0.81)	3.62 (1.20)	6.90 (Z)	<0.001	***
	Fluency/Doability	5.56 (1.03)	4.56 (1.31)	2.39 (Z)	0.0231	*
	Calmness/Serenity	5.06 (1.24)	5.00 (1.51)	0.13 (Z)	0.8987	-
	Flow/Timelessness	5.31 (1.20)	4.12 (1.41)	2.57 (Z)	0.0153	*
	Intrinsic Motivation	5.62 (1.09)	4.06 (1.06)	4.11 (Z)	<0.001	***
(RQ3) Perception of *Echo*	Convenient to express opinions (DG1)	5.62 (1.09)	5.00 (1.51)	1.35 (Z)	0.1884	-
	Good understanding by AI (DG1)	5.62 (0.96)	4.00 (1.37)	3.90 (Z)	<0.001	***
	Control the discussion progress (DG2)	5.38 (1.31)	4.31 (1.82)	1.90 (Z)	0.0673	+
	Generate new ideas (DG3)	6.25 (0.86)	4.12 (1.50)	4.92 (Z)	<0.001	***
	System caters to users (DG3)	4.38 (1.15)	4.94 (1.57)	-1.16 (Z)	0.2562	-
	Mind map assistance (DG4)	5.88 (1.36)	3.88 (1.59)	3.83 (Z)	<0.001	***
	Thinking in Macro Dimensions (DG4)	5.75 (1.13)	4.88 (1.36)	1.98 (Z)	0.0566	+

effectively helped the AI capture and confirm user intent, making users feel their ideas were accurately received. P9 (M, 24)'s feedback corroborates this: *"Visualizing the thought process helps AI alignment better,"* even feeling *"The system's line of thought is very similar to mine,"* effectively addressing the challenge of insufficient AI understanding (**C1**) found in the formative study.

6.3.2 Discussion Control (DG2). Regarding the goal of providing fine-grained control and structured exploration (**DG2**), users felt a slightly enhanced sense of control over discussion progress with *Echo*, approaching statistical significance (M=5.38 vs Baseline M=4.31; p=0.0673). This indicates that *Echo*'s interaction methods (e.g., focusing discussion by clicking tree nodes or responding to cards) somewhat empowered users to guide the conversation, alleviating the issue of difficult-to-control AI output mentioned in **C2**. Some users explicitly felt this control, like P9 (M, 24) stating *"The user can control the agent,"* and P10 (F, 22) appreciating the ability to *"Break down ideas, the user can control the details."* While the improvement wasn't extremely large, *Echo* demonstrably moved towards giving users more agency in the discussion.

6.3.3 Mutual Inspiration (DG3). A core goal of *Echo* was to promote active mutual inspiration (**DG3**), not just passive responses.

Data strongly supports *Echo*'s success here: users generated new ideas about the problem significantly more with *Echo* than Baseline (M=6.25 vs Baseline M=4.12; p<0.001). This aligns with the high "Sense of Insight" score observed in **RQ2**. User feedback echoed this, with comments like P5 (M, 26) praising its *"high-quality and inspiring responses,"* P11 (M, 25) feeling the discussion *"diverged even more than my original thoughts,"* and P4 (F, 22) agreeing the system *"can make one's thinking diverge."* Interestingly, on the item "System catered to me," *Echo* scored non-significantly lower than Baseline (M=4.38 vs Baseline M=4.94; p=0.2562). This suggests *Echo* might not have simply agreed with users, but rather provided moderate challenges or new perspectives through its interactions (like AI questions), thereby better stimulating user thinking and effectively overcoming the lack of constructive questioning noted in **C3**.

6.3.4 Visualized Reflection (DG4). To address the difficulty of global overview and reflection in linear interfaces (**C4**), *Echo* introduced the Dynamic Thought Tree (**DG4**). User feedback strongly validated this design's value. "Mind map (thought tree) assistance" received significantly higher ratings under the *Echo* condition (M=5.88 vs Baseline M=3.88; p=0.0006), indicating users found the tree very effective for recording ideas and reducing cognitive load. As P3 (M, 25) said: *"The Dynamic Thought Tree helps us review previous discussions,"* and P10 (F, 22) mentioned the value of *"global view visualization."* Concurrently, users also felt *Echo* somewhat promoted thinking from a macro perspective (M=5.75 vs Baseline M=4.88; p=0.0566, approaching significance). This shows the Dynamic Thought Tree not only offloaded the burden of note-taking but also, through its non-linear structured presentation, helped users better grasp the overall discussion, connect different viewpoints, and engage in higher-level reflection.

7 Discussion

A key contribution of this study is revealing that effective Human-AI Deliberate Discussion requires synergy between the AI's underlying intelligence needs and the interaction support mechanisms. The design and evaluation of the *Echo* system validate this point, and user feedback provides concrete directions for future systems. We distill these insights into the following design guidelines:

7.1 Design Guidelines for Synergistic Human-AI Deliberation

Guideline 1: Anchor and structure discussion around questions. User feedback (*"The question is more important than the content"* - P3; *"Forgot what the initial question was"* - P1 (M, 24)) reveals the limitations of linear dialogue. Therefore, effective deliberation systems must anchor the discussion structure around the user's core inquiry (the question), not just the content flow. This requires AI intelligence capable of identifying, tracking, and understanding the evolution of this intent, going beyond simple text responses. Correspondingly, the interface must explicitly visualize the current discussion focus (question/intent) and provide navigation based on a question chain rather than an information flow, allowing users to easily jump between question nodes and related dialogues to maintain coherence.

Guideline 2: Enable adaptive AI behavior to match diverse discussion needs. User feedback (P3/P6/P14) indicates that deep discussion requires the AI to move beyond a single response mode. An ideal AI should possess the intelligence to understand different discussion needs (e.g., seeking information vs. seeking inspiration) and adapt its behavior accordingly (e.g., direct answer vs. Socratic questioning; neutral stance vs. taking a position). Interaction design must provide mechanisms for users to guide or select the AI's behavior mode and clearly communicate the AI's current role or strategy to effectively manage expectations and collaboration.

Guideline 3: Empower users with control and editing rights over the discussion space. Automatically generated visualizations often become messy and misaligned with users' mental models (P2/P3/P9). Therefore, systems must grant users direct control and editing capabilities over the shared discussion space. This requires not only intuitive and flexible editing tools in the interface (P6/P12), especially supporting dynamic branch selection and pruning (P11/P16), but also AI intelligence that can understand and adapt to user's structural edits, collaboratively maintaining a discussion space that truly serves the user's thinking.

Guideline 4: Ensure transparent, efficient, and cognitively friendly interaction loops. Interaction friction and high cognitive load severely impact the discussion experience. Users need clear feedback (P1: *"Tell me before changing the graph"*), efficient input (especially voice - P14 (F, 24), P16 (F, 25)), instant orientation (P3/P11), and appropriate information density (P6 (F, 24): *"Common model issue...adds unnecessary content"*; P9 (M, 24): *"Control word count"*; P14 (F, 24): *"The gap with the user is too large..."*). This requires AI intelligence that generates responses of appropriate length and complexity, and provides transparent previews or explanations for its actions (like modifying the graph - P1). Simultaneously, interaction design must ensure timely and clear feedback, smooth and interruptible input (especially voice - P3 (M, 25)), presentation methods that actively manage cognitive load (P6/P14), and tight, efficient coupling between dialogue and visualization.

7.2 Limitation

First, the limited interaction duration and specific discussion topics mean *Echo*'s visualization (Dynamic Thought Tree) might face scalability issues and information overload in longer or more complex discussions. Second, *Echo*'s interaction mechanisms need refinement. While the tree and cards showed potential, users desired stronger visual editing control (e.g., free restructuring, branch pruning) and more natural, adaptive interactions for alignment and inspiration (currently reliant on prompts and heuristics, sometimes feeling mechanical). Finally, and crucially, *Echo*'s performance is fundamentally limited by the underlying LLM's capabilities. While interaction design can effectively guide and compensate, it cannot fully overcome the model's inherent limitations in deep understanding, content appropriateness, or logical consistency.

7.3 Future Work

Future work will focus on further deepening the potential of collaborative human-AI discussion:

Quantifying and Guiding Divergence: We currently assess discussion inspiration qualitatively. In the future, we plan to explore computational metrics like knowledge entropy to quantify discussion divergence. This could enable more objective evaluation of interaction strategies and allow the AI to adapt its inspiration tactics based on real-time entropy, actively guiding users to broaden their thinking when stuck.

More Natural Interaction Modalities: To reduce interaction cost and enhance immersion, we will explore more natural input/output methods. This includes enhancing voice interaction for both content input and structural operations/queries on the thought tree, and investigating gaze tracking as an implicit intent signal (e.g., automatically focusing branches or triggering information based on gaze) for seamless collaboration.

User-Controllable AI Stance: To better stimulate critical thinking and perspective exploration, we will allow users to explicitly control the AI's stance during discussion. Beyond simple style adjustments, users could assign specific roles (e.g., critic or supporter) or even extreme scenarios (e.g., having the AI oppose the user's points 100%) to help users rigorously examine their arguments.

8 Conclusion

Current large language models struggle to support deep discussions requiring alignment and mutual inspiration. Through a formative study, we identified key challenges and user needs, leading to the design of the *Echo* system. *Echo* uses Question Cards for real-time alignment and a Dynamic Thought Tree for structured visualization and reflection. A user study confirmed *Echo* significantly enhances discussion depth, novelty, and user experience. Crucially, our work highlights a core principle: effective human-AI deliberation relies not just on better models or novel interfaces alone, but on the indispensable synergy between them. This principle, along with our design guidelines, paves the way for future systems promoting truly mutually illuminative human-AI collaboration.

References

[1] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[2] Ángel Alexander Cabrera, Adam Perer, and Jason I Hong. 2023. Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–21.

[3] Pei-Yu Chen. 2022. AI alignment dialogues: An interactive approach to AI alignment in support agents. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 894–894.

[4] Xiang "Anthony" Chen, Chien-Sheng Wu, Lidiya Murakhovs' ka, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2023. Marvista: exploring the design of a human-AI collaborative news reading tool. *ACM Transactions on Computer-Human Interaction* 30, 6 (2023), 1–27.

[5] Anthony Diamond. 2025. PRISM: Perspective Reasoning for Integrated Synthesis and Mediation as a Multi-Perspective Framework for AI Alignment. *arXiv preprint arXiv:2503.04740* (2025).

[6] Nicholas Duran, Amie Paige, and Sidney K. D'Mello. 2024. Multi-Level Linguistic Alignment in a Dynamic Collaborative Problem-Solving Task. *Cognitive science* 48 1 (2024), e13398. doi:10.1111/cogs.13398

[7] Simret Araya Gebreegziabher, Elena L Glassman, and Toby Jia-Jun Li. 2024. MOCHA: Model Optimization through Collaborative Human-AI Alignment. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–4.

[8] G. Gigerenzer. 2023. Psychological AI: Designing Algorithms Informed by Human Psychology. *Perspectives on Psychological Science* 19 (2023), 839 – 848. doi:10.1177/17456916231180597

[9] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[10] Dongyu Gong. 2023. Assessing Working Memory Capacity of ChatGPT. *ArXiv* abs/2305.03731 (2023). doi:10.48550/arXiv.2305.03731

[11] Chahna Gonsalves. 2024. Generative AI's Impact on Critical Thinking: Revisiting Bloom's Taxonomy. *Journal of Marketing Education* (2024). doi:10.1177/02734753241305980

[12] Nitesh Goyal, Minsuk Chang, and Michael Terry. 2024. Designing for Human-Agent Alignment: Understanding what humans want from their agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[13] Alicia Guo, Pat Pataranutaporn, and Pattie Maes. 2024. Exploring the impact of AI value alignment in collaborative ideation: Effects on perception, ownership, and output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.

[14] Jennifer L Heyman, Steven R Rick, Gianni Giacomelli, Haoran Wen, Robert Laubacher, Nancy Taubenslag, Max Knicker, Younes Jeddi, Pranav Ragupathy, Jared Curhan, et al. 2024. Supermind Ideator: How scaffolding Human-AI collaboration can increase creativity. In *Proceedings of the ACM Collective Intelligence Conference*. 18–28.

[15] Matt-Heun Hong, Lauren A Marsh, Jessica L Feuston, Janet Ruppert, Jed R Brubaker, and Danielle Albers Szafir. 2022. Scholastic: Graphical human-AI collaboration for inductive and interpretive text analysis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–12.

[16] Patricia K. Kahr, G. Rooks, M. Willemsen, and Chris C. P. Snijders. 2024. Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions. *ACM Trans. Interact. Intell. Syst.* 14 (2024), 29:1–29:30. doi:10.1145/3686164

[17] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on AI - An experimental study on the extent and costs of overreliance on AI. *Comput. Hum. Behav.* 160 (2024), 108352. doi:10.1016/j.chb.2024.108352

[18] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[19] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective explanations: Leveraging human input to align explainable ai. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.

[20] Jingyang Li, Shengli Song, Yixin Li, Hanxiao Zhang, and Guangneng Hu. 2024. ChatMDG: A discourse parsing graph fusion based approach for multi-party dialogue generation. *Inf. Fusion* 110 (2024), 102469. doi:10.1016/j.inffus.2024.102469

[21] Jessy Lin, Nicholas Tomlin, Jacob Andreas, and J. Eisner. 2023. Decision-Oriented Dialogue for Human-AI Collaboration. *Transactions of the Association for Computational Linguistics* 12 (2023), 892–911. doi:10.1162/tacl_a_00679

[22] Inês Lobo, Janin Koch, Jennifer Renoux, Inês Batina, and Rui Prada. 2024. When Should I Lead or Follow: Understanding Initiative Levels in Human-AI Collaborative Gameplay. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 2037–2056.

[23] Inês Lobo, Janin Koch, Jennifer Renoux, Inês Batina, and Rui Prada. 2024. When Should I Lead or Follow: Understanding Initiative Levels in Human-AI Collaborative Gameplay. *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (2024). doi:10.1145/3643834.3661583

[24] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2024. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. *arXiv preprint arXiv:2403.16812* (2024).

[25] Erik Miehling, Manish Nagireddy, Prasanna Sattigeri, Elizabeth M Daly, David Piorkowski, and John T Richards. 2024. Language models in dialogue: Conversational maxims for human-ai interactions. *arXiv preprint arXiv:2403.15115* (2024).

[26] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[27] Iwan Purnama, Firman Edi, Syahrul, Risnawati Agustin, and Nuridin Widya Pranoto. 2023. GPT Chat Integration in Project Based Learning in Learning: A Systematic Literature Review. *Jurnal Penelitian Pendidikan IPA* (2023). doi:10.29303/jppipa.v9ispecialissue.6712

[28] Jeba Rezwana and Mary Lou Maher. 2023. Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–28.

[29] Michal Rinott and Orit Shaer. 2024. Temporal Aspects of Human-AI Collaborations for Work. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–7.

[30] M. A. Rusandi, Ahman, Ipah Saripah, Deasy Yunika Khairun, and Mutmainnah. 2023. No worries with ChatGPT: building bridges between artificial intelligence and education with critical thinking soft skills. *Journal of public health* (2023). doi:10.1093/pubmed/fdad049

[31] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[32] 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).

[33] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264* (2024).

[34] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. 2024. Valuecompass: A framework of fundamental values for human-ai alignment. *arXiv preprint arXiv:2409.09586* (2024).

[35] Kacper Sokol and Peter Flach. 2024. Interpretable representations in explainable AI: from theory to practice. *Data Mining and Knowledge Discovery* 38, 5 (2024), 3102–3140.

[36] Zhendan Sun and Ruibin Zhao. 2024. LLM Security Alignment Framework Design Based on Personal Preference. In *Proceeding of the 2024 International Conference on Artificial Intelligence and Future Education*. 6–11.

[37] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate chatbots to facilitate critical thinking on youtube: Social identity and conversational style make a difference. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

[38] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. Interactive AI alignment: specification, process, and evaluation alignment. *arXiv preprint arXiv:2311.00710* (2023).

[39] S. Valtolina and Lorenzo Neri. 2021. Visual design of dialogue flows for conversational interfaces. *Behaviour Information Technology* 40 (2021), 1008 – 1023. doi:10.1080/0144929X.2021.1918249

[40] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

[41] 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.

[42] Eirini Zormpa, Antje S. Meyer, and Laurel E Brehm. 2023. In conversation, answers are remembered better than the questions themselves. *Journal of experimental psychology. Learning, memory, and cognition* (2023). doi:10.1037/xlm0001292

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009