SenseSync: Supporting Collaborative Information-Seeking with the Involvement of Large Language Models

Mohammad Hasan Payandeh
mpayandeh@uwaterloo.ca
School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

Jian Zhao
jianzhao@uwaterloo.ca
School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

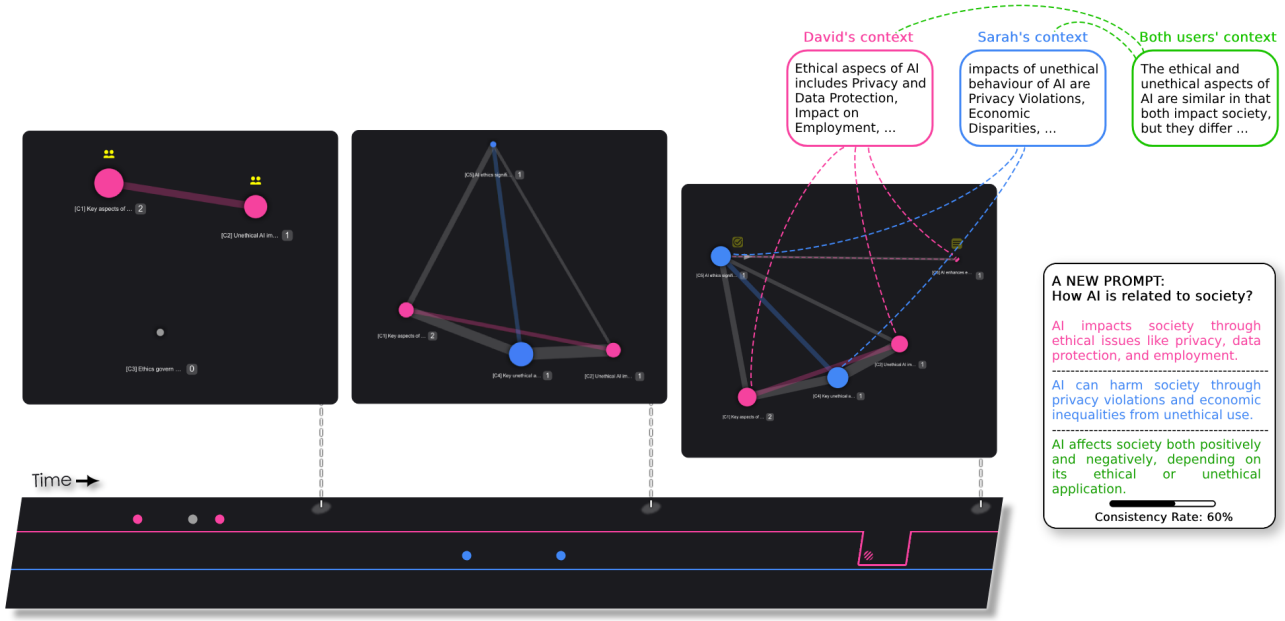


Figure 1: SenseSync is an interactive tool designed to facilitate collaboration work with LLMs by offering multiple perspectives. It includes a dynamic graph that visualizes both individual and shared conversations with LLMs, as well as a timeline that helps users explore collaborative information spaces over time. The tool is also enriched with contextual data and tailored features for LLM-assisted information-seeking. To help users verify the accuracy of LLM-generated information, SenseSync provides a consistency rate, which measures the similarity of responses from different LLM contexts linked to each collaborator when responding to a specific prompt.

ABSTRACT

Recently, tools driven by Large Language Models (LLMs), such as ChatGPT, have been extensively used for gathering information. While LLMs improve efficiency in individual tasks, new challenges emerge in collaborative information-seeking when user groups collect data from their conversations with AI that have various contexts. To fill this knowledge gap, we investigate these challenges and reflect on them via the design, development, and evaluation of SenseSync. SenseSync supports collaborative work involving LLMs from different perspectives, featuring a dynamic graph to display individual and shared conversations with LLMs and a visual timeline for exploring collaborative activities over different periods. Moreover, SenseSync is enriched with contextual information and specific support for LLM-assisted information-seeking. A summative study was conducted to explore how pairs of participants used the tool, enriching our understanding of LLM-assisted collaborative information-seeking tasks.

GI'25, May 26–29, 2025, Kelowna, British Columbia, Canada

https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Human-centered computing** → **Visualization systems and tools**; **Interactive systems and tools**; • **Information systems** → **Collaborative search**.

KEYWORDS

Information Seeking, Collaboration, Sense-making, Large Language Models (LLMs), Graph Visualization, Temporal Visualization.

ACM Reference Format:
Mohammad Hasan Payandeh and Jian Zhao. 2025. SenseSync: Supporting Collaborative Information-Seeking with the Involvement of Large Language Models. In *Proceedings of Graphics Interface*. ACM, New York, NY, USA, 19 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Information-seeking is one of the most critical activities that people engage in daily work and life, and collaboration is often needed when tasks are complex [46, 61]. Due to time and location constraints, remote asynchronous collaborative information-seeking remains more common in practical settings [45, 62]. People rely on various tools to gather information in this process, and recently, Large Language Model (LLM) based tools, such as ChatGPT, have been extensively used [23], both in individual and collaborative scenarios. While LLMs improve the efficiency of information-seeking in individual tasks, new issues emerge in collaborative scenarios because of the diverse context of each conversation with AI for collecting information. Specifically, each collaborator may converse with LLMs multiple times for different pieces of information to gather, and these different conversational contexts may make it more difficult to make sense of each other's information as well as their interpretations. For example, cross-functional teams conducting innovation sprints utilize LLMs to retrieve information on emerging trends, user needs, and creative design approaches from unique perspectives. Similarly, students working together on a course project might utilize ChatGPT to explore a topic by approaching it at different times and employing varied prompting strategies or setups. In both cases, the users should eventually share, consolidate, and reconcile the gathered information.

There exists a large body of literature on understanding the challenges of collaborative information-seeking and creating tools to address them [14, 25, 37, 38, 42, 46, 55, 61, 62], but only in the cases without LLMs. Thus, there remains a gap in understanding whether collaborators utilizing LLMs encounter similar challenges or new ones, as well as how an effective tool can be designed to accommodate LLMs in collaborative information-seeking.

To address the gap, this paper aims to provide insights into information-seeking with LLM involvement for remote asynchronous collaboration and explore new techniques to support it. We first conducted a formative study with eight (four pairs) participants performing collaborative information-seeking tasks with ChatGPT (e.g., collaboratively exploring ethical AI topics to write a report) to understand the challenges. The results show that participants faced various issues in making sense of LLM-curated information, including difficulties in identifying overlaps and gaps due to variations in writing style and the overwhelming amount of text generated by LLMs. They also expressed concerns in building a shared understanding because of lacking LLM-specific contextual information and support, as well as in switching, recalling, and resuming activities with LLMs. Further, participants pointed out the trustworthy issues of LLM-generated responses, as inconsistencies or hallucinations in outputs created mistrust and slowed down the collaboration.

Drawing from these findings, we designed SenseSync, an interactive tool that facilitates asynchronous collaborative information-seeking with LLMs from both spatial and temporal perspectives. Compared to existing tools for similar tasks, SenseSync was carefully enriched with crucial LLM-specific information related to users' conversations with LLMs and their collaboration. First, it incorporates a dynamic graph visualization for showing individual and shared LLM-based conversations in a collaborative view with the necessary context. Second, it equips a timeline visualization that enables users to browse and manage activities with LLMs in different time spans. Working together, they allow users to make sense of each other's information and understand their interpretation as well as to identify overlaps and gaps in collaborative information-seeking. SenseSync also includes features such as note-taking, task assignment, auto-summarization of shared conversations, and suggestions for exploring particular LLM's responses. To address the trustworthiness of LLM-generated information, SenseSync provides users with a consistency rate that shows the similarity of responses from various LLM contexts associated with each collaborator to similar prompts.

At last, a mixed-method summative study was conducted with 14 participants (seven pairs, with three new and four returning ones from the formative study), to assess the effectiveness of SenseSync. The results indicate that SenseSync effectively supported participants in addressing all identified challenges. We also identified other interesting findings such as participants' perception of SenseSync's LLM consistency rate and insights into the future development of tools to support remote synchronous collaborative information-seeking. In summary, our paper has the following contributions:

- A formative study that explores the challenges of collaborative information-seeking when utilizing LLMs.
- An interactive tool, SenseSync, that represents conversations with LLMs as a dynamic graph, enhancing collaborative sensemaking, exploration, task management, and validation of AI responses.
- Empirical knowledge learned from our summative study that highlights the potential of SenseSync in supporting LLM-assisted collaborative information-seeking and implications for designing similar tools.

2 RELATED WORK

This section reviews the literature on collaborative information-seeking and sensemaking as well as systems and techniques supporting these tasks, including those with LLM assistance.

2.1 Collaborative Information-seeking and Sensemaking

Information-seeking involves the actions and strategies used to find information that meets one's needs [57]. This can range from a simple lookup search [27], such as finding the capital of a country, to complex search tasks [2, 56] that occur when searchers are uncertain about what and how to fulfill their information needs. White and Roth's exploratory search model [56] describes two strategies

for complex tasks: exploratory browsing and focused searching. In exploratory browsing, searchers issue broad queries to expand their knowledge and identify relevant information. Researchers advocated that relevant information should be automatically revealed [18], particularly when dealing with the large volumes of information generated by LLMs [22]. As understanding improves and uncertainty decreases, searchers shift to focused searching to retrieve specific information, requiring systems to provide flexibility for their exploration [60]. In both strategies, effective sensemaking is essential to varying degrees [28]. It involves gathering information, representing it in a useful schema, developing insights, and creating knowledge or actions based on these insights [40].

Collaboration, whether synchronous or asynchronous, and co-located or remote [45, 62], is often required for information-seeking tasks that are complex for a single individual to handle [46]. While collaboration offers benefits, such as bringing diverse perspectives to a search task [48], it presents challenges. Collaborators need to share not only information but also their understanding of it [38, 42, 53]. Also, collaborators must be able to organize the information and identify overlaps and gaps in the shared information [42]. Additionally, being aware of others' activities is a significant challenge. Such awareness involves the understanding and knowledge that team members have about each other's activities, goals, and progress within a collaborative environment [4, 16, 46, 49]. Lastly, it is noteworthy that the nature of collaboration is temporal, and it is important to highlight the history of collaborative sensemaking [42] and the evolvement of information [38].

Informed by these theories and empirical knowledge, our goal is to identify the unique challenges users face during collaborative information-seeking when utilizing LLMs. This focused investigation aims to provide insights that will inform the design of future collaborative tools, emphasizing the specific context of LLM usage.

2.2 Information-seeking with LLM Assistance

Search engines have traditionally served as primary tools for information-seeking. However, with the advancement of Large Language Models (LLMs), generative AI tools such as GPT-4 [8], have now been extensively used for collecting information, questionnaire answering, and summarizing contents [23]. Numerous studies have been conducted to compare these information-seeking approaches (i.e., search engine-based and LLM-based) [9, 59, 64].

While there are some benefits of using LLMs such as efficiently generating information that is based on user-specific context, coherent, and human-like, it introduces challenges centered around prompting, evaluating and relying on outputs, and optimizing workflows [52], all of which impose substantial metacognitive demands on users. First, LLM responses highly depend on inputting prompts. It requires users to clearly define their goals and break down tasks into effective inputs for the AI, necessitating continuous self-monitoring and adjustment [1]. Second, making use of LLM generation involves assessing the quality and validity of the outputs, demanding users to critically evaluate their confidence in the trustworthiness of the results. This is because LLMs blend fact with fiction and generate non-factual content, which is known as hallucinations [5, 6, 12]. Third, leveraging LLMs requires users to strategically integrate AI into their processes, balancing automation with manual efforts [17] and continuously adjusting their approaches based on the AI's performance.

These challenges highlight the need of individuals for enhanced metacognitive support and system designs that facilitate better user control and understanding, ultimately aiming to improve the interaction between humans and AI by addressing the cognitive demands imposed by these advanced technologies [52]. Different approaches have been proposed to support users in assessing the information generated by LLMs, such as searching relevant databases and the web for matching sources [12, 43]. However, existing research primarily focuses on individual information-seeking tasks involving LLMs. Our study shifts the attention to collaborative settings that have been left under-explored. We investigate both persistent and new challenges that arise in collaborative information-seeking, considering the unique context each collaborator has with their LLMs. Our goal is to provide fresh insights into the effective integration of LLMs in these environments.

2.3 Systems and Techniques for Supporting Information-seeking

There exists an extensive body of research proposing systems and techniques to support information-seeking. We classify these works based on two aspects: whether they are designed for individual or collaborative tasks, and whether they utilize LLMs as an information source.

Individual tasks without LLM involvement. Researchers have long focused on designing systems for individual tasks without LLM or AI agent involvement. Some systems support the sensemaking of information retrieved from the web by capturing, organizing, and visualizing information to help individuals understand their findings [15, 30, 31]. Others facilitate exploratory search activities in digital libraries, such as refining search queries, organizing documents using workspaces, or discovering new documents by providing interactive features [3, 19, 39]. However, traditional search engines and information retrieval systems may struggle to fully grasp the nuances of an individual's specific needs or provide tailored information based on the user's previous context.

Collaborative tasks without LLM involvement. Even without AI agents or LLMs, various tools are available to support collaborative tasks, each enabling different aspects of collaboration. For example, CoSense [37] and Coagmento [14] focus on improving collaborative information-seeking by developing systems that offer interactive features to enhance sensemaking and support various aspects of collaborative search activities. KTGraph [62] utilizes techniques to capture and encode tacit aspects of the investigative process and streamline handoffs in asynchronous collaborative analysis. CLIP [26] provides a shared space, visualized in a graph, where users can record, organize, and share externalizations to improve awareness and coordination during sensemaking. However, while these systems effectively facilitate collaboration when the users utilize traditional information sources such as Google Search, they may require new or adapted features to support information generated by LLMs. This is because AI-generated information requires specific consideration, as it is shaped by the unique context

each collaborator has with their LLMs, bringing unique challenges such as concerns about validity or personalization.

Individual tasks with LLM involvement. With the recent rise of generative AI, in many individual tasks, users have shifted from traditional search methods to LLM-powered tools for gathering information [64]. This trend strongly motivates HCI researchers to understand user behaviors and design systems that address their challenges. For instance, Memory Sandbox [21] and Memolet [60] deal with managing and reusing conversational memory, giving users affordances to control how LLMs recall past information. Graphologue [22] and Knownet [58] use graphical representations (node-link diagrams or knowledge graphs) to make LLM responses more accessible and organized, helping users better explore and comprehend information. Sensecape [51], Marco [10], and Gero et al.'s system [13] focus on enhancing how users handle complex tasks with LLMs to aid sensemaking and task management. While these systems effectively address the unique challenges of individual information-seeking associated with LLMs, they lack essential considerations needed to support collaborative work.

Collaborative tasks with LLM involvement. As an emerging topic, there currently exists few studies on supporting users in collaborative information-seeking with LLMs. To design an effective system for this purpose, it is necessary to understand whether collaborators utilizing LLMs encounter similar challenges or new ones. Our research highlights that organizing, identifying overlaps and gaps, and summarizing LLM-generated information is crucial for effective collaborative sensemaking and exploration. Additionally, building a shared understanding through contextual data and specific support is essential for facilitating activities such as switching, recalling, and resuming tasks. Also, concerns about trusting LLM-generated information can be addressed through collaboration by monitoring inconsistencies in LLM responses.

3 FORMATIVE STUDY

While the literature has investigated challenges in collaborative sensemaking without LLM assistance [14, 37, 62], it was unclear whether these challenges are similar or not and if there are new challenges when LLMs are present. We thus conducted a formative study aiming to understand the challenges that collaborators face when using LLMs for information-seeking tasks.

3.1 Participants

We recruited eight participants (five men and three women; 20-28 years old), divided into four pairs using the university's mailing list. The participants were graduate students (four PhDs and four Master's) with experience using LLMs and prior involvement in collaborative information-seeking tasks. We used a pre-screening questionnaire to ensure that they met the inclusion criteria. To facilitate a comfortable working environment, we requested that each potential participant choose and bring a partner who also met the inclusion criteria. Of all the participants, five specialized in Software Engineering, two in Human-Computer Interaction, and one in Artificial Intelligence. Regarding their use of LLMs, four reported using them "multiple times a day," two "several times a week," and two "once a week." Participants' experience with collaborative tasks ranged widely from "a lot" to "not much," with the majority having "some" or "quite a bit." Table 1 and Table 2 in Appendix A outline the tasks that participants normally perform with LLMs and in collaboration. The study was approved by the institutional research ethics office.

3.2 Procedure

We conducted the study via video conferencing software with each pair of participants. After signing the consent form, participants were presented with five random topics (see Appendix A) from diverse domains (e.g., from ethics of AI to advanced materials for space exploration) and were asked to choose those that were unfamiliar to them. This was crucial because we wanted the designed information-seeking tasks to require complex search strategies, particularly exploratory browsing, where users are uncertain about what to search for and how to approach the information-seeking process [56]. Then, participants engaged in an open-ended information-seeking task (see Appendix A), which involved freely exploring the topic and was divided into two exploration phases. The rationale for asking participants to perform the task was twofold: 1) to familiarize the participants with using LLMs in collaborative information-seeking tasks, ensuring that their answers are specific to LLMs and based on real experiences rather than imagination; and 2) to observe their behavior and interactions with the system, and take notes accordingly.

In the first phase of the task, participants performed individual exploration and were not allowed to communicate via the conferencing software. This ensured that they brought diverse perspectives to the topic without imposing their search behaviors on each other [48]. In the second phase, they were instructed to share and discuss their individual findings and then continue their exploration collaboratively. However, we did not mandate how they collaborate, seeking to understand participants' organic needs. We did not require them to produce a report of their findings, as the search was exploratory with no specific best findings, focusing instead on identifying the challenges they faced while using LLMs during the search process. They were instructed to use ChatGPT to retrieve information and an online tool such as Google Docs to save and share individual findings. Next, a semi-structured interview was conducted with each pair of participants concurrently to collect their qualitative feedback. Participants were furnished with guiding questions (see Appendix A) designed to elicit information regarding their challenges and needs. Meanwhile, they were given the flexibility to discuss the specifics of their activities and the particular challenges and needs encountered while performing the information-seeking tasks. Also, we observed users' behavior during both individual and collaborative sessions, incorporating notes into our analysis later. Each participant received $20 for their time and effort.

3.3 Challenges

We transcribed all interview sessions and conducted thematic analysis in three stages: familiarization, coding, and theme development. In familiarization, we imported the audio transcripts and text observations into NVIVO and familiarized ourselves with the data.

During coding, we identified 365 initial codes, which were then organized into 67 broader categories to facilitate the identification of core themes. In theme development, we synthesized these categories to pinpoint four core themes related to the challenges faced by participants. While some results replicate existing challenges in collaborative tasks without LLMs, our analysis contributes to the emerging insights into collaborative environments integrated tightly with LLMs, which magnifies the challenges and brings new ones. Based on the challenges, we proposed design guidelines (Section 3.4), which drove the development of SenseSync.

C1: Making sense of LLM-curated information (118 codes). Individual exploration before collaborative work was deemed essential, as highlighted by P7: *"This was a good pattern to follow, to do like things first on our own and then like get together"*, which is also supported by the literature [42, 47, 53]. Doing this allows collaborators to contribute diverse perspectives to an information-seeking goal. However, before effective exploration can occur, it is crucial to make sense of the generated information. Participants expressed several key challenges regarding information sensemaking such as identifying overlaps and gaps in shared findings, summarizing and organizing information, and understanding others' outcomes.

These challenges were further complicated by the involvement of LLMs. Specifically, participants found it more difficult to identify overlaps and gaps when the information was generated by LLMs. Although the core content of the responses might be similar, variations in writing style made it challenging to discern whether there were overlaps by merely comparing the texts. As P8 noted: *"The answers are based on the style of questioning and are inconsistent. When we review and process these answers, it's not clear that they are the same, so they're not easily comparable."* Regarding summarizing and organizing information, participants observed that while LLMs could generate large amounts of text in seconds, reading, summarizing, and organizing this information is much more time-consuming. They expressed a preference for concise summaries over lengthy outputs. As P5 remarked: *"I didn't want to read so much, so I just asked ChatGPT for a TLDR."*

C2: Achieving a shared understanding of generated contents by LLMs (165 codes). Collaborative work involves activities collaborators perform to reach the shared goal, which is prominently discussed in the literature [26, 42]. Participants found that they needed to coordinate to align their efforts with the overall goal: *"We are going to merge and consolidate information in different ways, based on different goals."* -P2 To achieve their goal, they needed to have a shared understanding, aligning with observations from previous studies [38]. This could pose difficulties, as mentioned by P7: *"It is hard to make the other person understand why our point is right."*

When LLMs are used to generate information, the lack of context—such as the prompt used or the conversation history fed to the model—creates challenges in achieving a shared understanding: *"More information about the context would be helpful, like when we had that difference in understanding regarding whether it was ethical or unethical."* -P2 Also, participants noticed that collaborators might use different ways of prompting, potentially increasing the difficulty of reaching a shared understanding. P8 expressed the need to address this challenge: *"We mostly approached the same question from different perspectives. The first step is always to get everyone on the same page and exchange ideas and understandings."* -P8 They also tried to align themselves during the task while interacting with LLMs individually: *"When I had a different understanding of something, and he had a different one, we decided to help each other understand why our points were valid and why they should be considered."* -P5 All the above observations underscored the new issues in reaching a good shared understanding in collaboration, caused by the uncertainty of LLMs and the AI's blackbox-like behaviors.

C3: Switching, recalling, and resuming activities with LLMs (29 codes). We observed that the nature of collaboration is inherently temporal, meaning that participants engaged in distinct sub-activities throughout the task and needed switching between these different sub-activities over time. Literature also highlights that sensemaking is temporal in nature [38]. These sub-activities included validating specific LLM-generated responses, further exploring or clarifying aspects of the research, and discussing topics to enhance mutual understanding. Moreover, resuming from where they left off and continuing to achieve their goals can be challenging. Participants noted the need to recall previous progress made by collaborators: *"I make more notes while I'm working over a long period of time."* -P7 as well as to stay informed about what others have done during their absence: *"Because I didn't know what he had read, it was hard for me to write about his findings."* -P5

While switching, recalling, and resuming activities during collaboration may appear similar to traditional search scenarios, the inclusion of LLMs introduces unique considerations. For example, issuing the same prompt at different times could yield varying responses, as the context provided to the LLMs changes over time. Therefore, it became crucial to know the temporal LLM-specific contextual information, such as when other collaborators initiated their prompts. For instance, P4 remarked to P3: *"I'm getting a different response to the same prompt. What time did you ask?"* P4 further commented *"He went through all the prompts and responses and explained them to me one by one."*

C4: Investigating the trust with LLM-generated responses (51 codes). Some participants expressed that they in general did not trust the responses generated by LLMs: *"We should not trust the results. There's a line between common knowledge and domain-specific knowledge, and I wouldn't trust ChatGPT for anything that requires domain-specific expertise."* -P6 *"One thing that was constantly in my head while doing research with our GPT was the concern that it might hallucinate, so, I don't know if I'm getting the right information."* -P5 Additionally, there were inconsistencies in the responses to the same prompt from LLM agents in different conversational contexts. *"It was interesting how my ChatGPT was not giving me unethical stuff, but P1's was."* -P2

Hallucination [5, 6, 12] is an important issue of LLMs. It impacts both individual and collaborative experience, which was observed from our study, because reliable shared information is essential for successful teamwork. Also, we observed that mistrust in the shared information can slow down collaboration. When participants doubted the validity of LLM-generated information, they hesitated to accept or act on shared information by others, leading to more back-and-forth discussions and validation efforts. Additionally, inconsistencies in LLM-generated responses across different contexts

may further complicate collaboration. When one participant received different or even conflicting information from another, it led to confusion and reduced the effectiveness of collaboration.

3.4 Design Guidelines

Drawing from the above challenges discovered in our formative study, we derive the following design guidelines (**DG[n]**), each of which corresponds to the previously identified challenge (**C[n]**).

DG1: Enable the organization, comparison, and summarization of LLM-generated information. To support both individual and collaborative exploration, users need to make sense of their individual information, select relevant ones to share in a common space with other members, and collaboratively make sense of shared information [48]. This requires maintaining separate individual and collaborative spaces, with the ability to switch between them. To facilitate more effective sensemaking and exploration, it is important to illuminate the relationships among different information pieces. Specifically, identifying overlaps and gaps depends on revealing the similarities between these pieces of generated information. When working with LLMs, the above aspects should be properly designed to tailor the unconventional workflows with prompts and responses. It is also essential to facilitate prompting LLMs collaboratively: *"[P4] is way better at prompting GPT than I am. She could get what she wants in one prompt."* -P5

DG2: Provide LLM-specific context alongside generated information for enhanced shared understanding. When involving LLMs in collaborative work, it is important to improve a user's awareness of the responses to a prompt in another user's conversational context, going beyond just sharing LLM-generated content. For example, one should know what prompt was used to generate the responses shared in the collaboration: *"I think having the prompt will be helpful in the future when we want to reference where each piece of information came from."* -P8 Also, the sequence and timing of prompts can influence the generated responses, as the context provided to the LLM may differ. P7 commented, *"I started with a very simple question and then moved into more detailed ones. If someone sees my later questions before the first one, they might not understand the context. Why is that?"* P4 also spot the same issue, *"If you ask a question and receive a different answer than someone else, you might ask, 'What time did you ask your question?'"*

DG3: Support tracking collaborators' activities and interactions with LLMs over time. Collaboration involves many task switching, recalling, and resuming, and these activities become more complex when involving LLMs because a user works with collaborators and the AI at the same time. This awareness can be enhanced by providing interactive timeline visualizations of activities [24], which can be further integrated with interactions with LLMs, such as the number of prompts (P3) and the history of the prompts (P4), as well as progress cues, such as note-taking (P1) and task assignment (P4).

DG4: Offer collaborative approaches to assess trustworthiness of LLM-generated responses. To address LLM's hallucination and trust issues, existing approaches often involve comparing generated results with information from the web or other sources [12, 43]. In collaborative settings, it is important for comparing responses to a specific prompt across different conversational contexts and examine whether the responses are consistent. Low consistency could alert collaborators that the information might not be trustworthy and requires further verification: *"If he receives a response, I'd like to prompt ChatGPT with the same question to check if the response remains consistent or varies."* -P5 Beyond the consistency assessment, trustworthiness can be verified through discussion, combination, and cross-validation of LLM-generated content together. It is crucial to allow different collaborators to ask LLMs for clarification or further exploration, such as one continuing another user's conversation with AI. *"We can validate which elements are common and which might differ by talking to each other."* -P1

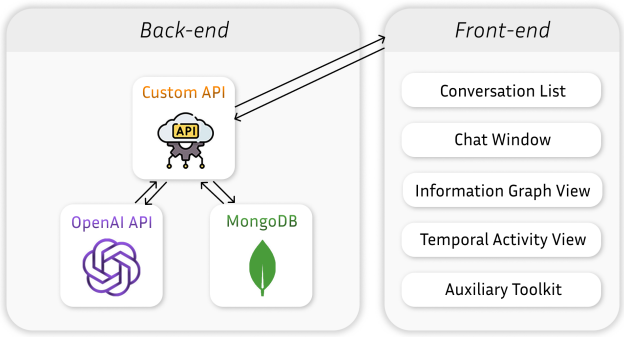


Figure 2: SenseSync consists of a back-end with a custom API, MongoDB Atlas, and OpenAI's API, and a front-end with five interface components.

4 SENSESYNC

This section introduces the design and implementation of SenseSync, followed by a scenario to illustrate how SenseSync can be used in practice.

4.1 System Overview

Guided by the aforementioned design goals, we developed the SenseSync system that consists of a back-end and a front-end (Figure 2). The back-end contains a custom API, MongoDB Atlas, and OpenAI's API. The custom API is implemented using Node.js [11] and Express.js [50] to handle tasks such as user management and the storage and retrieval of conversations. MongoDB Atlas [29] serves as the cloud-based database for storing data. Two models from OpenAI's API [8] are employed: the text-embedding-3-small for generating text embeddings and the GPT-4o mini for generating information. The front-end is developed using React [41], MaterialUI [44], and D3.js [34]. It comprises two primary workspaces: individual and collaborative, each offering different components to facilitate collaborative information-seeking (Figure 3). The collaborative workspace includes a Conversation List, Chat Window, Information Graph View, Temporal Activity View, and Auxiliary Toolkit, whereas the individual workspace features only a Conversation List, Chat Window, and Information Graph View.

The Conversation List (Figure 3-A) displays a list of ongoing and past conversations, allowing users to navigate between different

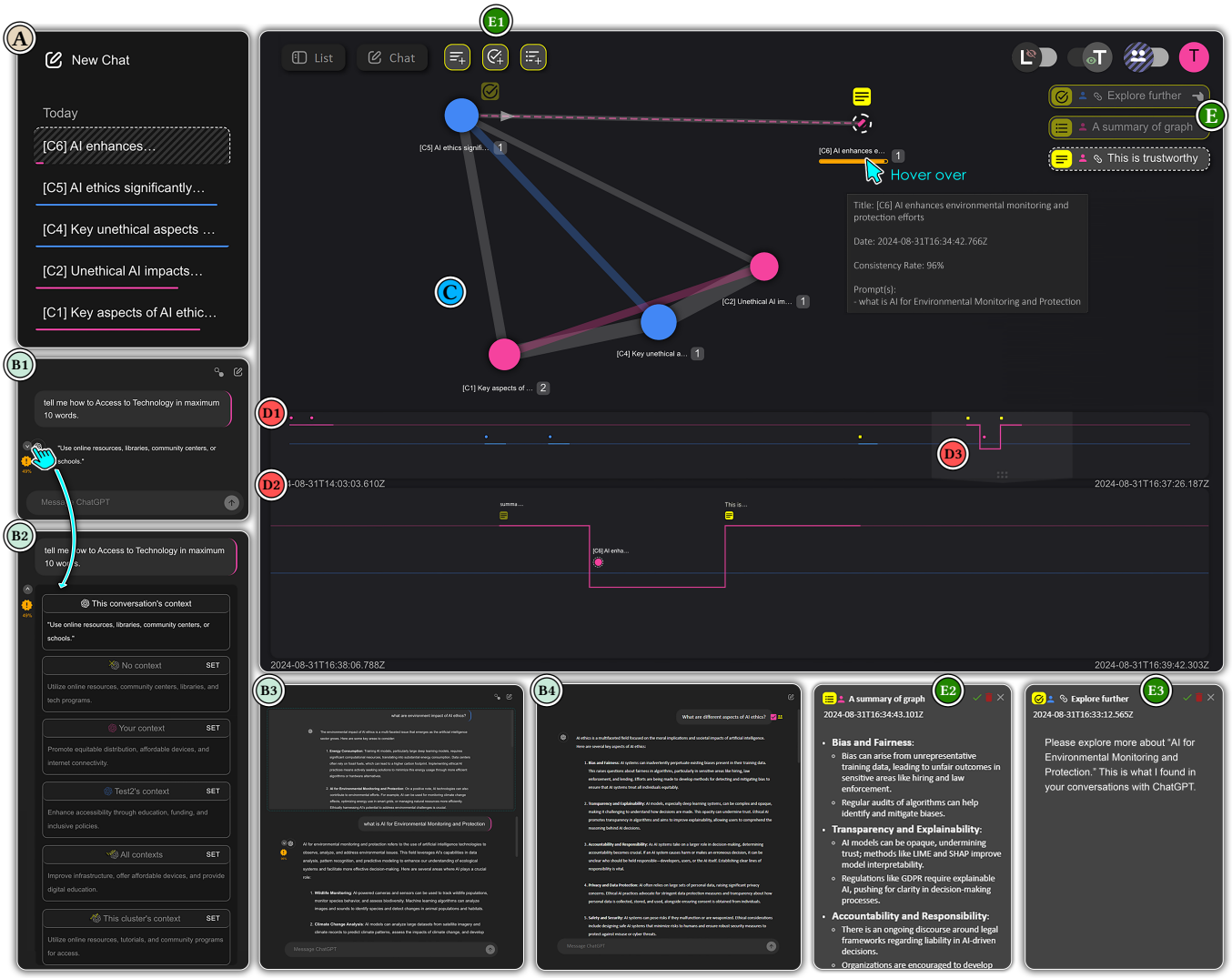


Figure 3: The front-end interface of SenseSync includes: (A) A Conversation List displaying conversation IDs and titles; (B1-4) Different Chat Windows that show conversation content, allows initiating new prompts, supports sharing/unsharing conversations (individual workspace (B4)), and integrates Consistency Rate and Context-Based Response Suggestions B1 features (collaborative space (B1, B3)); (C) An Information Graph View illustrating the similarity between different conversations; (D1-3) Minimap and zoomed sub-views of Temporal Activity View that show collaborator activities and allows adjustment of the timespan for the Information Graph View by moving the viewport (D3) horizontally; and (E) A Auxiliary Toolkit presenting a list of temporal Notes, Tasks, and Auto-summarizations generated by the Auto-Summarization feature (E3 corresponds to the content of a note, while E2 represents a summary. E1 represents buttons to add different activities.)

conversations. The Chat Window (Figure 3-B1 to B4) shows the content of conversations, including prompts and responses, and allows users to interact with ChatGPT. As shown in Figure 3-B2, the Chat Window also includes two main features (**DG4**): Context-based Response Suggestion, which offers varied responses based on collaborators' conversations with ChatGPT, and Consistency Rate, which helps evaluate the trustworthiness of responses by comparing similarities between suggested responses. The Information Graph View (Figure 3-C) includes a graph visualization where conversations are represented as nodes and their similarities as links. This graph highlights overlaps and gaps, provides a holistic view as a starting point for collaboration, and aids in conversations' organization (**DG1**). The Temporal Activity View leverages a temporal visualization, consisting of a minimap and a zoomed subview (Figure 3-D1 and D2), that showcase collaborators' activities and interactions with LLMs (**DG3**). Both the graph and timeline visualizations are enriched with contextual data on interactions between collaborators and LLMs (**DG2**). The Auxiliary Toolkit (Figure 3-E) offers two key features: Notes/Tasks and Auto-Summarization. The Notes/Tasks feature lets users annotate nodes with their thoughts or comments and assign tasks, such as seeking clarification or further exploration, to their partners, which aids in

recalling and resuming tasks. (**DG3**). The Auto-Summarization feature generates summaries of the entire graph or specific clusters of nodes using ChatGPT, helping users quickly review past activities (**DG1**). Together, these features enhance awareness of task progress and history, facilitating smoother transitions between tasks and collaboration with LLMs.

All the features consist of LLM-specific aspects absent in traditional collaborative information-seeking tools or baseline LLM interfaces (e.g., ChatGPT). The Conversation List employs LLM-derived visual encodings (e.g., colored lines for user contributions), unlike baseline LLM interfaces. The Chat Window uniquely addresses AI-generated content variability—a challenge unmitigated in baseline LLMs—via Context-based Response Suggestions (blending user/cluster-specific LLM contexts) and Consistency Rate (semantic similarity scoring to flag hallucinations). The Information Graph View automatically maps relationships of LLM-derived conversations, transcending manual tagging or metadata-based linking. The Temporal Activity View visualizes shifts in LLM conversational contexts over time, correlating AI interactions with collaborative progress. The Auxiliary Toolkit leverages LLMs for Auto-Summarization of multi-conversation outputs, unlike static user-authored notes.

4.2 SenseSync System

SenseSync is equipped with a collaborative group management feature. Upon logging in, users need to specify a task title by clicking on the  icon. To switch between individual and collaborative workspaces, users can click the  icon. Additionally, they can toggle the Conversation List () or Chat Window () to manage the screen space each view occupies. In the following, we introduce the main view components of SenseSync in detail.

4.2.1 Conversation List. The Conversation List (Figure 3-A) is designed to resemble OpenAI's ChatGPT interface [36]. However, it includes additional visuals, such as a static ID included as a prefix in conversations' titles to help collaborators locate and revisit conversations in other views as well as colored lines beneath each title, where pink represents the current user and blue indicates the user's partner. Also, the size of these lines indicates the relatedness of a conversation to all the shared information in comparison to other conversations. When users click or hover over a node in this view, the corresponding conversation is highlighted across all other views, making it simple to locate a specific conversation.

To measure relatedness, we first transform the information in each conversation into high-dimensional vectors using OpenAI embeddings. We then calculate the similarity between these vectors using cosine similarity [54]. The relatedness of a conversation is determined by the ratio of its average similarity to all other conversations, divided by the sum of the average similarities of all conversations to each other.

4.2.2 Chat Window. While the Chat Window (Figure 3-B1 to B4) is also designed to replicate OpenAI's ChatGPT interface, it incorporates specific visual encodings and features tailored to each workspace. In the individual workspace (Figure 3-B4), the  indicates whether a prompt or response is shared with others, with the option for users to uncheck a pink checkbox to avoid sharing. The Chat Window in the collaborative workspace (Figure 3-B1 to B3 and Figure 7), consists of two primary features: Context-based Response Suggestion and Consistency Rate.

Context-based Response Suggestion provides varied responses to a specific prompt by leveraging different contexts derived from combinations of nodes in the graph (Figure 3-B2). These contexts include: 1) the conversation's context, where the response is generated using only the information from the node currently being viewed; 2) no context, which means providing no additional context to the model, effectively treating the prompt as if answered without any contextual input; 3) different users' contexts, where the generated response is enriched by information from nodes associated with various users to offer a diverse perspective; 4) all contexts, where the model incorporates information from every node across the entire graph; and 5) this cluster's context, which consider nodes within a specific cluster to generate responses. Users can set the active response for a conversation to any of these different responses. This selection affects the shape of the graph, as the node similarity is calculated based on the chosen response.

Consistency Rate, displayed in the chat panel under an exclamation mark icon () and on the graph using an orange bar (), reflects the similarity between different responses generated by the Context-based Response Suggestion feature.

Users can interact with ChatGPT in three distinct ways: First, they can create a new conversation, with the system positioning it based on its similarity to other nodes. Second, they can start a new conversation linked to an existing one by clicking on  at the top of the Chat Window. Third, they have the option to continue chatting within the same conversation. In this collaborative workspace, users can view standard ChatGPT responses (B3), along with a Context-based Response Suggestion (B2) and a Consistency Rate.

4.2.3 Information Graph View. SenseSync employs a graph visualization (Figure 3-C) with various visual encodings to highlight overlaps and gaps, provides a holistic view as a starting point for collaboration, and aids in conversations' organization. First, solid patterns and diagonal hatch patterns  distinguish between individual and collaborative workspaces, respectively. Filled circle shapes  represent conversations, with the size of each circle indicating the conversation's relatedness to all of the shared information (the whole graph) in comparison to other nodes. If the circle's shape is outlined  , it means that the conversation has been removed. Solid  and dashed  links between two nodes illustrate their similarity or a semantic connection between them that has been made by the user, respectively. The links' thickness and length signify the degree of similarity. If two nodes are connected semantically to each other by the user, an arrowhead  demonstrates which conversation is connected to which. If there are multiple nodes semantically connected to each other, this arrowhead reveals the flow of exploration (see Figure 7-a). A dashed border  highlights conversations being hovered over by the mouse.

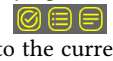
When hovering over some visual objects, the same visual encoding appears across the Conversation List and Temporal Activity

View as well. For example, hovering over a node in the graph illustrates where each encoding appears in other views (Figure 3). This uniformity aids users in locating objects across different views. Also, hovering over a node reveals contextual data, including the consistency rate, the number of prompt/response pairs in the conversation, the full titles of the prompts, and the date the conversation was issued (Figure 3-C).

To position nodes on the graph, we first generated embeddings from conversations using OpenAI embeddings and then applied MultiDimensional Scaling (MDS) [20] to perform dimensionality, projecting the 1536-dimensional vectors into 2-dimensional ones that determine the nodes' positions. We then employed D3-force[35] to fine-tune the nodes' positions, ensuring a clear and organized layout for collaborative work. Additionally, to indicate the relationship between nodes, we set a similarity threshold empirically; nodes with similarity scores exceeding the threshold are linked on the graph.

4.2.4 Temporal Activity View. SenseSync equips timeline visualization (Figure 3-D1,D2,D3) consists of: a minimap (Figure 3-D1) and a zoomed view (Figure 3-D2). The minimap shows the overall pattern of collaborators' activities, including the date and time when they issued a conversation, left a Note, assigned a Task, or generated an Auto-summarization, by considering a line to each user colored with the users' identity color. It shows if a conversation, Note, Task, or Auto-summarization is removed by outlining that element visually. Also, it reveals the point in time when a user was working on the search task by setting the line opacity to full intensity. Lastly, it indicates when users worked on each other's nodes by drawing the user's line close to and parallel with the other user's line (Figure 3-D1,D2). The system considers a user worked on others' nodes if a user semantically connected a node to another user's or they chatted with LLM in a node that is related to another user.

A viewport (Figure 3-D3) is included in the minimap, which users can drag horizontally to reveal details in the zoomed view, such as titles of conversations, notes, tasks, or summaries. Hovering over an element reveals contextual data, including the number of prompt/response pairs in that conversation, the full titles of the prompts, the date the conversation was initiated, and the element's representation in both the Graph and Conversation List.

4.2.5 Auxiliary Toolkit. This view includes two features: Notes/Tasks and Auto-summarization, all of which are unified into a single component at the implementation level and shown in one single list (Figure 3-E). A different icon is designed for each feature, with filled icons specifying items as not done () and outlined ones as done (). A hand icon indicates if a task is assigned to the current user. Users can view the details of these activities (Figure 3-E2 and E3) by clicking on them from the list. To add Notes, Tasks, or Auto-summarizations, users can click the yellow-outlined buttons (Figure 3-E1). Summaries are generated using the LLM on different levels, including the whole graph (if no node is selected) and a cluster (if nodes are selected). Users can mark Notes, Tasks, or Auto-summaries as done, remove them, or close the details section using  .

4.3 Usage Scenario

Suppose that David and Sarah, two university students, are collaborating on a report about the "Ethical Implications of AI" for their course project. Due to conflicting schedules, they decide to use SenseSync to support asynchronous collaborative information-seeking process of writing the report (Figure 1).

1) David's individual exploration. David begins by using the Chat Window to start a conversation with ChatGPT, asking, "What are the different aspects of AI ethics?" (Figure 3-B4). The system generates a response, [C1], which is represented as a node in the Information Graph View. He follows up with two additional prompts, resulting in three nodes: [C1], [C2], and [C3]. David notices that [C1] and [C2] are connected, indicating a strong relationship between the nodes. However, [C3], titled "Ethics govern moral behavior and decision-making," is not connected to the other two nodes and is represented by a small circle, suggesting it is less relevant. As a result, David decides to unshare [C3] so it will not appear in the collaborative workspace.

2) Sarah's individual exploration. Later, Sarah begins her exploration by asking, "What are the unethical aspects of AI?"—a broad prompt similar to David's, but focused on the negative aspects. After following up with an additional prompt, two new nodes, [C4] and [C5], appear in the Information Graph View.

3) Sarah's exploration in the collaborative workspace. Once Sarah completes her individual exploration, she decides to review David's findings and compare them with her own. Switching to the collaborative space, she notices that David's nodes, [C1] and [C2], are connected to her nodes, [C4] and [C5]. This overlap reveals that [C4] is highly similar to both [C1] and [C2]. However, [C5], while connected to all other three nodes, is farther from the others, indicating less similarity. Curious about [C5], Sarah clicks on it to read the corresponding ChatGPT response and discovers a new aspect of AI ethics: "AI for Environmental Monitoring and Protection." Uncertain about the validity of this information, she assigns a task to David, asking him to explore this aspect further.

4) David's exploration in the collaborative workspace. When David logs back into SenseSync, he opens the collaborative workspace to see the graph, now showing both his and Sarah's findings. Short on time, he uses the Auto-Summarization feature to get an overview of the shared information (Figure 3-E2). Looking at the Auxiliary Toolkit list, David notices the , indicating that Sarah has assigned him a task titled "further explore." Clicking on the task (Figure 3-E3), he reads that Sarah wants him to investigate the validity of the "AI for Environmental Monitoring and Protection" aspect, which is associated with node [C5]. After reviewing [C5] and understanding Sarah's uncertainty, David prompts with "What is AI for Environmental Monitoring and Protection?" to investigate further (Figure 3-B3). The system shows a high Consistency Rate, suggesting the response is trustworthy. To be sure, David reviews the Context-based Responses and concludes the information is reliable. He adds a note titled "This is trustworthy" to inform Sarah of his findings. Figure 1 shows how Information Graph View looks at this stage from David's perspective.

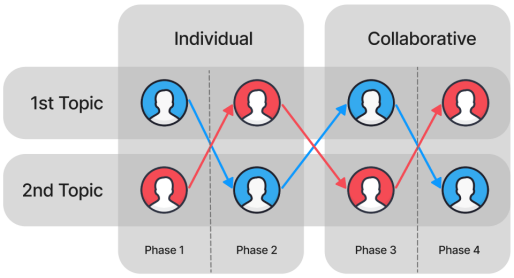


Figure 4: Study task design for simulating asynchronous collaboration, where different colors distinguish the participants working together.

5) Further exploration and collaboration. David and Sarah continue to collaborate through additional phases. To recall and resume their task, they utilize the Information Graph View and Temporal Activity View. As an example, Figures 3-C and 3-D illustrate how the workspace looks to David. SenseSync enabled David and Sarah to become familiar with various aspects of the ethical implications of AI, gain a deeper understanding of each, and choose key and reliable aspects to incorporate into their report.

5 SUMMATIVE STUDY

We conducted a summative study to investigate whether SenseSync can support the challenges identified in the formative study and learn about the strengths and weaknesses of the system.

5.1 Participants

Through mailing lists of a local university, we recruited fourteen participants (seven pairs; seven men and seven women; aged 20-39), of which eight were PhDs and six were Master's. Of all the participants, eight (four pairs) were involved in the formative study. We invited them back to see if they believed their challenges would be addressed by SenseSync or not. The six new participants (three pairs; two men and four women; aged 23-39; four PhDs and two Master's) were recruited using the same inclusion criteria and recruitment procedure as for the formative study. Of all the participants, five specialized in Software Engineering, four in Human-Computer Interaction, two in Artificial Intelligence, one in Data Engineering, one in Bioinformatics, and one in Transportation Engineering (not Computer Science). Regarding their use of LLMs, seven reported using them "multiple times a day," one "once a day," four "several times a week," and two "once a week." Participants' experience with collaborative tasks ranged widely from "a lot" to "not much," with the majority having "quite a bit" or "a fair amount." The study was approved by the institutional research ethics office.

5.2 Task and Procedure

The study task was designed to simulate asynchronous collaborations with each pair of participants. It was a collaborative information-seeking task using SenseSync for two different topics (Appendix B), which were then divided into individual and collaborative exploration phases [48]. Specifically, for each pair, one participant experienced four distinct phases: 1) individual exploration of the first topic, 2) individual exploration of the second topic, 3) collaborative exploration of the first topic, and 4) collaborative exploration of the second topic. Meanwhile, the other participants followed the exact same process, but during each phase, they worked on the other topic simultaneously (Figure 4). This setup not only simulated asynchronous collaboration but also eliminated the need for participants to wait for one another to complete work on a specific topic. Moreover, it enabled the collection of twice as many data points from their interactions during each phase.

The study was conducted via video conferencing software with each pair of participants. After signing the consent form, participants were introduced to the task. Then, participants were given a training video and task to familiarize themselves with all the functionalities of SenseSync. They were then asked to select two topics from five randomly chosen subjects spanning diverse domains (e.g., ethics of AI and advanced materials for space exploration), ensuring that the topics were unfamiliar to them. This was crucial for ensuring that the information-seeking tasks required complex search strategies and exploratory behaviors. Next, participants performed the task in the structure mentioned above. They were allocated 5 minutes for each individual phase and 10 minutes for each collaborative phase. Participants were instructed not to communicate directly with each other while performing the tasks and were required to use SenseSync's collaboration features for communication. Upon completion of the task, participants completed a questionnaire consisting of questions from UES-SF [33] to measure user engagement, along with additional questions to assess the perceived helpfulness of SenseSync's features. All questions were rated on a 7-point Likert scale [32]. Finally, a semi-structured interview was conducted with each pair of participants to collect qualitative feedback. The topics included general impressions and how SenseSync addressed the challenges identified in the formative study. The interviews were audio-recorded and the interactions with various components of SenseSync were logged. The whole study session lasted about 90 minutes and each participant received $25 for their time and effort.

6 RESULTS

We transcribed all interview sessions and applied deductive coding using NVIVO [7] to identify 180 codes, each corresponding to a specific research question. In the following, we first report the general impression of SenseSync and then discuss both quantitative and qualitative results in the context of the four themes guided by the challenges (C1-4) identified earlier, where old participants are denoted as P1-8 and new participants as N1-6.

6.1 General User Experience

Overall, participants felt that SenseSync was a useful tool to support collaborative information-seeking tasks with the involvement of LLMs. Figure 5 presents participants' ratings on different aspects of SenseSync, with most median ratings being 6 or higher. Ratings to each UES question as well as the average of UES scores ($MD = 6$, $IQR = 0$) show that SenseSync was highly user-engaging. This is further confirmed by the interaction statistics shown in Figure 6, which illustrate participants' interactions with all key components of SenseSync.

Notably, Figure 6 reveals that both the average number of interactions and the time spent on the task increased from the first to

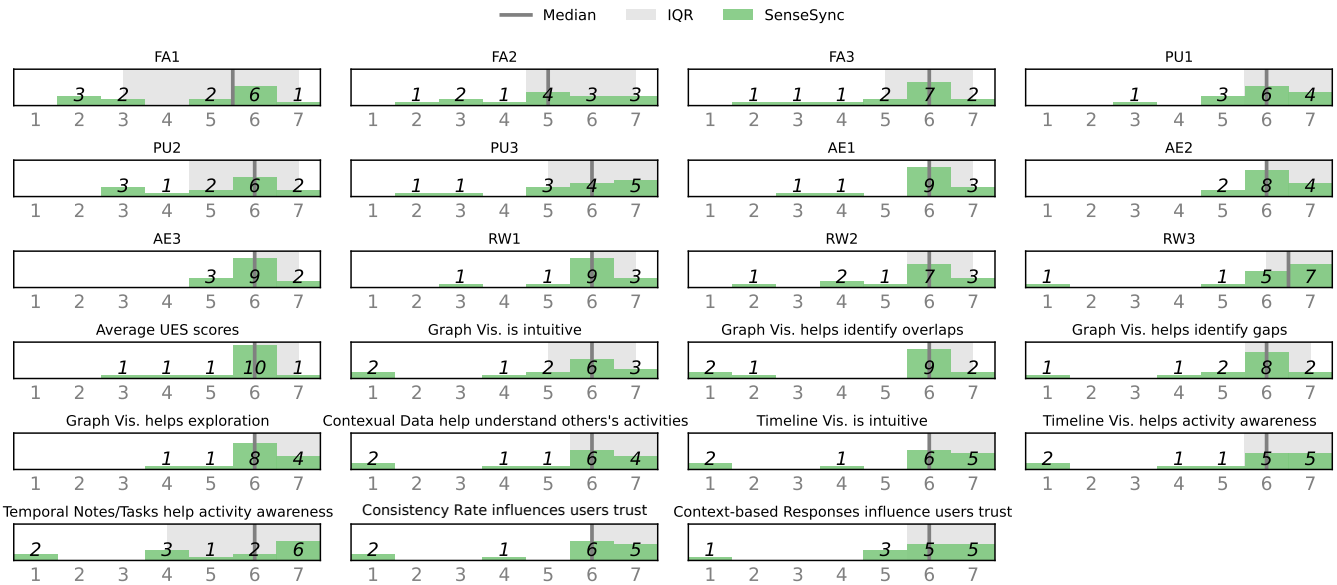


Figure 5: Participants' ratings based on a questionnaire including the UES-SF [33] to measure user engagement, and additional questions to evaluate the perceived helpfulness of SenseSync's features. The UES-SF questionnaire covers four aspects: Focused Attention (FA), Perceived Usability (PU), Aesthetic Appeal (AE), and Reward (RW).

the second phase of collaboration. This increase occurred as participants had more to work on, such as nodes, tasks, and notes, which increased their engagement with the key features. The data suggest that all key components were interacted with by participants, particularly Graph Visualization, Timeline Visualization, Context-based Responses, Auto-summarization, and Temporal Notes/Tasks. While the Graph Visualization was utilized in both individual and collaborative phases, participants used it significantly more during the collaborative phases. This difference highlights the graph's usefulness when conversations from multiple users are displayed, enabling participants to better understand the relationships between their findings. The Timeline Visualization, specifically designed for the collaboration phases, revealed a distinct usage pattern across two collaboration phases: Phase 3 – Collaboration and Phase 4 – Collaboration. Participants used this visualization more in the latter phase. This aligns with observations that, during the first collaboration phase, participants were encountering the collaborative space for the first time. At this stage, they were more curious to explore relationships between their findings and those of others using the graph, as well as assign tasks or notes. In contrast, by the second collaboration phase, the shared conversation had already been reviewed by their partners, and some tasks and notes had been assigned. As a result, participants increasingly relied on Temporal Visualization to understand the contextual information behind those tasks or notes, which provided them with valuable insights into their partner's thought process.

Both returning participants and new participants highlighted distinct aspects of SenseSync's utility in collaborative tasks. Returning participants, reflecting on their prior challenges, emphasized improvements in usability and workflow coherence. *"Compared to last time, this is wonderful. Last time was a complete loss, but this time the UI is pretty nice, and everything goes through smoothly."* -P8 Newcomers, while lacking comparative context, focused on the system's exploratory affordances, with one noting, *"I think the overall experience is very interesting because the diagram motivates me to explore more since it dynamically changes. I would like to see more changes by interacting with it."* -N5

6.2 SenseSync Helps Explore, Organize, and Leverage LLM-mediated Information (C1)

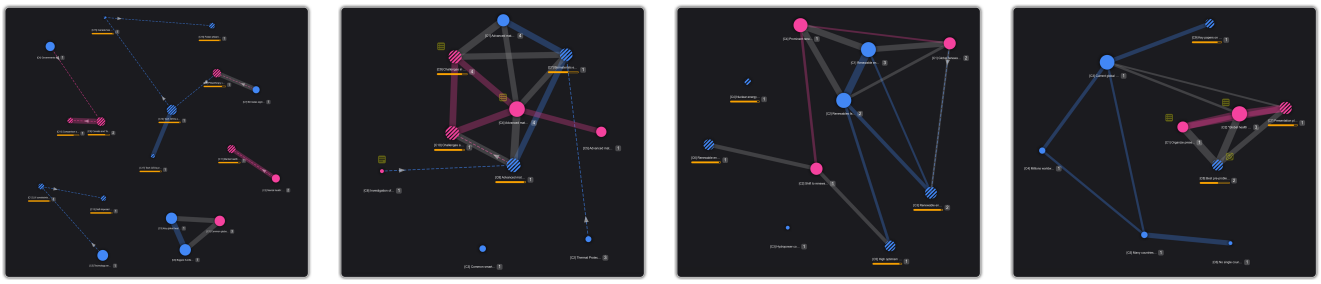
The Graph Visualization was rated highly intuitive by participants ($MD = 6$, $IQR = 1$). Figure 6 shows that participants engaged with the graph during both individual and collaborative phases. We can observe that they interacted with the graph during both collaborative phases but this occurred in only half (14 out of 28) of the individual phases. Participants mentioned that they had not had a chance to dive in the connections in the obtained information due to the short amount of time allocated for the individual phases.

Moreover, they thought the graph to be useful in identifying overlaps ($MD = 6$, $IQR = 0$). They noted that knowing about overlaps before reviewing their partner's LLM-generated information would save time by avoiding redundant reading. *"I find this function very useful, as there are many similar questions and answers, and I don't have time to read all of them."* -P1 Participants also rated the graph to be helpful in identifying gaps ($MD = 6$, $IQR = 0.5$) in two main ways. First, reviewing similar information from their partner helps identify new aspects to explore: *"Based on the nodes connected to mine, I also found that there are some interesting topics that I hadn't discovered."* -P1 Second, observing disconnected clusters of nodes reveals potential missing links: *"There were four separate graphs. By noticing the structure of the graph, you could think about how something might connect these separate graphs into a single unified graph."* -P2 In addition to the graph, some participants mentioned

Figure 6: Visualization of participants' interaction logs across four plots, each representing a task phase. The x-axis labels ([X]-[Y]Topic) indicate the participant ID (X) and first or second topic (Y). The y-axis shows the time spent by each participant on each topic. Data points, which signify interactions such as clicking or hovering, are color-coded based on the SenseSync components interacted with. The total number of interactions is shown at the end of each row. Examples of general components include task selection and workspace switching.

that auto-summarization helped them to identify the overlaps: *"By summarizing everything, I could see most of what P3 had asked."* -P4.

Participants rated the graph as helpful for collaborative exploration with LLMs ($MD = 6$, $IQR = 0.5$). This exploration can be enabled in four ways. Figure 7 presents examples of graphs created by participant pairs, highlighting various exploration strategies and patterns in shared information. First, they mentioned that being able to identify gaps using the graph helped them determine which aspects required further exploration: *"The graph helped me identify explored topics and focus on gaps with fewer connections between nodes."* -P2. Second, they found that SenseSync improved prompting by allowing them to view and be inspired by others' prompts: *"The fact that I can see what P3 is asking, and then ask follow-up questions myself, while she can see what I am asking and design her prompts similarly, is very helpful."* -P4 Third, seeing connections between collaborators' nodes highlighted relevant information and encouraged further exploration: *"When I see my partner's nodes connected to mine, it suggests her content might relate to my ideas, motivating me to explore her conversation and find new directions that could integrate with my own work."* -N6. Lastly, some participants stopped exploring upon seeing overlaps to save time. *"If a lot of people use the system, I think there will be a significant decrease in repeated work."* -P2



(a) (b) (c) (d)

Figure 7: Examples of graphs created by participant pairs using SenseSync, showcasing diverse patterns in the shared information. (a) P3 and P4, Topic 1: Depth vs. breadth in exploration. (b) N5 and N6, Topic 2: Further exploration of an isolated node in the collaborative workspace connects it to a cluster of nodes. (c) N1 and N2, Topic 2: Revealing the convergence in information shared by collaborators. (d) P7 and P8, Topic 1: Revealing the divergence in information shared by collaborators

6.3 SenseSync Enhances the Interpretation of Others' Work with LLMs (C2)

Participants agree that the contextual data integrated into the Graph and Timeline Visualizations was helpful in grasping others' understanding of LLM-generated information and collaborative work ($MD = 6$, $IQR = 1$). There are different types of contextual data that helped them in this regard.

First, knowing about the prompts in addition to the information generated and enabling users to view this information by hovering over the nodes in the graph helped them better understand their partner's generated information: *"I will be more curious about the prompting strategies she used. I will hover over each of the nodes to see her previous prompts and understand them better."* -N6 Second, the chain of prompts used for generating information, visualized as the flow of nodes using the dashed lines on the graph (Figure 7), was helpful for them to understand each other's thought process: *"I like the dashed part with the arrows; it was good that they show direction, indicating which one was created first and how the information is related."* -N3 Third, while the consistency rate was primarily intended to help users assess their trust in LLMs, a surprising finding emerged. Participants used the consistency rate to determine if they had a similar understanding with their partner. If most bars across different nodes indicated a high consistency rate, it suggested that the contexts were aligned, meaning collaborators might explore the topic similarly: *" I like the bar on the graph because it's clear and intuitive. High scores show that we've reached a consensus and that the content is consistent."* -N6 Fourth, participants found the date of issuing conversations or notes—encoded by their position in the Timeline Visualization—helpful for understanding the reasons behind their partner's decisions: *"By looking at the timeline, I noticed that P1 shifted his focus from Canada to Taiwan. I used the timeline to see when and why he changed his mind."* -P2 Lastly, participants found the overlapping lines in the timeline visualization, which indicates if a user worked on their partner's nodes, particularly helpful for understanding user activities and following their thought processes to explore new directions: *"When I see the straight lines followed by overlaps with other users' lines, I become curious about the cause of these overlaps. I then look into the specific notes on the Timeline Visualization to see if I can find any new direction."* -N6

6.4 SenseSync Eases Activity Switching, Recalling, and Resuming (C3)

Participants rated the Timeline Visualization as highly intuitive ($MD = 6$, $IQR = 1$). They rated both the Timeline Visualization ($MD = 6$, $IQR = 1.5$) and the notes/tasks feature ($MD = 6$, $IQR = 3$) highly useful in making them aware of their own and others' activities, and interactions with LLMs over time. In specific, regarding switching tasks, P5 noted, *"A bit later on, after I had done my research, I gave myself a task. Once I was done, and then I came back to the task."* P6 commented on recalling activities, *"I can use the Timeline Visualization to understand what was my thought process."* For resuming tasks, P1 mentioned, *"The timeline is useful because whenever I go back to the task, I can see a kind of summary of what the partner's activities have been."* -Lastly, they thought these features helped them track subtopics during the collaboration. *" We decomposed the task into subtopics, with each person working on one. The timeline helped track the progress of these subtopics."* -P1

Participants also highlighted other scenarios where the Timeline Visualization proved useful. For instance, when there are no time constraints for task completion: *"I would like to click into her past conversations and skim through what GPT's answers were"* -N6; when the task is a long-term multi-session one: *"If we were researching a topic together and then continued at the end of the month, we might come to different conclusions by then, as we would be researching differently"* -P4; when more than two collaborators are contributing: *"Maybe there are more than two people, perhaps ten, collaborating with each other. In that situation, this system might be more helpful"* -P4; and for synchronous tasks: *"I could see that the timeline is a bit more useful in a synchronous setup"* -P2.

6.5 SenseSync Facilitates the Assessment of LLM-generated Information (C4)

Participants rated the consistency rate provided in SenseSync as highly helpful in assessing their trust in LLM-generated information ($MD = 6$, $IQR = 1$). They mentioned it notified them when the information might not be trustworthy: *"The percentage helps you understand when you have doubts that the answer you're reading from ChatGPT is not what you expect."* -N3 After getting notified about potential trustworthiness issues, they also mentioned they would

verify the information either by asking the LLM for further clarification or by assigning the task to their partner: *"The percentage helped me decide whether I should go for another prompt [for verification purposes], or assign a task to the partner based on that."* -N3 In addition, P1 thought that if the information was about a factual topic, validation is necessary: *"I would say that the consistency rate is helpful when there are a lot of factuals or validations."* -

Also, participants rated the context-based response suggestion as useful in assessing their trust in LLM-generated information ($MD = 6$, $IQR = 1.5$). Figure 6 reveals that the participants interact with this feature in more than half of the participant-topic pairs conditions (17 out of 28). We found that a low consistency rate motivate the participants to use the context-based responses suggestion. They said if they saw a clue in the responses and feel the response generated by LLMs was not trustable, they could read alternative responses generated using other contexts to see if they could trust or not. *"I could not fully trust the system because in its answer it used the word 'potential,' which felt like a possible AI hallucination. I followed up by reviewing the context and other responses to verify responses remained consistent."* -P5

However, a surprising perception of the consistency rate was observed from participants. Some participants noted that they preferred not to read all responses with varying contexts to save time: *"If this number is low and I don't have time to read the responses, this number could help me skip some of them. I was receiving so many questions and answers."* -P1 In this case, a low consistency rate can be a signal of potential new useful information and participants would be motivated to explore the other responses to uncover valuable insights: *"If there is anything new, especially when the consistency scores are comparatively lower than those for other questions, then I know there might be some new content appearing in the collaborative space. I think that lower consistency scores actually motivate me to click into those responses."* -N6

7 DISCUSSION

7.1 Reflections

From our study, we have learned several insights into collaborative information-seeking when LLMs are actively involved. One of the main observations is that when using LLMs, the challenges in collaboration are amplified. This is because LLMs can generate diverse information based on different prompts and contexts. When collaborators engage in exploratory browsing—where they are unsure of what exactly they are looking for or how to effectively search for it—they often rely on LLMs to explore the topic. This process, however, generates a large volume of information, making it time-consuming to digest the information and collaborate. The interactive graph in SenseSync facilitated participants in this regard by connecting similar information, allowing them to focus on unique or dissimilar information. This was further empowered by the Context-based Responses to uncover new information and the Consistency Rate to determine if there is new information.

Additionally, it becomes harder to understand others' interpretation of the LLM-generated information, without contextual data. Contextual data integrated into the Graph Visualization such as prompts, the number of prompts, the flow of conversations, and the consistency rate, as well as Timeline Visualization, such as overlapping users' lines and dates of issuing conversations and notes, were found to be helpful in understanding others' thought processes, information, and collaborative work. However, our exploration of the effective features is just a start, and tools that effectively track and manage the generated information and their context need to be developed.

The lack of trust in LLMs, worsened by the higher risk of hallucinations in diverse conversational contexts, has signified another challenge in our study. Collaboration helps users spot inconsistencies in LLM responses and assess similarities. The integrated consistency rate in SenseSync flagged potential hallucinations by comparing response consistency across varied contexts, with consistent responses indicating potential trustworthy information. It would be interesting to investigate how the consistency rate and context-based response suggestions can be further enhanced to offer more comprehensive benefits to collaborators.

7.2 Potential Use Cases

SenseSync's design principles and features make it adaptable to diverse collaborative scenarios involving LLMs. We list some example use cases below, however, we believe SenseSync may be useful for a wider range of scenarios.

Graduate students or interdisciplinary research teams often collaborate on complex topics (e.g., climate change mitigation strategies or AI ethics) requiring extensive literature reviews and synthesis of diverse perspectives. SenseSync's dynamic graph can visualize overlapping insights from team members' LLM queries (e.g., "ethical frameworks for AI governance"), while the timeline tracks iterative explorations. The consistency rate helps flag conflicting LLM responses (e.g., differing definitions of "AI transparency"), prompting cross-validation. The auto-summarization and task assignment features are streamlined, enabling teams to efficiently compile the findings into cohesive reports.

Market research teams analyzing emerging trends (e.g., consumer behavior shifts in renewable energy adoption) can use SenseSync to coordinate asynchronous inquiries across global teams. The graph identifies gaps between regional insights (e.g., "solar panel adoption rates in Europe vs. Asia"), while the timeline reveals temporal patterns in data collection. Contextual prompts (e.g., "How did inflation impact EV sales in 2023?") and consistency checks ensure reliable inputs for strategic decision-making. Notes and task features facilitate handoffs between analysts, reducing redundancy and aligning stakeholders on actionable insights.

Medical teams managing rare disease diagnoses or treatment protocols could employ SenseSync to harmonize LLM-generated evidence (e.g., "latest CRISPR therapies for genetic disorders"). The graph highlights consensus or discrepancies in literature summaries (e.g., conflicting drug efficacy studies), while the timeline tracks diagnostic milestones. Clinicians can assign tasks (e.g., "verify side effects of Drug X") and use consistency rates to prioritize trustworthy sources. This fosters shared understanding among specialists (e.g., oncologists, pharmacologists) despite asynchronous workflows, improving patient outcomes through coordinated, evidence-based care.

7.3 Limitations and Future work

Our work is not without limitations. First, while we showed SenseSync's value in two-user collaborations with simple graphs over limited durations, its scalability—such as support for larger datasets, extended timelines, or multi-user scenarios—remained uninvestigated. Collaborative systems often require handling evolving use cases where visualization density or interaction inefficiencies could compromise usability. In such situations, both Graph and Timeline Visualizations may require more granular optimizations to maintain clarity and utility. To address this, Graph Visualization could incorporate dynamic scaling, hierarchical clustering, and interactive filtering to manage complexity, alongside progressive disclosure to reduce overload. For Timeline Visualization, integrating aggregated activity in minimaps and zoomable views could balance detail and overview, while redesigning layouts to accommodate additional users. These enhancements, coupled with future studies in complex scenarios, would strengthen the system's scalability and real-world applicability.

Second, our study design constrained the participants to complete tasks within certain time limits. While in reality, similar situations can occur during time-sensitive tasks, the time constraint for the exploration phases may restrict participants' ability to fully explore the topics and utilize all the features of the system. Therefore, our insights and observations from the results might be limited. Future studies could address this by extending the task duration to facilitate more thorough exploration and interaction, especially the behaviors working with LLMs.

Third, due to the study design, participants conducted all the tasks within a single session and between two users, which may not have captured the complete dynamics of real-world multi-user multi-session collaborative information-seeking with LLMs. Future research should consider how the system supports collaboration over extended periods, with multiple sessions and evolving information needs, in the wild with a deployment study. Moreover, while SenseSync is designed to support asynchronous collaboration scenarios, many of its features and visualization are useful for synchronous collaboration as well. However, we have not studied this case because our focus in this research is remote asynchronous collaborative information-seeking, which more frequently happens in our real-world tasks [63]. It would be interesting to conduct further studies on assessing how well SenseSync supports real-time synchronous collaboration with LLMs involvement and identifying additional challenges in such scenarios.

Fourth, SenseSync currently uses an empirically determined threshold for linking the nodes in the graph visualization. However, setting a proper threshold can be difficult. A lower threshold may result in an excessive number of connections, potentially overwhelming users, while a higher threshold could overlook significant overlaps. Future research should focus on exploring adaptive approaches that adjust the threshold based on factors such as the number of nodes or the complexity of the topics.

Last, while the promoted conversational context data in SenseSync was useful for participants to understand others' work, establishing the optimal number of nodes to provide as context for generating responses poses a challenge. Having fewer nodes for generating the context may not be able to offer enough information for participants to utilize. Including more nodes may enhance the validity of the generated information, but could lead to reduced user experience due to longer processing times. Future research should investigate methods for determining the ideal amount of context needed for LLMs to generate responses that effectively balance performance and accuracy.

8 CONCLUSION

This paper has presented a formative study on examining the challenges of collaborative information-seeking involving LLMs, which are influenced by the diverse contexts in which AI is used for information generation. To support users facing these challenges, we designed and developed SenseSync, an interactive system featuring a dynamic graph and a timeline visualization, both enriched with LLM-specific contextual data. This combination, which also includes specific features designed to enhance the collaborative experience, allows users to explore the collaborative conversation with LLMs across different time spans. Through a summative study, we gained insights into how users utilized SenseSync to address these challenges, which led to implications for designing future tools that utilize LLM-specific contextual data to enhance collaborative information-seeking experience.

ACKNOWLEDGMENTS

This work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant #RGPIN-2020-03966 and Alliance Grant #576742-22 as well as the Canada Foundation for Innovation (CFI) John R. Evans Leaders Fund (JELF) #42371. We acknowledge that much of our work takes place on the traditional territory of the Neutral, Anishinaabeg, and Haudenosaunee peoples. Our main campus is situated on the Haldimand Tract, the land granted to the Six Nations that includes six miles on each side of the Grand River.

REFERENCES

[1] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 304, 18 pages. https://doi.org/10.1145/3613904.3642016

[2] Marcia J. Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online Information Review* 13, 5 (1989), 407–424. https://doi.org/10.1108/eb024320

[3] Abbas Pirmoradi Bezanjani and Orland Hoeber. 2024. Enabling Exploratory Browsing using Dynamic Search Result Tagging, Highlighting, and Filtering. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (Sheffield, United Kingdom) *(CHIIR '24)*. Association for Computing Machinery, New York, NY, USA, 334–339. https://doi.org/10.1145/3627508.3638295

[4] John M. Carroll, Dennis C. Neale, Philip L. Isenhour, Mary Beth Rosson, and D.Scott McCrickard. 2003. Notification and awareness: synchronizing task-oriented collaborative activity. *International Journal of Human-Computer Studies* 58, 5 (2003), 605–632. https://doi.org/10.1016/S1071-5819(03)00024-7 Notification User Interfaces.

[5] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. RELIC: Investigating Large Language Model Responses using Self-Consistency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 647, 18 pages. https://doi.org/10.1145/3613904.3641904

[6] Hyo Jin Do, Rachel Ostrand, Justin D. Weisz, Casey Dugan, Prasanna Sattigeri, Dennis Wei, Keerthiram Murugesan, and Werner Geyer. 2024. Facilitating Human-LLM Collaboration through Factuality Scores and Source Attributions. https://doi.org/10.48550/arXiv.2405.20434 arXiv:2405.20434 [cs.HC]

[7] Andrew Edwards-Jones. 2014. Qualitative data analysis with NVIVO. *Journal of Education for Teaching* 40, 2 (2014), 193–195. https://doi.org/10.1080/02607476.2013.866724

[8] OpenAI et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[9] Marcos Fernández-Pichel, Juan C. Pichel, and David E. Losada. 2024. Search Engines, LLMs or Both? Evaluating Information Seeking Strategies for Answering Health Questions. arXiv:2407.12468 [cs.IR] https://arxiv.org/abs/2407.12468

[10] Raymond Fok, Nedim Lipka, Tong Sun, and Alexa F Siu. 2024. Marco: Supporting Business Document Workflows via Collection-Centric Information Foraging with Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 842, 20 pages. https://doi.org/10.1145/3613904.3641969

[11] OpenJS Foundation. 2024. *Node.js*. https://nodejs.org/

[12] Boris A. Galitsky. 2023. Truth-O-Meter: Collaborating with LLM in Fighting its Hallucinations. *Preprints* (July 2023). https://doi.org/10.20944/preprints202307.1723.v1

[13] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 838, 21 pages. https://doi.org/10.1145/3613904.3642139

[14] Roberto González-Ibáñez and Chirag Shah. 2011. Coagmento: A system for supporting collaborative information seeking. *Proceedings of the American Society for Information Science and Technology* 48, 1 (2011), 1–4. https://doi.org/10.1002/meet.2011.14504801336

[15] David Gotz. 2007. The ScratchPad: sensemaking support for the web. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 1329–1330. https://doi.org/10.1145/1242572.1242834

[16] Preben Hansen and Kalervo Järvelin. 2005. Collaborative Information Retrieval in an information-intensive domain. *Information Processing Management* 41, 5 (2005), 1101–1119. https://doi.org/10.1016/j.ipm.2004.04.016

[17] Hossein Hassani, Emmanuel Sirimal Silva, Stephane Unger, Maedeh TajMazinani, and Stephen Mac Feely. 2020. Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future? *AI* 1, 2 (2020), 143–155. https://doi.org/10.3390/ai1020008

[18] Paul M. Herceg, Timothy B. Allison, Robert S. Belvin, and Evelyne Tzoukermann. 2018. Collaborative exploratory search for information filtering and large-scale information triage. *Journal of the Association for Information Science and Technology* 69, 3 (2018), 395–409. https://doi.org/10.1002/asi.23961

[19] O Hoeber, M Harvey, M Momeni, A Pirmoradi, and D Gleeson. 2024. Exploratory search in digital humanities: a study of visual keyword/result linking. In *Proceedings of the Association for Information Science and Technology*. Wiley-Blackwell.

[20] Michael C. Hout, Megan H. Papesh, and Stephen D. Goldinger. 2013. Multidimensional scaling. *WIREs Cognitive Science* 4, 1 (2013), 93–103. https://doi.org/10.1002/wcs.1203 arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.1203

[21] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen Macneil. 2023. Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 97, 3 pages. https://doi.org/10.1145/3586182.3615796

[22] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 3, 20 pages. https://doi.org/10.1145/3586183.3606737

[23] Dinesh Kalla, Nathan Smith, Fnu Samaah, and Sivaraju Kuraku. 2023. Study and Analysis of Chat GPT and its Impact on Different Fields of Study. *International Journal of Innovative Science and Research Technology* 8, 3 (March 2023), 1–N/A. https://ssrn.com/abstract=4402499 Available at SSRN.

[24] Bum chul Kwon, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, Ji Soo Yi, and David S. Ebert. 2012. Evaluating the Role of Time in Investigative Analysis of Document Collections. *IEEE Transactions on Visualization and Computer Graphics* 18, 11 (2012), 1992–2004. https://doi.org/10.1109/TVCG.2012.89

[25] LMS Lau, Vania G Dimitrova, Fan Yang-Turner, Manolis Tzagarakis, N Karacapilidis, and Spyros Christodoulou. 2014. Understanding collaborative sensemaking behaviour using semantic types in interaction data. In *Smart Digital Futures 2014*. IOS Press, 190–199. https://doi.org/10.3233/978-1-61499-405-3-190

[26] Narges Mahyar and Melanie Tory. 2014. Supporting Communication and Coordination in Collaborative Sensemaking. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 1633–1642. https://doi.org/10.1109/TVCG.2014.2346573

[27] Gary Marchionini. 2006. Exploratory search: From finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46. https://doi.org/10.1145/1121949.1121979

[28] Gary Marchionini. 2019. Search, sense making and learning: closing gaps. *Information and Learning Science* 120, 1 (2019), 74–86. Copyright - © Emerald Publishing Limited 2018; Last updated - 2024-04-18; SubjectsTermNotLitGenreText - Indexing; Information Retrieval; Developed Nations; Research and Development; Probability; Electronic Equipment; Information Needs; Workstations; Information Seeking; Search Engines; Information Services.

[29] MongoDB. 2024. *MongoDB: The Developer Data Platform | MongoDB*. https://www.mongodb.com/

[30] Phong H. Nguyen, Kai Xu, Andy Bardill, Betul Salman, Kate Herd, and B.L. William Wong. 2016. SenseMap: Supporting browser-based online sensemaking through analytic provenance. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 91–100. https://doi.org/10.1109/VAST.2016.7883515

[31] Phong H. Nguyen, Kai Xu, Ashley Wheat, B.L. William Wong, Simon Attfield, and Bob Fields. 2016. SensePath: Understanding the Sensemaking Process Through Analytic Provenance. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 41–50. https://doi.org/10.1109/TVCG.2015.2467611

[32] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education* 15 (2010), 625–632. https://doi.org/10.1007/s10459-010-9222-y

[33] Heather O'Brien. 2016. *Theoretical Perspectives on User Engagement*. Springer International Publishing, Cham, 1–26. https://doi.org/10.1007/978-3-319-27446-1_1

[34] Observable. 2024. *D3 | The JavaScript library for bespoke data visualization*. https://d3js.org

[35] Observable. 2024. *D3-force*. https://d3js.org/d3-force

[36] OpenAI. 2024. *ChatGPT*. https://openai.com/chatgpt/

[37] 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* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1771–1780. https://doi.org/10.1145/1518701.1518974

[38] Sharoda A. Paul and Madhu C. Reddy. 2010. Understanding together: sensemaking in collaborative information seeking. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 321–330. https://doi.org/10.1145/1718918.1718976

[39] Mohammad Hasan Payandeh, Miriam Boon, Dale Storie, Veronica Ramshaw, and Orland Hoeber. 2023. Drag-and-Drop Query Refinement and Query History Visualization for Mobile Exploratory Search. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 432–437. https://doi.org/10.1145/3576840.3578282

[40] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[41] Meta Platforms. 2024. *React - A JavaScript library for building user interfaces*. https://legacy.reactjs.org/

[42] Yan Qu and Derek L Hansen. 2008. Building shared understanding in collaborative sensemaking. In *Proceedings of CHI 2008 Sensemaking Workshop*.

[43] Nitin Rane, Saurabh Choudhary, and Jayesh Rane. 2024. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. *Performance, Architecture, Capabilities, and Implementation (February 13, 2024)* (2024). https://doi.org/10.2139/ssrn.4723687

[44] Material UI SAS. 2024. *MUI: The React component library you always wanted*. https://mui.com/

[45] Chirag Shah. 2009. Toward Collaborative Information Seeking (CIS). arXiv:0908.0709 [cs.IR] https://arxiv.org/abs/0908.0709

[46] Chirag Shah. 2010. *Collaborative Information Seeking: A Literature Review*. Advances in Librarianship, Vol. 32. Emerald Group Publishing Limited, 3–33. https://doi.org/10.1108/S0065-2830(2010)0000032004

[47] Chirag Shah. 2010. Collaborative Information Seeking: A Literature Review. In *Advances in Librarianship*, Anne Woodsworth (Ed.). Vol. 32. Emerald Group Publishing Limited, 3–33. https://doi.org/10.1108/S0065-2830(2010)0000032004

[48] Chirag Shah. 2014. Collaborative information seeking. *Journal of the Association for Information Science and Technology* 65, 2 (2014), 215–236. https://doi.org/10.1002/asi.22977 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22977

[49] Chirag Shah and Gary Marchionini. 2010. Awareness in collaborative information seeking. *Journal of the American Society for Information Science and Technology* 61, 10 (2010), 1970–1986. https://doi.org/10.1002/asi.21379 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21379

[50] StrongLoop and IBM. 2024. *Express.js*. https://expressjs.com/

[51] 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* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. https://doi.org/10.1145/3586183.3606756

[52] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 680, 24 pages. https://doi.org/10.1145/3613904.3642902

[53] Yihan Tao and Anastasios Tombros. 2017. How collaborators make sense of tasks together: A comparative analysis of collaborative sensemaking behavior in collaborative information-seeking tasks. *Journal of the Association for Information Science and Technology* 68, 3 (March 2017), 609–622. https://doi.org/10.1002/asi.23693

[54] Tan Thongtan and Tanasanee Phienthrakul. 2019. Sentiment Classification Using Document Embeddings Trained with Cosine Similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi (Eds.). Association for Computational Linguistics, Florence, Italy, 407–414. https://doi.org/10.18653/v1/P19-2057

[55] Karthikeyan Umapathy. 2010. Requirements to support collaborative sensemaking. In *CSCW CIS Workshop*, Vol. 10.

[56] Ryen W. White and Resa A. Roth. 2009. *Exploratory Search: Beyond the query-response paradigm*. Morgan and Claypool Publishers, San Rafael, CA. https://doi.org/10.2200/S00174ED1V01Y200901ICR003

[57] Thomas Wilson. 1999. Models in information behaviour research. *Journal of Documentation* (1999). https://doi.org/10.1108/EUM0000000007145

[58] Youfu Yan, Yu Hou, Yongkang Xiao, Rui Zhang, and Qianwen Wang. 2024. KNOWNET: Guided Health Information Seeking from LLMs via Knowledge Graph Integration. https://doi.org/10.48550/arXiv.2407.13598 arXiv:2407.13598 [cs.HC]

[59] Ryan Yen, Nicole Sultanum, and Jian Zhao. 2024. To Search or To Gen? Exploring the Synergy between Generative AI and Web Search in Programming. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 327, 8 pages. https://doi.org/10.1145/3613905.3650867

[60] Ryan Yen and Jian Zhao. 2024. Memolet: Reifying the Reuse of User-AI Conversational Memories. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 1, 12 pages. https://doi.org/10.1145/3654777.3676388

[61] Jian Zhao, Mingming Fan, and Mi Feng. 2022. ChartSeer: Interactive Steering Exploratory Visual Analysis With Machine Intelligence. *IEEE Transactions on Visualization and Computer Graphics* 28, 3 (2022), 1500–1513. https://doi.org/10.1109/TVCG.2020.3018724

[62] Jian Zhao, Michael Glueck, Petra Isenberg, Fanny Chevalier, and Azam Khan. 2018. Supporting Handoff in Asynchronous Collaborative Sensemaking Using Knowledge-Transfer Graphs. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 340–350. https://doi.org/10.1109/TVCG.2017.2745279

[63] Jian Zhao, Michael Glueck, Petra Isenberg, Fanny Chevalier, and Azam Khan. 2018. Supporting Handoff in Asynchronous Collaborative Sensemaking Using Knowledge-Transfer Graphs. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 340–350. https://doi.org/10.1109/TVCG.2017.2745279

[64] Tao Zhou and Songtao Li. 2024. Understanding user switch of information seeking: From search engines to generative AI. *Journal of Librarianship and Information Science* 0, 0 (2024), 09610006241244800. https://doi.org/10.1177/09610006241244800 arXiv:https://doi.org/10.1177/09610006241244800

A FORMATIVE STUDY DETAILS

A.1 Participant Characteristics

Table 1: Typical tasks for LLM usage by participants.

Task Category	Sub-category	# of Participants
Writing	General writing	5
	Grammar correction	4
	Paraphrasing	3
	Outlining	1
Information-seeking	Clarification	4
	Learning	2
	General IS	1
	Travel planning	1
Programming	General programming	6
Data analysis	General data analysis	1

Purple represents information-seeking tasks.

Table 2: Typical collaborative tasks by participants.

Use of LLMs	Task	# of Participants
Without	Course projects	4
	Programming	2
	Research	1
	Writing a document	1
With	Learning	1
	Programming	1

Purple represents information-seeking tasks.

A.2 Topics

They were instructed to choose one of the following topics that they found challenging to explore and were uncertain about.

- Global health challenges
- Ethics of artificial intelligence
- Quantum computing applications in everyday life
- Advanced materials for space exploration
- Renewable energy sources

A.3 Tasks

In a collaborative information-seeking endeavor, imagine your team, consisting of you and your partner(s), has been tasked with exploring and finding information on **[TOPIC]** for a presentation to your company's manager. Utilizing ChatGPT as your main tool to gather relevant data and insights, Zoom as a communication tool, and another tool for saving, sharing, and organizing the conversations from ChatGPT with your partner (such as Google Docs), your team's goal is to gather comprehensive information on various **[TOPIC]**. As a general guideline for the task, you should begin by individually working with ChatGPT to gather initial insights (for about 10 minutes). Following this, share your findings with your partner and continue the exploration collaboratively (for about 20 minutes).

A.4 Interview Questions

A.4.1 Main Questions.

1. What are your general thoughts and insights?
2. Can you identify any significant challenges you encountered?
3. What were possible system requirements/features you felt were necessary to enhance your collaboration and task completion?
4. - I'll list some of the sub-tasks you've performed. I'd like to hear about your approach, any difficulties you encountered, and any suggestions or requirements you have for improving each subtask:
   - Sharing of individual conversations with your partner
   - Aggregation of individual conversations
   - Sharing understanding and thought processes
   - Collaboratively evaluation of shared findings
   - Identifying overlaps and gaps to be able to achieve your goal, after the aggregation of the conversations
   - Collaborative navigation and exploration
5. Do you have any further thoughts and insights? Additionally, do you have any suggestions for enhancing task completion?

A.4.2 Example Follow-up Questions).

1. Do you believe that metadata (data about the conversations with ChatGPT), such as the time or the order of conversation issuance, can help you make sense of other users' findings? If yes, can you think of any types of metadata that would be helpful?
2. How do you perceive the effectiveness of incorporating various data types (such as images, voice recordings, URLs) or structures (like tables, lists) in sharing of the understanding of the information?
3. Suppose you need to collaborate with your partner to achieve your goal over a long period and multiple sessions. What would be your approach to resuming the task effectively? What challenges do you think you would face?

B SUMMATIVE STUDY DETAILS

B.1 Topics

Participants were instructed to select two topics from the provided list that they found challenging to explore. If they had participated in the formative study, they were advised not to select the same topic they had chosen during that study.

- Global health challenges
- Ethics of artificial intelligence
- Quantum computing applications in everyday life
- Advanced materials for space exploration
- Renewable energy sources

B.2 Tasks

In a collaborative information-seeking endeavor, your team, consisting of you and your partner(s), has to utilize a system to perform a task about exploring and finding information on various aspects of a topic that you need for a report. The task consists of two phases (individual and collaborative exploration) and you do this task with two different topics. Here is the general procedure of the study: Explore the first topic, individual phase (5 minutes) Explore the second topic, individual phase (5 minutes) Explore the first

topic, collaborative phase (10 minutes) Explore the second topic, collaborative phase (10 minutes) During either the individual or collaborative phases of the task, you are not allowed to talk to each other. However, during the collaborative phase of the task, you need to use the collaboration features of the system to coordinate with your partner. Later in the study procedure, you need to choose two topics.

B.3 Post-task Questionnaire

Participants were asked to respond to each question using a 7-point Likert scale, ranging from Strongly Disagree (1) to Strongly Agree (7), to indicate the extent to which they agreed with the statements.

1. I lost myself in this experience.
2. The time I spent using the interface just slipped away.
3. I was absorbed in this experience.
4. I felt frustrated while using this interface.
5. I found this interface confusing to use.
6. Using this interface was taxing.
7. This interface was attractive.
8. This interface was aesthetically appealing.
9. This interface appealed to my senses.
10. Using the interface was worthwhile.
11. My experience was rewarding.
12. I felt interested in this experience.
13. The visual representations of data in the "Graph Visualization" are intuitive.
14. The "Graph Visualization" is helpful for identifying overlaps.
15. The "Graph Visualization" is helpful for identifying gaps.
16. The "Graph Visualization" is helpful for further exploration.
17. The "Context-based Response Suggestion" is helpful for further exploration.
18. The visual representations of data in the "Temporal Visualization" are intuitive.
19. The "Temporal Visualization" is helpful for understanding others' activities.
20. The "Temporal Visualization" is helpful for resuming the task.
21. The "Temporal Note Taking/Task Assignment/Summaries" is helpful for understanding others' activities.
22. The "Temporal Note Taking/Task Assignment/Summaries" is helpful for resuming the task.
23. The "consistency Rate" is helpful for assessing ChatGPT-generated responses.
24. The "Temporal Note Taking/Task Assignment" is helpful for assessing ChatGPT-generated responses.

B.4 Interview Questions

1. What are your general thoughts and insights regarding your experience with the system and performing the task?
2. How did you find the system in helping you make sense of your and your partner's findings?
3. Were you able to identify the overlaps between your partner's findings and your own using the system? Why or why not?
4. Were you able to identify the gaps between your partner's findings and your own using the system? Why or why not?
5. Were you and your partner able to explore new information or areas that needed further investigation using the system? Why or why not?
6. How did you find the Context-based Response Suggestion feature in supporting your exploration?
7. How did the visual elements in the graph related to the contextual data specific to LLMs (e.g., the number of prompt/response pairs in a conversation or consistency rate) support you make sense of the information specific to LLMs? Why did or why didn't?
8. How did you find the system in supporting collaborative work?
9. Were you able to gain a good understanding of your partner's activities? Why or why not?
10. Were you able to resume the task during the collaborative phase? Why or why not?
11. How did the visual elements in the temporal visualization related to the contextual data (e.g., the order and date that conversation or notes/tasks/summaries initiated) support you to resume your task? Why did or why didn't?
12. How did you find the system in supporting your assessment of ChatGPT responses?
13. How did the consistency rate affect your assessment of the ChatGPT responses? Please explain.
14. Were you able to utilize the note-taking/task assignment feature to assess the ChatGPT responses? Why or why not?
15. Is there anything that the system can improve?