Enhancing Immersive Sensemaking with Gaze-Driven Recommendation Cues

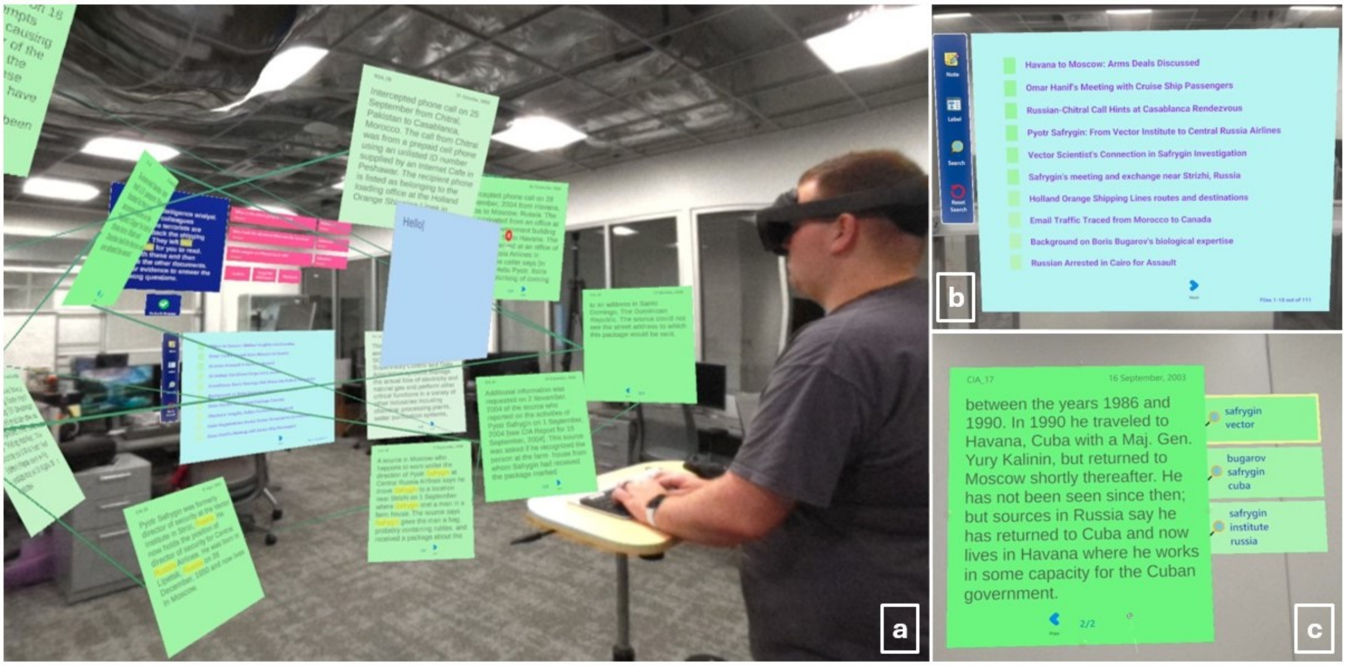
Ibrahim Asadullah Tahmid
Computer Science
Virginia Tech
Blacksburg, Virginia, USA
iatahmid@vt.edu

Chris North
Department of Computer Science
Virginia Tech
Blacksburg, Virginia, USA
north@vt.edu

Kylie Davidson
Center for Human-Computer Interaction
Virginia Tech
Blacksburg, Virginia, USA
kyliedavidson@vt.edu

Kirsten Whitley
Department of Defense
College Park, Maryland, USA
visual@tycho.ncsc.mil

Doug Bowman
Center for Human Computer Interaction
Virginia Tech
Blacksburg, Virginia, USA
dbowman@vt.edu



- Havana to Moscow: Arms Deals Discussed
- Omar Hanif's Meeting with Cruise Ship Passengers
- Russian-Chitral Call Hints at Casablanca Rendezvous
- Pyotr Safrygin: From Vector Institute to Central Russia Airlines
- Vector Scientist's Connection in Safrygin Investigation
- Safrygin's meeting and exchange near Strizhi, Russia
- Holland Orange Shipping Lines routes and destinations
- Email Traffic Traced from Morocco to Canada
- Background on Boris Bugarov's biological expertise
- Russian Arrested in Cairo for Assault

between the years 1986 and 1990. In 1990 he traveled to Havana, Cuba with a Maj. Gen. Yury Kalinin, but returned to Moscow shortly thereafter. He has not been seen since then; but sources in Russia say he has returned to Cuba and now lives in Havana where he works in some capacity for the Cuban government.

Figure 1: EyeST offers a set of recommendation cues to help analysts in the sensemaking task. a) Document color represents their global interest to the analyst and green threads represent connections between documents. b) The list of unread documents is sorted by their global interest. c) Recommended documents related to a given document are represented as tabs. The tabs display the words that are shared with the recommended document. The background color of the tabs represents their global interest.

Abstract

Sensemaking is a complex task that places a heavy cognitive demand on individuals. With the recent surge in data availability, making sense of vast amounts of information has become a significant challenge for many professionals, such as intelligence analysts. Immersive technologies such as mixed reality offer a potential solution by providing virtually unlimited space to organize data. However,

IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712103

the difficulty of processing, filtering relevant information, and synthesizing insights remains. We proposed using eye-tracking data from mixed reality head-worn displays to derive the analyst's perceived interest in documents and words, and convey that part of the mental model to the analyst. The global interest of the documents is reflected in their color, and their order on the list, while the local interest of the documents is used to generate focused recommendations for a document. To evaluate these recommendation cues, we conducted a user study with two conditions: a gaze-aware system, EyeST, and a "Freestyle" system without gaze-based visual cues. Our findings reveal that the EyeST helped analysts stay on track by reading more essential information while avoiding distractions. However, this came at the cost of reduced focused attention and perceived system performance. The results of our study highlight the need for explainable AI in human-AI collaborative sensemaking to build user trust and encourage the integration of AI outputs into the immersive sensemaking process. Based on our findings, we offer a set of guidelines for designing gaze-driven recommendation cues in an immersive environment.

CCS Concepts

• **Human-centered computing → Interaction techniques**; **User interface design**; *Visualization*; • **Information systems → Recommender systems**; Personalization; • **Computing methodologies → Distributed artificial intelligence**.

Keywords

Immersive Sensemaking, AI-Mediated Sensemaking, Human-AI Collaboration, Gaze-Driven Recommendations, Rich Semantic Interaction

ACM Reference Format:
Ibrahim Asadullah Tahmid, Chris North, Kylie Davidson, Kirsten Whitley, and Doug Bowman. 2025. Enhancing Immersive Sensemaking with Gaze-Driven Recommendation Cues. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3708359.3712103

1 Introduction

Visual Analytics (VA) is *"the science of analytic reasoning facilitated by interactive visual interfaces"* [74]. By coupling human intuition with computational analysis, VA enables humans to address challenges that would otherwise be difficult due to their scale and complexity. The primary objective of VA is to provide tools that: (a) synthesize information and extract insights, (b) detect expected patterns and reveal unexpected correlations, (c) deliver timely, defensible, and understandable assessments, and (d) effectively communicate these assessments [41].

Building on the concept of visual analytics, the field of Immersive Analytics (IA) aims to enhance the interaction between analysts and data by offering an embodied experience in a fully-surrounding three-dimensional space. IA extends beyond traditional interfaces, creating a three-dimensional environment where analysts can explore, analyze, and derive insights more efficiently. The primary goal of immersive analytics is to "*explore the applicability and development of emerging user-interface technologies to create more engaging and immersive experiences and seamless workflows for data analysis applications*" [12]. Skarbez et al. further linked IA to the sensemaking process, defining it as "*the science of analytic reasoning facilitated through immersive human-computer interfaces*," [71]. These interfaces are designed to support the synthesis of information through abstract data visualizations while taking advantage of embodied interaction. Unlike traditional VA systems, IA systems introduce new possibilities, such as enhanced depth cues, a 360-degree interaction space with six degrees of freedom, fewer distractions, improved spatial comprehension, and integrated eye-tracking sensors [8, 54].

One thread of IA research called *Immersive Space to Think (IST)* focuses on using immersive space to organize and reason about documents, notes, and labels [49]. The immersive spatial layout lets the analyst project their mental model into a virtual space, allowing them to externalize and interact with their thoughts, much like handling a physical object.

While the IST concept enhances the sensemaking process with memory externalization and improved recall [44], sensemaking can still be overwhelming as the dataset expands. Hence, IA tools need to evolve to provide additional intelligent aids to analysts to help them manage and understand larger datasets. Wenskovitch et al. envisioned future VA/IA tools as *"a mutually intelligible communication channel between humans and AI/ML* [1] *models, where human and machine teammates are in sync with their roles and responses to each other's actions"* [80]. Their vision highlights two-way communication in which the AI learns from analyst interactions to improve human-machine collaboration and enhance team performance. Through explicit or implicit feedback from the analyst, the machine develops an understanding of the analyst's strategy, allowing it to act as a teammate, rather than an all-knowing entity.

To better define the roles of humans and machines, Sheridan and Verplank proposed the concept of Level of Automation (LoA) in human-centered AI decision-making tasks [70], which was later refined by Parasuraman et al. [60]. Their taxonomy describes ten levels, ranging from full human control for *LoA(1)* to complete machine autonomy for *LoA(10)*. Mackeprang et al. demonstrated that users perform better but face confusion with lower LoA systems, while higher LoA systems provide a seamless but error-prone experience [53], causing them to suggest choosing LoA(5) as the sweet spot for human-AI collaborative ideation tasks. However, finding the sweet spot for automation still depends largely on the domain.

Building on the search for the optimal LoA, VA tools have introduced *semantic interaction*, where the system learns from explicit analyst actions during sensemaking, such as highlighting text or grouping documents, to build an interest model for the analyst. The model can be used later for automatic layout adjustments [24], and/or generating smart recommendations aligned with the analyst's topic of interest [9]. IA systems, with their integrated sensors, offer the potential to advance this concept by introducing *rich* semantic interaction, in which the system could model analyst interest based on implicit, embodied interactions, such as gaze. Gaze can be an implicit, non-intrusive way to gather valuable insights about a person's mind, revealing their thought process during cognitively heavy sensemaking tasks, and even indicating the

[1] AI=Artificial Intelligence, ML=Machine Learning

perceived relevance of different topics of interest within a complex, interconnected dataset [73].

We utilized the gaze-based metric presented by Tahmid et al. [73] in an IST-like IA system to develop an intelligent recommendation system, driven by analyst's gaze data during sensemaking in the immersive space. We call this **Eye**-Enhanced Immersive **S**pace to **T**hink, or EyeST for short. We developed two levels of automation for different kinds of recommendations. First, EyeST tracks the analyst's overall interest in real time, reordering the list of unread documents in decreasing order of the analyst's global interest, and color-coding all documents based on their global interest. This cue operates at a high LoA, as it is unobtrusive and requires little to no analyst intervention to be of use. Second, EyeST generates a list of recommendations for each document, based on both the analyst's global and local interest in that specific document. It is up to the analyst to decide which of these recommendations they want to read. These local recommendation cues operate at a low LoA, offering some automated support, but leaving the analyst to make the final decision based on the cues.

We conducted a user study to compare the gaze-based recommendation system (EyeST) with a control condition offering no recommendation cues (Freestyle IST). Our findings demonstrate that the recommendation cues helped analysts manage their time more efficiently by reading more essential documents, whereas Freestyle analysts were often sidetracked by distractor information. However, EyeST analysts were often confused by the local recommendations, leading to reduced attention during sensemaking and a decrease in perceived task performance. Feedback from analysts allowed us to identify the design challenges of the local recommendations that led to these results. We conclude by offering guidelines for improving gaze-based recommendation cues in immersive analytics tools. In summary, the following are the contributions of this work.

(1) A gaze-driven intelligent recommendation system for IA tools.
(2) The design of recommendation cues for IA to reflect the analyst's global and local interest.
(3) An understanding of the benefits and challenges of gaze-driven recommendation cues.
(4) Design guidelines for improving recommendation cues in IA tools.

2 Related Work

In this section, we review prior research related to immersive sensemaking and gaze-driven recommendations.

2.1 Sensemaking and Semantic Interaction

Pirolli and Card defined sensemaking as a complex cognitive task that involves browsing unstructured information, extracting meaningful evidence, and synthesizing new insights about one or more topics in a set of documents [64]. This series of actions can be divided into two major parts. First, an analyst collects evidence by browsing the dataset (foraging). Second, the analyst spends time on organizing and synthesizing information (sensemaking) [78]. Visual Analytics (VA) focuses on developing interactive tools to support sensemaking by allowing analysts to read, annotate, organize, and synthesize in a visual, often spatial layout. To address the handling of large amounts of data, researchers have used large, high-resolution displays that can become part of the distributed cognitive process, providing both external memory and a semantic layer [4]. However, for real-world sensemaking tasks, the dataset still may not be able to fit on large two-dimensional displays. Lisle et al. [49] proposed the use of an immersive, expansive, three-dimensional space to analyze large multimedia datasets, and called it Immersive Space to Think (IST). The immersive experience allowed analysts to follow creative spatial organization strategies [50] during different stages in the sensemaking process [21], and improve their overall understanding of the dataset [49].

Space solves the issue of seeing and organizing the whole dataset. However, analysts still need assistance to find information and synthesize information from the data. ForceSPIRE [24] proposed statistical models steered by *semantic interaction* where the model learns from the analyst's actions such as searching, highlighting, annotating, and repositioning documents during sensemaking, and co-creates the spatial layout with the analyst. StarSPIRE [9] built on this idea to utilize the analyst-perceived information relevance from the semantic interaction to develop a relevance-based foraging model. The underlying assumption for these models is that if an analyst highlights or searches for a term, it is considered 'relevant' to their cognitive process, and they would be more interested in exploring documents about similar topics [63]. Here, 'relevance' reflects the perceived closeness in meaning between the term and the task at hand. Analysts are also prone to keep 'similar' documents in close proximity in both large 2D displays [9] and immersive spaces [47, 72]. All of these analytic tools rely on explicit analyst interactions to enhance the analyst's sensemaking process. With mixed reality headsets with built-in sensors such as eye-trackers, IA research has a new window of opportunity to use subtler, more implicit user interactions to build intelligent models for sensemaking [11, 52].

2.2 Rich Semantic Interaction

One of the primary motivations behind developing semantic interactions was to keep the analyst focused on their cognitive process while the system takes care of the intermediate steps such as spatial organization [24], and information retrieval [9]. With traditional computing setups, the only way for the system to infer the analyst's intent is through cursor movement or keyboard input, but human behaviors have much more nuance that could be leveraged. Researchers have shown that a system can infer the user's intent from a variety of human behaviors such as motion/gesture [6, 68], speech [22, 27], eye gaze patterns [19, 48], and brain activity [13, 58]. Most of these actions, if not all, can be captured by a singular device with Mixed Reality technology. This opens up a new avenue in the field of semantic interaction where the system can infer the analyst's intent from natural interactions in an immersive environment. We call this rich semantic interaction, defined as follows.

> ***Rich semantic interaction*** *is a mode of user-system interaction where the system can infer the user's intentions from a wide range of natural human interactions*

> *in the immersive space, such as motion, speech, eye gaze, and brain signals.*

One aspect of the sensemaking process that can be enhanced through such high-level inference is information retrieval [9], which helps analysts browse the dataset more efficiently and effectively. Such an inference would involve the system being able to identify the user perception of each document they are going through, their topics of interest, and their path of reasoning towards the solution of the sensemaking task. We suggest that eye gaze can provide insight into each of these.

2.3 Role of Eye Gaze in Reading

Eye gaze has been shown to be closely linked to cognitive processing [38, 39, 75], mental workload [62], reading comprehension [1], and emotional expression [79]. Just and Carpenter [38] found negligible lag between eye fixations and cognitive processing, causing them to suggest that what we see is also what we think about. We can infer a lot about a person's thinking process from their gaze behavior. Fixation duration alone is a strong measure to distinguish novice users from experts [2, 34], infer student engagement in extracting and processing information from a set of given sources [32], distinguish reading behaviors of users for different tasks such as comprehension and proofreading [40], and even predict query terms during information processing with high accuracy [18]. Fixation count has also been associated with fixation duration in cognitively processing a word [33, 66], identifying readers with high recallability [77], and identifying the relevance of specific areas of interest to readers [45]. More recent studies found the effect of more sophisticated gaze measures such as increased pupil size for novice users compared to experts [2], and gaze velocity being able to predict users' intent to interact [19]. All of these studies, however, focus on analyzing users' gaze behaviors while reading single sentences or single documents.

Research on users' gaze behaviors while reading and processing information from multiple sources [35] is relatively underexplored. During everyday sensemaking tasks, in addition to comprehending individual pieces of text, people spend a lot of time searching for information from diverse sources and integrating them to answer questions [10]. Thus, in real-world scenarios, the ability to make predictions from eye gaze measures based on reading individual pieces of text is challenging [35], even with advanced deep neural networks [1]. In addition, solely analyzing eye movements does not provide insight into the user's reasoning process for multiple documents as it introduces frequency bias [37]. In a dataset with multiple documents, a word can appear in different documents in different contexts, not all of which are relevant to the users. Due to its high frequency, the word may end up having a larger total fixation duration regardless of how the user perceives that word. Tahmid et al. addressed this issue by proposing a gaze metric for sensemaking tasks involving multiple inter-connected documents, where they handled the frequency bias and document length bias, and combined the fixation duration and fixation count to measure user-perceived information relevance during sensemaking tasks [73]. In our work, we aim to leverage this inferred-relevance by generating intelligent recommendations for sensemaking analysts.

2.4 User-Centered Information Retrieval

It comes as no surprise that systems with intelligent document retrieval features have been studied extensively [9, 14, 24, 81, 84]. They showed that a user's interactions have implicit meanings that help to reveal their information-seeking strategy [83]. For instance, users' interactions with a list of searched documents can provide an understanding of how the searcher's information needs change over time [81]. The searched term itself can help a system to determine which documents the user would be interested in [14]. This may introduce a term-matching problem where the searched term and the index terms may not match exactly. *Phrasier* [36] tried to address this problem by automatically exploiting predetermined keyphrases from the source documents to create links to similar documents. However, this approach puts the machine in the driving seat and takes some control away from analysts.

Another approach is to rely more on implicit user actions such as reading time [42, 56], browsing patterns [69], scrolling time [17, 42], or mouse movement [17] to estimate the user-perceived relevance of terms in the documents. Our hypothesis is that the gaze measures in the immersive space could be used in a similar way. Fixation time, for instance, is quite effective in predicting the relevance of individual Web pages [26], and predicting relevant search terms [18, 81] in information-retrieval tasks. The underlying assumption is that the time spent reading a word reflects the user's cognitive processing of that word [33, 65, 66]. This principle has been proven effective in predicting relevance for words [51], paragraphs [7], and documents [28] read by users. McNamara et al. [55] used eye tracking to measure user attention to objects of interest and place labels in an information-rich environment. However, Drusch et al. [23] showed that the user's area of interest changes over time and requires dynamic visualization techniques to accurately represent the user's interest.

In summary, eye gaze data can predict analysts' perception of information relevance. However, these predictive models cannot be directly transferred to sensemaking tasks due to the interconnected nature of the dataset, and the analyst's evolving interest during a given task. In our work, we aim to address this gap by exploring ways to integrate a gaze-based recommendation model into the immersive sensemaking process, and evaluating its effect on the analyst's overall sensemaking process involving multiple interconnected documents.

3 System Design

In this section, we will detail the design of our system with general sensemaking features and proposed recommendation cues.

3.1 Sensemaking Features

Let us first detail the basic sensemaking features available in both Freestyle and EyeST, inspired by previous IST prototypes [49, 50]. All the available documents for the sensemaking task are given as a two-dimensional list in a three-dimensional space (see Figure 1b). Analysts could browse the documents through pagination buttons. On each page, analysts could see the headlines for up to ten documents. They could use the controller ray to aim at the headlines, and press the trigger button to open it in detail. The analyst could also use their controller ray to grab and move the documents around in the three-dimensional space. We allowed the analysts to create notes and labels to externalize their thoughts. We also allowed the analysts to search for keyword(s) in the dataset. All three features utilized a physical keyboard on a rolling cart (see Figure 1a). We also implemented a quick search [20] feature where the analyst can aim their controller ray at a word and press a predefined button to search, reducing the step of manually typing the word.

3.2 Gaze-Driven Interest Model

In a sensemaking task involving multiple interconnected documents, single gaze measures such as fixation duration or fixation count are not enough to infer the interest of a document perceived by the analyst [35]. Tahmid et al. proposed a gaze-derived metric, GazeScore, to address this issue by combining the duration and dwell values on each word and document [73]. GazeScore also addressed the frequency bias by incorporating the inverse document frequency (IDF) [16] of each word in a dataset. In our work, we built on their findings to develop a recommender based on the analyst's gaze data. Compared to the previous study, which only focused on capturing the analyst's overall (global) interest at the end of a sensemaking session, this study will explore two types of interest, local and global interest in documents, *during* sensemaking sessions. Equation 1 below represents the GazeScore (GS) for an entity x, which could be either a word or a document:

$$GS_x = \frac{\frac{GD_x - \mu_{GD}}{\sigma_{GD}} + \frac{UD_x - \mu_{UD}}{\sigma_{UD}}}{2} * IDF_x \tag{1}$$

Here, μ and σ denote the mean and standard deviation of Gaze Duration (GD) and Unique Dwell (UD), respectively. GD quantifies the time an analyst spends on an entity, while UD measures the frequency of an analyst revisiting an entity. Following are the ways we used this equation to derive local and global document interest.

Local interest of a word is defined as the significance of a word to an analyst relative to other words within the same document. In a sensemaking task, a single word can appear across multiple documents, each time in a different context. By considering local interest, we can gain insights into the contextual significance of a word to an analyst. For calculating the local interest of a word, we used the following parameters in Equation 1:

$$GD_w = \textit{reading duration of a word within a document}$$
$$UD_w = \textit{reading frequency of a word within a document}$$
$$IDF_w = 1$$

This gives us the local interest vector (LIV) for each document, D, with Equation 2.

$$LIV_D = \{GS_{w_1}, GS_{w_2}, GS_{w_3}, \cdots, GS_{w_M}\} \tag{2}$$

where M is the number of words in the vocabulary. EyeST computes the similarity between two documents by taking the cosine similarity between their LIVs.

Global interest refers to how relevant a word or a document is to an analyst relative to the entire dataset. Global interest allows us to get an understanding of the analyst's mental model at any time during a sensemaking session, and hints at their path of approach to the solution of the task. To calculate the global interest of a word, we used the following parameters in Equation 1:

$$GD_w = \textit{reading duration of a word within the entire dataset}$$
$$UD_w = \textit{reading frequency of a word within the entire dataset}$$
$$IDF_w = \log \frac{N}{n_w}$$

Here N is the total number of documents in the dataset, and n_w is the number of documents that contain the word w. This allows us to derive the Global Interest Vector (GIV) for the analyst following Equation 3.

$$GIV = \{GS_{w_1}, GS_{w_2}, GS_{w_3}, \cdots, GS_{w_M}\} \tag{3}$$

where M is the number of words in the vocabulary. We can compute the global interest of a document by taking the cosine similarity of its LIV and the analyst's GIV.

The primary objective of this study was not to devise a new algorithm for a recommendation system, but rather to assess the effectiveness of gaze-driven visual cues in suggesting relevant documents. The recommendation model for the system was thus kept simple. Our approach drew inspiration from the TFIDF algorithm [67], which transforms each document into a vector of double values containing the TFIDF value. We substituted TFIDF with GazeScore (Eq 1), subsequently converting them into LIVs (for each document) and a GIV (for the analyst), both of which encompass every word in the vocabulary, ensuring the same size for all vectors. EyeST updated these values each time the analyst looked away from a document after reading it for at least two seconds.

3.3 Recommendation Visual Cues

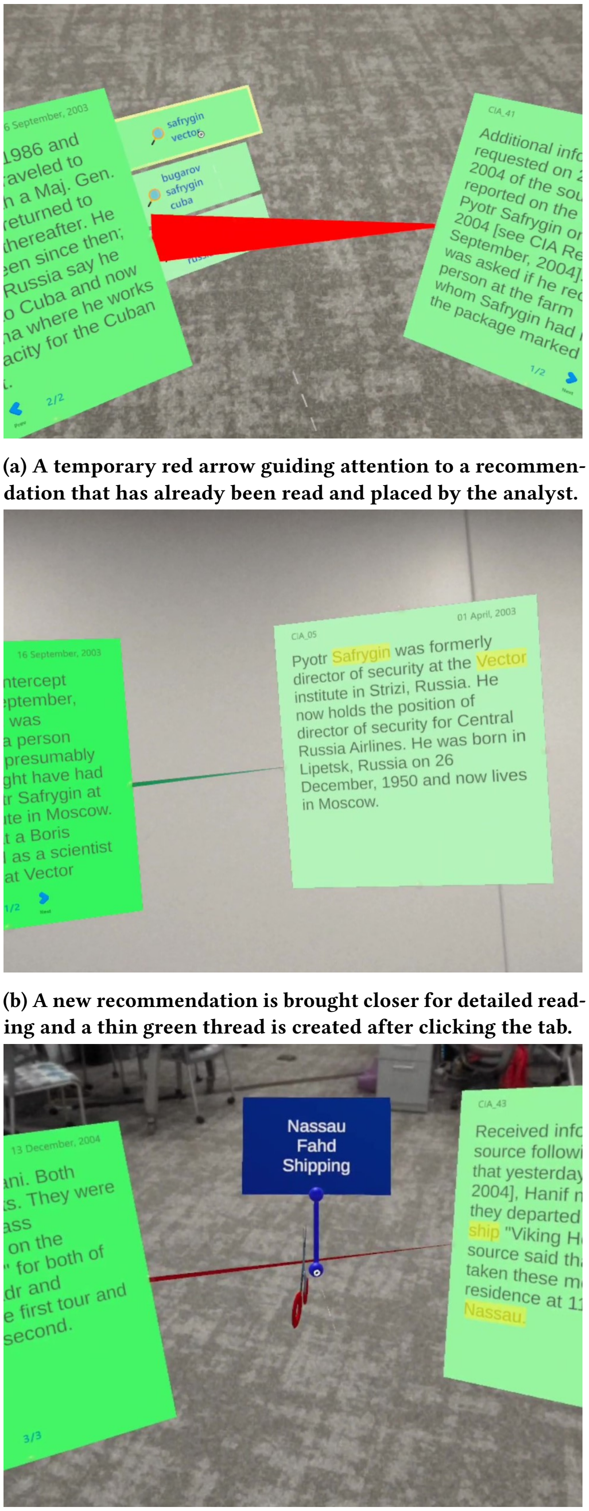
3.3.1 *Global Recommendation Cues.*

EyeST presents two cues to represent the documents' **global interest**. First, it changes the **document background color** to reflect its global interest, ranging from bright green for the most interesting to white for the least interesting to the analyst. Second, it **reorders the list** of unread documents so that the first document is always the one with the highest global interest to the analyst, and the following documents are presented in decreasing global interest order (see Figure 1b). Both of these cues helped in guiding the analyst's attention to the more interesting documents in an otherwise cluttered environment.

3.3.2 *Local Recommendation Cues.*

EyeST calculates the average of the similarity and global interest for each document pair. It then sorts these averages in descending order, and provides the top four as **local recommendations** to each document of the pair. This helps analysts expand their understanding of a document by linking it with others that are also of high interest overall. By factoring in global interest, we ensure that analysts stay focused and are not misled by information from any single document.

To avoid overwhelming the analyst with excessive information, we provide the recommendations with an overview first, leaving the analyst to decide if they want the details. The overview is shown as a list of tabs, each providing some context about the recommendation (see Figure 1c). The tab color signifies the recommended document's global interest; it also contains up to three words with the most interest within the context of these two documents. Finally, the tab has a yellow border if the recommendation has already been read



(a) A temporary red arrow guiding attention to a recommendation that has already been read and placed by the analyst.

Pyotr Safrygin was formerly director of security at the Vector institute in Strizi, Russia. He now holds the position of director of security for Central Russia Airlines. He was born in Lipetsk, Russia on 26 December, 1950 and now lives in Moscow.

(b) A new recommendation is brought closer for detailed reading and a thin green thread is created after clicking the tab.

(c) The analyst can review the shared words by hovering over a link. They can also sever a connection by clicking on the thread.

Figure 2: Interactions with the local recommendation tabs

by the analyst. The analyst can hover their controller ray over the tab to see the document's spatial position with a temporary red arrow (see Figure 2a). They can click the tab to make it disappear and create a more permanent thread attached to the recommended document (see Figure 1a). A tab without the border represents a new document for the analyst to read. The analyst can click the tab to bring the recommended document into view, and read it in detail. This will also result in a thin thread connecting the two documents (see Figure 2b), and the shared words being highlighted in the recommended document. Once a thread is created, the analyst can review the shared words by hovering over the thread. They can also click it to sever the connection (see Figure 2c).

4 Experiment Design

We present the details of the experiment we ran to evaluate EyeST's effectiveness in enhancing immersive sensemaking.

4.1 Research Questions and Hypotheses

4.1.1 RQ1: How do the gaze-driven recommendation cues affect the analyst's task performance? We propose using the analyst's implicit gaze data to infer their perceived interest in documents, and leveraging it to generate intelligent recommendation cues. The goal of these cues is to help analysts navigate the complex interconnected documents in a sensemaking task in a more efficient and effective way. We hypothesize that, in the EyeST condition,

- **H1a:** Analysts will be more efficient, by spending more of their time analyzing essential information.
- **H1b:** Analysts will not be derailed by spending more time on reading distractor information in the dataset.
- **H1c:** Analysts will read more of the essential documents.
- **H1d:** Overall sensemaking score will not be significantly better, since the personalized recommendations are only as good as the analysts they are personalized for.

4.1.2 RQ2: How do the gaze-driven recommendation cues affect the analyst's foraging strategies? The recommendation cues aim to enhance the analyst's foraging capability by having them retrieve more essential information with less effort and time. We hypothesize that, in the EyeST condition,

- **H2a:** With the help of recommendation cues, analysts will read more documents in total.
- **H2b:** Since analysts will be exposed to more essential information worth reading, they will spend more time reading essential documents.

4.1.3 RQ3: How do the gaze-driven recommendation cues affect the analyst's physical and mental effort? In a cognition-heavy sensemaking task, the analyst is already overwhelmed with a lot of unstructured information that requires a considerable amount of mental effort. We want to see if the recommendation cues can help alleviate some of that effort by partially reflecting their mental model onto the spatial layout. We hypothesize that, in the EyeST condition,

- **H3a:** Analysts will report less mental effort.
- **H3b:** Analysts will report less physical effort.

4.1.4 RQ4: How do the gaze-driven recommendation cues affect the analyst's overall experience? With the recommendation cues, we are offering the analyst an AI assistant that constantly learns from the analyst and updates itself. We want to investigate how the analysts respond to such support in a cognition-heavy task. Here, the focus is not just on how well they perform the task, but rather, how they *feel* about their own performance, their strategies, and their experience of collaborating with an AI assistant. We will analyze these findings and discuss how they relate to the implementation of our recommendation cues. Based on these findings, we will present design guidelines for future researchers working with implicit gaze-based recommendation cues in IA tools. This question focuses on an exploratory analysis of the analyst's experience with the EyeST, and requires no hypothesis.

4.2 Conditions

This study aimed to explore how a gaze-enhanced IA tool, such as EyeST, affects the analyst's sensemaking ability, their information foraging strategies, and their overall experience. This results in two distinct conditions, which we varied between subjects:

In the **EyeST** condition, the IA tool kept learning about the participant's interest from their gaze data, and helped them with additional recommendation cues. However, we did not share how the recommendation cues were generated until the end of the session to prevent participants from intentionally influencing the recommendations. This also allowed us to keep consistency between the two conditions. In the **Freestyle** condition, the participant received no additional cues from the system. They could explore the dataset either by exhaustively reviewing each document or by using keyword searches to find essential information.

4.3 Dataset and Task

To evaluate the efficacy of the recommendation cues, we utilized analysis exercises developed by Frank Hughes of the Joint Military Intelligence College [31]. These exercises include a collection of 111 fabricated intelligence reports embedded within a master plot. Out of these, 65 documents are directly related to the master plot (we refer to these as 'essential' documents), while the rest serve as noise or deceptions ('distractors'). Our intention was to use a sufficiently large dataset to ensure that participants could not simply read all the information exhaustively to find the answers to the guiding questions. Instead, they would need to work efficiently to identify the most important information within the time provided.

Participants were tasked with exploring these documents to investigate the scenario. The documents consisted of reports from intelligence agencies, all presented in text format and identified by randomized numbers. To provide more context to the participants, we generated a headline for each document using the following prompt in ChatGPT-3.5 [2]: *"Generate a headline for the following text in no more than 50 characters."* Since participants would only interact with the dataset for a limited time, we gave them four documents to start with, all of which contained key insights about the plot. These starter documents also allowed EyeST to collect the participant's gaze data, and start developing an understanding of their interests. We also created a prompt containing specific questions about the master plot to give them a concrete set of objectives:

(1) When is the attack going to happen?
(2) Who hired the attackers?
(3) What weapon are they going to use?
(4) Find the leaders, suppliers, and deployers from the given list.

The first three questions were open-ended, while the last one required participants to classify 22 names into three specified roles. The task is appropriate for both students and professionals, as it does not require any specialized analytical skills or domain knowledge. We gave the participants 45 minutes to complete the task.

4.4 Measures

We recorded each participant's actions in a detailed log file, storing document positions, reading duration, unique dwell counts, note-taking, label creation, and keyword searches, among other interactions. Additionally, we logged the global interest updates and the local recommendations generated for all documents throughout the session. Recommendation cues were not displayed to participants in the Freestyle condition. Furthermore, we collected gaze origin and direction at a frequency of 60Hz.

After the study, participants completed a demographic questionnaire followed by the NASA TLX [29]. They then filled out a modified short-form User Engagement Scale [59] to provide feedback on their focused attention, perceived usability, and reward factor. To understand participants' perception of the AI, we collected responses on the system's trustworthiness (adapted from TIA [43]) and the system's performance (adapted from TOAST [82]). We also included custom questionnaires to gather their opinions on finding relevant information and their sensemaking quality.

4.5 Procedure

The procedure of the study was as follows. Participants began by signing a consent form approved by the IRB. Next, the experimenter introduced them to the 'case file', explaining the task and outlining the features of the IA tool they were assigned to use. Afterward, the experimenter measured the participant's Inter-Pupillary Distance (IPD), and adjusted the headset accordingly. Then the experimenter helped the participant put on the headset, and guided them through the eye-calibration steps [3] while wearing the headset. The application was launched after completing the calibration.

4.5.1 Readability Trials. Meta's eye calibration steps test gaze precision on a virtual sphere that is relatively larger than the text height used in our study. Hence, to reassess the participants' gaze precision in our study, we developed a series of 10 readability trials. In each trial, the participant first found the answer to a question from a paragraph of 2-3 sentences. Next, the participant focused on a predefined word and confirmed with a controller button. Throughout these trials, participants maintained a distance of $2'$ from the text. The entire process took approximately 7-10 minutes to complete.

4.5.2 Tutorial. The participants learned how to interact with the IA tool by completing a set of brief tasks, including creating notes, labeling, searching for keywords, and resetting searches. They also

[2] https://platform.openai.com/docs/models/gpt-3-5-turbo

[3] https://youtu.be/lP0OFFuzIEU?si=F84NPKWhe5c6ZVR4&t=120

learned to move and organize documents in a 3D spatial layout. This segment took approximately 5-10 minutes to complete.

4.5.3 Main Study. In the main study, the participants completed the task from Section 4.3. They had access to the four starter documents and the prompt questions. Before they read the starter documents, the participants could not search for keywords or access the recommendation cues. However, they could still externalize their thoughts with notes and labels. After reading the starter documents, they could access the other documents in the dataset and all the features became available (see Figure 3). The participants had 45 minutes in this session, with up to a five-minute grace period on request. The experimenter provided time updates at 15, 30, 40, and 45 minutes.

4.5.4 Post-Study Questionnaire. After completing the task, participants took off the headset. They completed the post-study questionnaires and sat for an audio-recorded semi-structured interview to discuss their experience with the IA tool. This step took 10-15 minutes to complete.

4.6 Apparatus

We used the Meta Quest Pro [4] with eye tracking enabled. The eyes were tracked with two in-device cameras at approximately 60 FPS. We applied smoothing and filtering to reduce noise and improve the overall accuracy of the gaze data. The participants walked freely around an obstacle-free space of $17'x14'$. We enabled passthrough so that the participants could see the real world. They could interact with the IA tool with Meta Quest Touch Pro 6-DOF controllers. We implemented the application in Unity v2021.3.12 with Mixed Reality Toolkit 2.

4.7 Participants

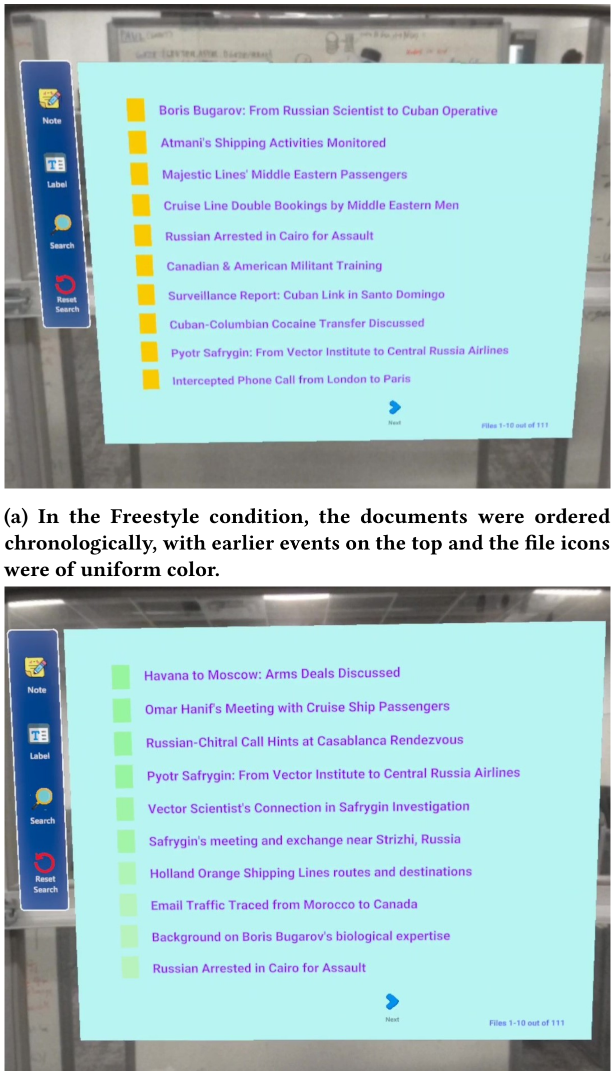
We ran a between-subject study with 26 participants (12F, 1 Non-Binary), with an average age of 23.28 (σ = 3.96, *max* = 31, *min* = 19). All participants had normal vision or corrected vision with contact lenses. All participants, except two, had used a mixed-reality headset at least once prior to this study.

5 Results

We analyzed the data from the experiment to test our hypotheses and understand the differences between the EyeST and Freestyle conditions. For all significance comparisons between the two conditions, we conducted the Shapiro-Wilk test to check for the normality of the data. For normal data distributions, we conducted the independent samples t-test, while running the Mann-Whitney U test for non-normal data.

5.1 Quality of Gaze Data

We started by analyzing the 'readability trials' (Section 4.5.1) to validate the gaze data collected from participants. First, we assumed that the participants would have more interest in the answers to the question. Out of 10 × 26 = 260 answer instances, the average interest for the answers was 81.1%, with a standard deviation of 12.3%, consistent with the findings from Tahmid et al. [73].



Boris Bugarov: From Russian Scientist to Cuban Operative
Atmani's Shipping Activities Monitored
Majestic Lines' Middle Eastern Passengers
Cruise Line Double Bookings by Middle Eastern Men
Russian Arrested in Cairo for Assault
Canadian & American Militant Training
Surveillance Report: Cuban Link in Santo Domingo
Cuban-Columbian Cocaine Transfer Discussed
Pyotr Safrygin: From Vector Institute to Central Russia Airlines
Intercepted Phone Call from London to Paris

Files 1-10 out of 111

(a) In the Freestyle condition, the documents were ordered chronologically, with earlier events on the top and the file icons were of uniform color.

Havana to Moscow: Arms Deals Discussed
Omar Hanif's Meeting with Cruise Ship Passengers
Russian-Chitral Call Hints at Casablanca Rendezvous
Pyotr Safrygin: From Vector Institute to Central Russia Airlines
Vector Scientist's Connection in Safrygin Investigation
Safrygin's meeting and exchange near Strizhi, Russia
Holland Orange Shipping Lines routes and destinations
Email Traffic Traced from Morocco to Canada
Background on Boris Bugarov's biological expertise
Russian Arrested in Cairo for Assault

Files 1-10 out of 111

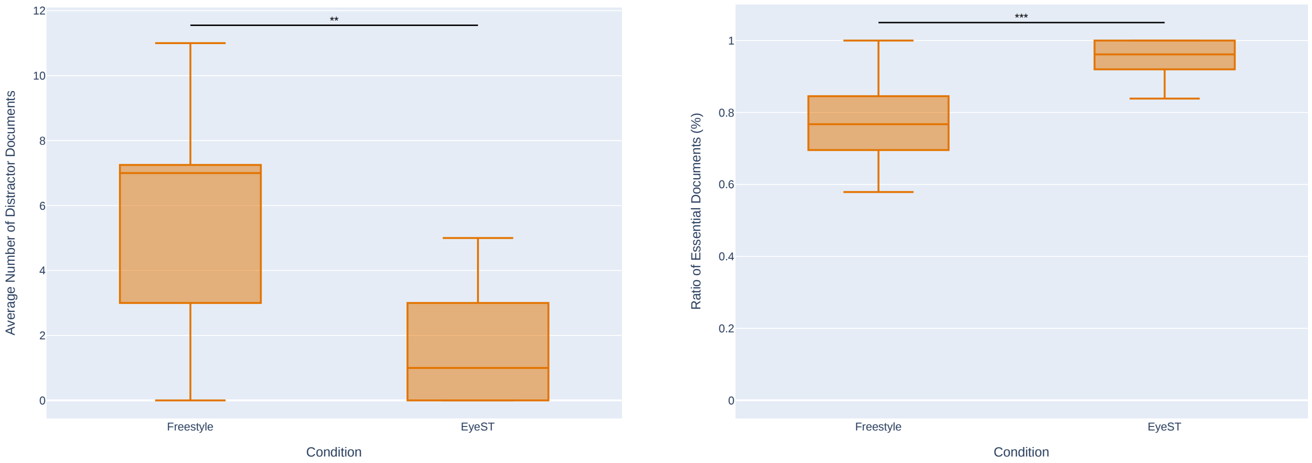
(b) In the EyeST condition, the documents were ordered based on their perceived global interest, which was also reflected in the icon color.

Figure 3: List of unread documents shown to participants in the Freestyle and EyeST conditions.

For the precision test, we measured whether the participant's gaze landed on a predefined word. We also measured the extent of deviation from the center of the word. Across 30×26 = 780 instances, the average precision was 75.7%, with a standard deviation of 18.4%. The low precision can be explained by the average gaze deviation from the target word, 0.72° ($\sigma = 0.26^\circ$), which is higher than the words' height (0.5°). However, the deviation still aligns with the 1.08° margin of error reported for Meta Quest Pro headsets [5].

The results from the readability trials emphasize that our gaze data were comparable with prior studies, allowing us to move forward with the data, and leveraging it for generating recommendation cues.

[4] https://www.meta.com/quest/quest-pro/



(a) Freestyle participants read more distractor documents than EyeST participants.

(b) EyeST participants read a higher percentage of essential documents than Freestyle participants.

Figure 4: Comparing the ground truth of documents read by participants in Freestyle and EyeST conditions.

5.2 Filtering Information

We evaluated how the participants from the two conditions filtered 'essential' information while avoiding 'distractors'.

5.2.1 Number of Essential Documents (rejects H1c, H2a). We compared the differences in the number of essential documents read by participants for the two conditions. We found no significant difference ($t(24) = -1.5, p = 0.15$) between the Freestyle ($\mu = 20.69$, $\sigma = 8.33$) and EyeST conditions ($\mu = 26.08$, $\sigma = 9.89$).

5.2.2 Number of Distractor Documents (supports H1b). We found that that participants in the Freestyle condition ($median = 7.0$, $IQR = 4.0$) read significantly more ($U = 145.5, p < 0.01, effect = 0.36$) distractor documents than the participants in the EyeST condition ($median = 1.0$, $IQR = 3.0$) (see Figure 4a).

5.2.3 Ratio of Essential Documents (supports H1a). Out of all documents read, the percentage of essential documents read by Freestyle participants ($\mu = 0.78, \sigma = 0.12$) was significantly less ($t(24) = -4.87, p < 0.001, effect = -1.91$) than same ratio for the EyeST participants ($\mu = 0.95$, $\sigma = 0.05$) (see Figure 4b).

5.2.4 Reading Duration of Essential Documents (supports H2b). We measured the time spent on reading essential documents by all participants. We found that Freestyle participants (μ=1446.73, σ=333.48) spent significantly less time reading essential documents ($t(24) = -3.10, p < 0.01, effect = -1.2$) compared to EyeST participants (μ=1810.12, σ=259.17) (refer to Figure 5a).

5.2.5 Reading Duration of Distractor Documents (supports H2b). We measured the reading duration of distractor documents by participants, revealing that participants in the Freestyle condition ($median = 222.01$, $IQR = 176.22$) spent significantly more time reading distractor documents ($U = 147.0, p < 0.01, effect = 0.37$) compared to participants in the EyeST condition ($median = 23.18$, $IQR = 82.57$).

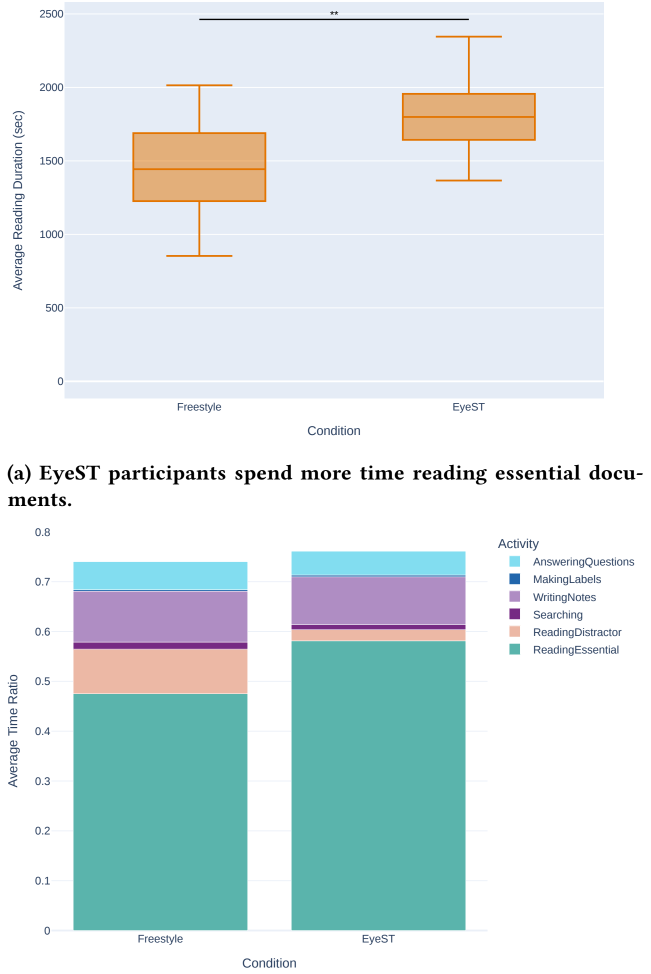
5.2.6 Time Distribution of Users (supports H1a). We compiled the duration of various activities during sensemaking and normalized them by the total time of the session. Figure 5b illustrates that EyeST participants spent roughly the same amount of time on essential documents as Freestyle participants did on both essential and distractor documents combined. Overall, EyeST participants spent slightly more time in the immersive space. We note that, while small, there were some overlaps between these actions. For instance, participants read documents while they were in the middle of writing notes. Also, we did not consider trivial actions such as moving documents, walking around, etc., hence the total time shown in the figure is less than 1.0.

5.2.7 Task Performance. To complete the task, participants had to answer a set of questions targeted towards the solution. We graded each of their responses according to a rubric, and gave one point for each correct answer. The highest possible points one could receive was 44. The results ($U = 83.0, p = 0.95$) revealed that there was no difference between Freestyle ($median = 9.0, IQR = 9.0$) and EyeST ($median = 10.0, IQR = 5.0$) performance, thus supporting H1d.

The results from this section allowed us to test the hypotheses for RQ1 and RQ2. We found that, while participants did not read more documents or perform better with EyeST, they used their time more efficiently by prioritizing essential documents over distractions. In the following sections, we will examine the role of local and global recommendation cues in supporting this efficiency and their influence on participants' interactions with the tool.

5.3 Evaluating Recommendation Cues

In order to explore the impact of local recommendation cues on locating essential documents, we counted how many of them were essential. We observed a precision of 90% with 43% recall, implying that EyeST was very selective about the recommendations (see Figure 6b). The confusion matrix cells represent the average percentage of documents across all EyeST participants. We observe that



(a) EyeST participants spend more time reading essential documents.

(b) Average time distribution of participants' activities across two conditions.

Figure 5: Comparison of time spent reading essential and distractor documents by participants in Freestyle and EyeST conditions.

EyeST only recommended documents if it was confident, allowing it to successfully prune the distractor ones. In contrast, Freestyle participants' strategy mostly involved searching keywords to find essential documents (see Figure 10a). Figure 6a demonstrates that their search results yielded higher recall (64%) at the cost of lower precision (67%) than the EyeST participants.

We also analyzed how many of the documents were essential when they were sorted by global interest, and found that EyeST maintained ≈ 100% precision for the top 15 documents throughout the session (see Figure 7). Towards the end of the session, the precision declined for documents ranked lower than 15. Interestingly, the decline started around the same time the participants were updated about passing two-thirds (30 minutes) of the session. One possible reason is that with the time update, participants began to rush through the documents. As a result, distractor documents started appearing more often, moving up the global interest hierarchy, and affecting the quality of global interest.

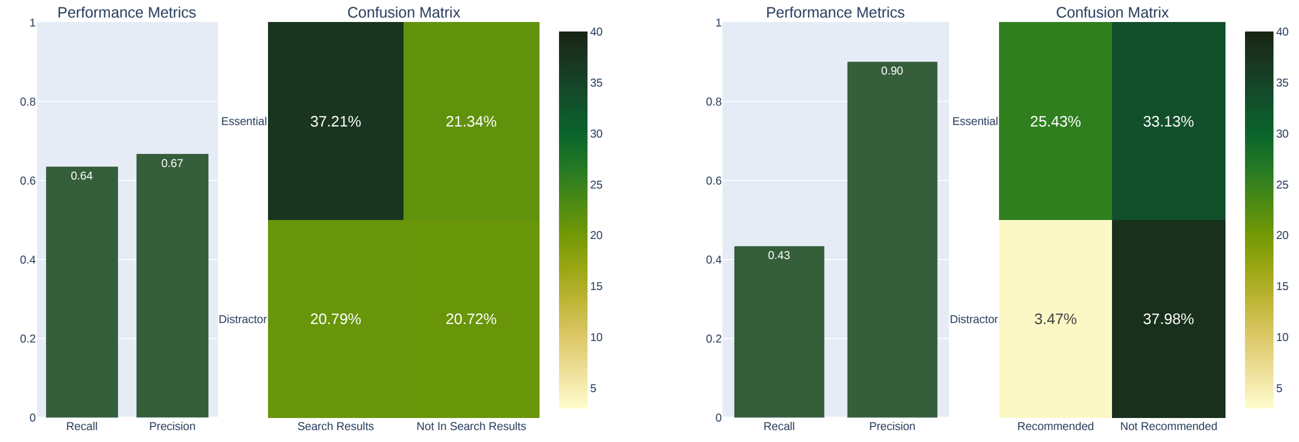
5.4 Dataset Exploration

Since we found that the recommendation cues were 'good' at guiding participants toward the ground truth, we wondered whether the cues were personalized for each participant. In other words, since everyone started with the same four essential documents, could we have simply recommended documents related to the starter documents, without considering the participant's gaze? We investigated this question by exploring how much of the dataset was explored by participants. Additionally, we wanted to see whether the recommendation cues guided all EyeST participants towards the same storyline, or if their gaze data personalized the experience enough to explore different storylines within the dataset.

5.4.1 Document Coverage. We analyzed all the documents read by participants in both conditions and counted how many participants read each one. Figure 8a shows the results in a heatmap where darker colors represent more readers and the documents in each group are sorted by their similarity to the starter documents. As expected, the starter documents are the darkest since all participants had to read them.

Interestingly, a few other documents are nearly as dark, suggesting they were popular with EyeST participants. These darker documents are mostly on the left side of the heatmap, indicating their content similarity with the starter documents. At first glance, it might seem like these documents were recommended only because of their similarity to the starter documents, regardless of their gaze-inferred interest. On closer inspection, we found that these darker documents (read by 10 or more EyeST participants) all focused on one suspect, *Safrygin*, who was mentioned once in four starter documents. It is worth noting that the starter documents included information about ten suspects, eight potential targets, and three suspicious organizations in total. Yet, most EyeST participants were interested in this one suspect, who was mentioned to be in contact with a 'bioweapon' expert. One of the prompt questions was directly about weapons which might explain why the participants showed more interest in this plot thread than the others. The other starter documents were about suspicious travel plans and a suspicious package being moved around the country. A few participants explored these threads of the plot, explaining the lighter shades near the starter group. From observation of the top-right corner, even though the EyeST participants read some distractors, they did not spend much time on those (see Figure 8b), reaffirming their tendency to avoid distractors.

In contrast, Freestyle participants, lacking recommendation cues, explored freely, and the colors are hardly distinguished between essential and distractor documents. There is one dark shade in the Freestyle row (around the middle), which was also the first document on the default unread document list. Hence, almost all Freestyle participants ended up opening and reading that document. In summary, while Freestyle participants explored documents with more diversity, EyeST participants were steered toward documents with high similarity to the one suspect from the starters, leading to a more focused exploration. We will discuss the implications of this in Section 6.



(a) Search results performance in the Freestyle condition.

(b) Local recommendation performance in the EyeST condition.

Figure 6: Comparing the performance of EyeST local recommendations with the Freestyle search results at finding essential documents while avoiding distractors.

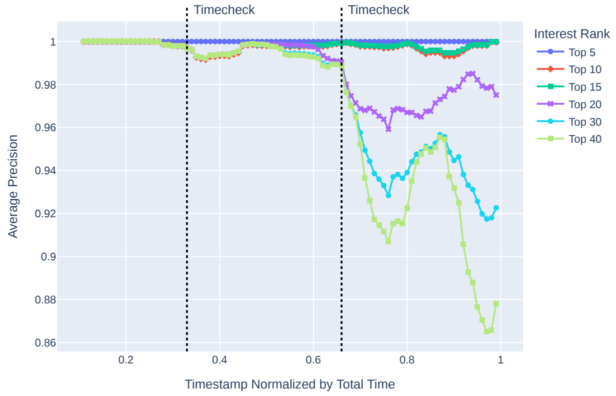


Figure 7: EyeST maintained a high precision for the top 15 documents (based on global interest) throughout the session. As participants started rushing through information in the latter part of the session, documents with lower global interest tended to start including distractors, lowering precision, but still staying mostly over 90%**. Interestingly, the decline aligns with the time the experimenter gave the time update at 30 minutes.**

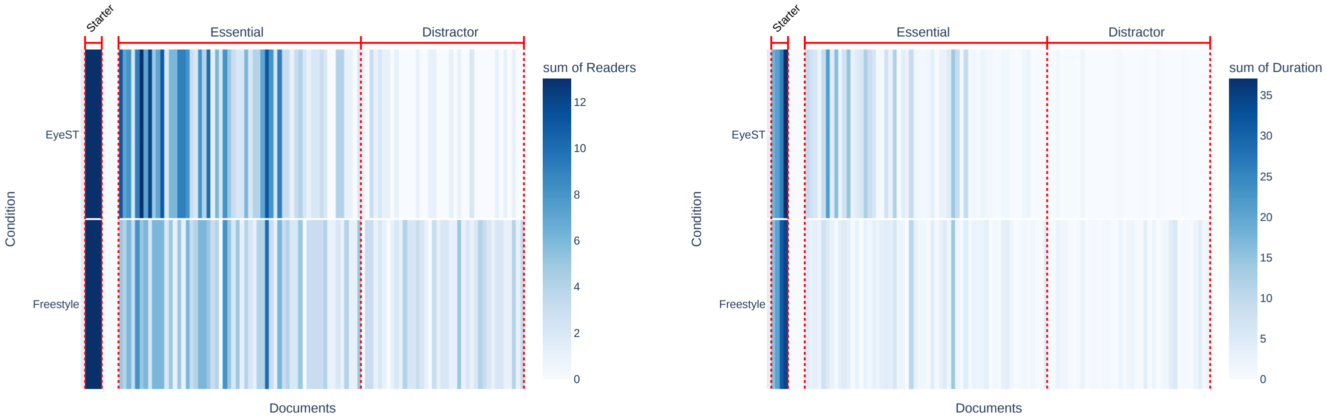
5.4.2 Effect of Gaze on Documents. To further investigate the personalization of gaze-derived interest, we studied how different EyeST participants perceived different documents, and how their gaze played a role before and after reading each document. Except for the starter documents, all EyeST participants read one other document, CIA_34, during their sensemaking session. We compared this document's global interest with a starter to see how the patterns differed. As Figure 9 illustrates, the starter document's interest gradually increased in small steps throughout the session for all participants, showcasing the participants' dependence on the starter documents' content. This argument aligns with the observation that EyeST participants read more essential documents that are similar to the starter documents. In contrast, for the non-starter, the global interest rose up and down at different stages of the session, hinting at the participants' evolving mental model.

5.5 Source of New Documents

To explore the dataset, the participants could browse the files in the list of unread documents (see Figure 3), and open documents to read. They could also search for keywords and find documents that way. The EyeST participants also had the option to read new documents by following the local recommendation tabs (see Figure 1c).

5.5.1 Freestyle. We conducted a Wilcoxon signed-rank test to compare the number of documents selected from the list ($median = 8$, $IQR = 3$) and through a search ($median = 20$, $IQR = 18$) in the Freestyle condition. The test revealed that participants relied on the search significantly more (($W = 12.5, p < 0.05, effect = 3.47$)) than the list to get new documents (see Figure 10a)

5.5.2 EyeST. We conducted a Friedman test to assess differences across the three sources of new documents (List, Search, and Recommendations) in the EyeST condition. The results showed a significant difference between sources ($\chi^2(2) = 9.69, p < 0.01$). Post-hoc Wilcoxon signed-rank tests with a Bonferroni correction revealed significant differences between List ($median = 14$, $IQR = 14$) and Recommendations ($median = 5$, $IQR = 5$) with $p < 0.001$



(a) Heatmap reflecting the total number of participants that read each document.

(b) Heatmap reflecting aggregated mean reading duration for each document.

Figure 8: There are a few darker lines on the essential documents for EyeST participants, mostly on the left side, meaning that they are very similar to the starter documents. In contrast, the Freestyle row contains more uniform shades. In both heatmaps, the top-right corner is relatively lighter, reflecting EyeST participants' tendency to avoid distractors.

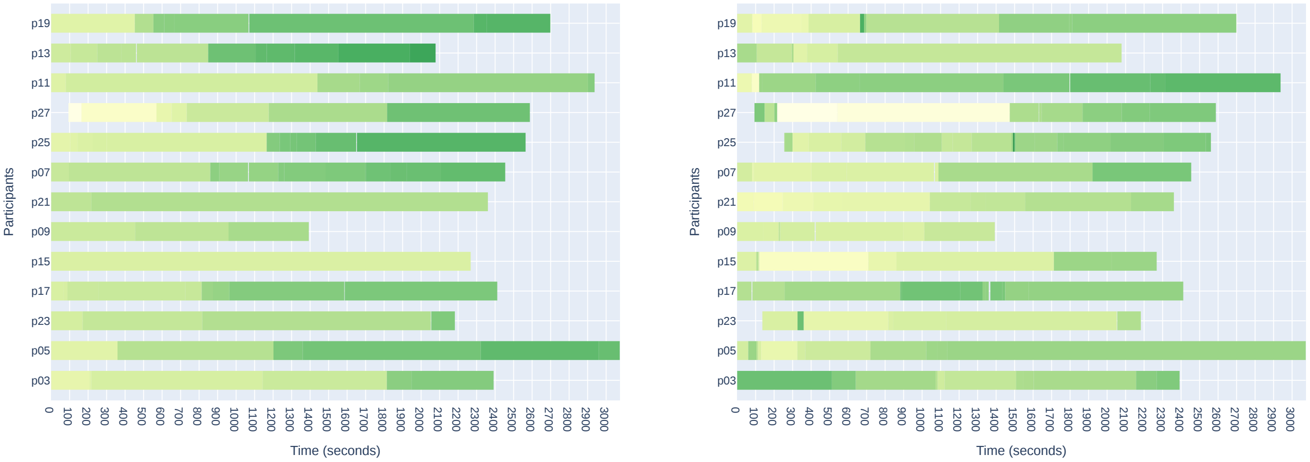


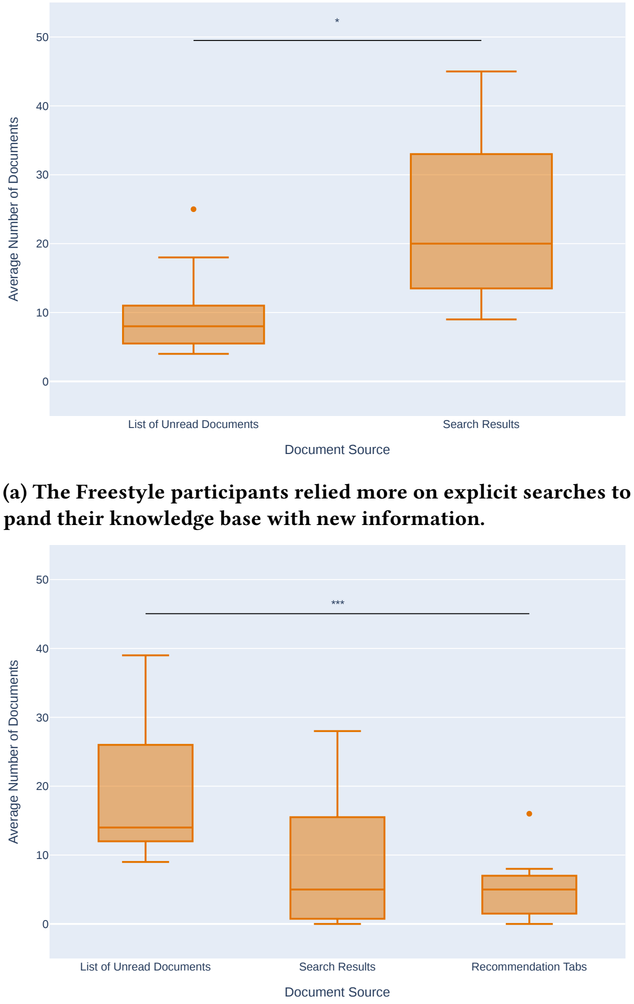
Figure 9: Global interest update timeline for a starter (left) and a non-starter (right) read by all participants. Participants are sorted by their performance. The global interest steadily increases for the starter document, implying the participants' reliance on the starter documents. In contrast, the global interest for the non-starter rises up and down based on participants' mental models throughout the session.

(see Figure 10b). No significant difference was found between List and Search ($p = 0.17$), or between Search and Recommendation ($p = 1.0$).

5.5.3 *Freestyle vs. EyeST.*

We conducted Mann-Whitney U tests to compare the document sources across the two conditions. The results ($U = 26, p < 0.005, effect = -0.35$) revealed that for new documents, the List was used more by the EyeST participants ($median = 14$, $IQR = 14$) than the Freestyle participants ($median = 8$, $IQR = 3$). On the other hand, the Search was used significantly more ($U = 140, p < 0.005, effect = 0.33$) by the Freestyle participants ($median = 20$, $IQR = 18$) than the EyeST participants ($median = 5$, $IQR = 14$).

To better understand how document sources influenced a sensemaking session, let us examine an example from each condition. Figure 11a presents a session from a Freestyle participant. The hourglass (⧖) at the top indicates each search performed by the user. It is evident that most of the documents were obtained from search



(a) The Freestyle participants relied more on explicit searches to expand their knowledge base with new information.

(b) The EyeST participants relied more on the list of unread documents to expand their knowledge base with new documents.

Figure 10: Comparing sources for new documents in the two conditions.

results. Figure 11a also shows the ground truth relevance of the documents on the color-coded y-axis labels. We observed that the participant ended up obtaining both essential (blue) and distractor (red) documents from these search results.

In contrast, Figure 11b illustrates the sensemaking timeline of an EyeST participant employing a similar strategy. This participant also engaged in frequent search operations but did not rely solely on search results. Instead, they also retrieved documents from the list and recommendations. This suggests that the user was utilizing recommendation cues to get new documents, while search functionality aided in reviewing and integrating their own ideas with system-suggested ones. Figure 11b also shows how the additional information helped them acquire essential documents (blue) almost exclusively throughout the session.

5.6 Qualitative Questionnaires

We did not find any significant differences for NASA TLX (rejecting H3a, H3b). We present the results from the questionnaires about user engagement and trust in the system, followed by two custom question sets on finding essential information and overall sensemaking.

5.6.1 User Engagement and Trust. We converted participants' responses into a 1-7 scale, where 1 represented "strongly disagree" and 7 represented "strongly agree." An independent samples t-test revealed significant differences in Focused Attention ($t(24) = 2.26, p < 0.05, effect = 0.89$) and System Performance ($t(24) = 2.25, p < 0.05, effect = 0.88$), as shown in Figure 12. We derived the Focused Attention score from the following three statements:

(1) I lost myself in the experience
(2) The time I spent in sensemaking just slipped away
(3) I was absorbed in my sensemaking task

The results indicate that Freestyle participants ($\mu = 5.51$, $\sigma = 1.0$) felt more focused on their sensemaking tasks compared to EyeST participants ($\mu = 4.33$, $\sigma = 1.6$). We calculated the System Performance based on the following four statements:

(1) The tool helped me achieve my goals
(2) The tool performed consistently
(3) I was rarely surprised by how the tool responded
(4) I felt comfortable relying on the information provided by the tool

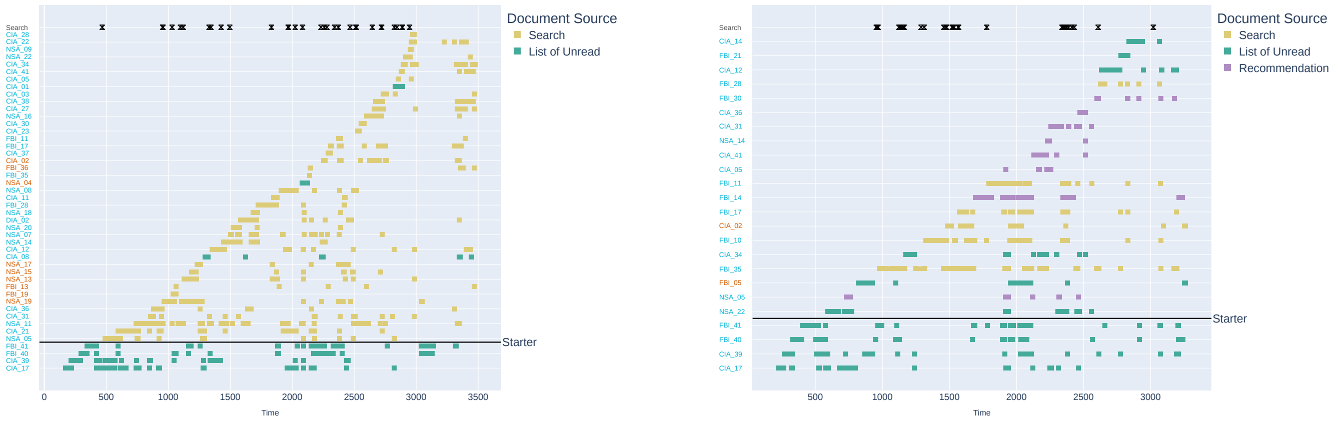
The results show that Freestyle participants ($\mu = 5.52$, $\sigma = 0.75$) rated the system's performance higher than EyeST participants ($\mu = 4.65$, $\sigma = 1.17$), suggesting greater confidence in the tool's outputs when they did not have any recommendation cues.

5.6.2 Finding Essential Information and Sensemaking Score. We asked the participants a set of five custom questions to understand their thoughts on finding relevant information, and a set of four custom questions to assess their sensemaking experience. We did not find any significant difference between the two conditions for finding relevant information (($t(24) = 0.2, p = 0.84$)), or for sensemaking ($t(24) = 1.64, p = 0.11$). However, looking at each statement individually, we noticed a significant difference for one statement, *'The tool helped me in exploring the dataset'* ($U = 131.5, p < 0.05, effect = 0.28$). Freestyle participants felt the tool helped them explore the dataset more than the EyeST participants.

5.7 Qualitative Analysis of the Interview

We analyzed the interview transcriptions, coding them into key themes and organizing them into distinct categories. By cross-referencing participant quotes with session logs and screen recordings, we gathered a comprehensive understanding of their experiences. We present the key findings from this analysis in the following subsections.

5.7.1 Pros and Cons of the Freestyle Condition. Almost all Freestyle participants hailed the search feature, particularly the quick search, as one of the most useful tools in exploring the dataset. As the reason for preferring the quick search over the typed search, the participants mentioned the challenges of switching context between



(a) Example of a Freestyle participant's (P12) timeline. Freestyle participants rely mostly on search results to get new documents and tend to read distractors (red labels on the y-axis) in addition to essential documents (blue labels on the y-axis). The black hourglasses (⧖) on the top represent each search operation.

(b) Example of an EyeST participant's (P19) timeline. EyeST participants rely on the list, local recommendations, and search results to get new documents. However, they still stay on track by reading only essential documents (blue labels on the y-axis) while avoiding distractors (red labels on the y-axis).

Figure 11: Timelines from two conditions showing the impact of new document sources in retrieving essential documents. Participants from both conditions tend to regularly go back to the starter documents to reaffirm their understanding of the dataset.

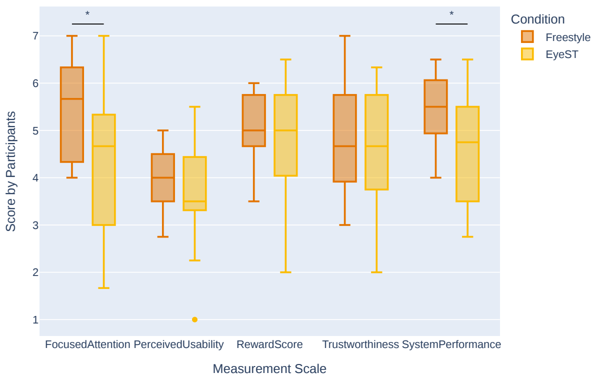


Figure 12: The Freestyle participants reported more focused attention and better system performance than EyeST participants.

controller and keyboard and losing their train of thought. Freestyle participants also expressed feelings of being overwhelmed with the sheer volume of information, and getting lost after a while when they ran out of keywords to search. P26 said, *"I had a strategy, but didn't know how to execute. Because I didn't know the documents I read were relevant or not."*. P20 reiterated, *"··· lots of information. Didn't know how much of that is relevant."*

For improvements, Freestyle participants suggested the addition of text-highlighting capabilities within documents as a way to narrow down their focus. Several participants requested the option to close or minimize documents to reduce visual clutter, while others wanted capabilities to draw connections between documents.

5.7.2 Pros and Cons of the EyeST Condition. Participants in the EyeST condition generally found the color-coding of document interest helpful, as the color gradient made it easier to identify which documents to prioritize while reviewing older information. Several participants found the thread between related documents as a valuable tool for synthesizing inter-connected information across documents. Many participants also found the quick search feature useful, especially when reviewing information from previously read documents. When search results were overwhelming, participants relied on the document's background color to narrow their focus.

However, the local recommendations were not as well-received. Participants found them less helpful compared to the color-coding or the reordered document list. P21 remarked, *"The keywords were vague and disjointed"*, while P13, who relied on both search and recommendations to expand their knowledge noted, *"The AI kept recommending documents about the same person, so I stopped using it and searched for other suspects."* Interestingly, the AI-suggested person was a key figure in the ground truth solution. Contrary to the Freestyle, EyeST participants did not feel lost and did not run out of ideas during the session. However, they felt their space was filled pretty quickly, some participants pinning the reason on the recommendation tabs that *"kept popping up even when I did not want it. (P25)"*

When asked how the AI could improve their experience, participants expressed a desire for more control over the recommendation cues, including the ability to guide the recommendations by integrating their own insights so they would not feel confined to the AI's suggestions. They also wanted the capability to manually highlight phrases in documents and create labeled threads to enhance their ability to organize and relate information.

6 Discussion

In this section, we will discuss the implications of our results, and provide a design guideline for gaze-based recommendation cues.

6.1 Validity of the Recommendations

There are two key ways to validate the effectiveness of the recommendation cues. First, it is essential to assess whether these recommendations are close to the ground truth, ensuring they can guide the participant toward the correct solution. We found that an average of 90% of the recommendations was essential to the ground truth. Considering the global interest, almost 100% of the documents in the top 15 were always essential. This, combined with the fact that EyeST participants preferred getting new documents from the list, where the top 10 documents with the highest global interest were always on the first page, explains why EyeST participants spent more time on essential documents. However, EyeST did not aid in making implicit connections between the documents which is a non-trivial task even for state-of-the-art generative AI models [85]. As a result, EyeST participants found the task highly complex (as demonstrated by NASA TLX results) and faced similar struggles to Freestyle participants in answering questions, leading to comparable task performance.

Second, it is important that the automated recommendation cues do not dominate the analyst's decision-making by always suggesting the same set of essential information to everyone. The analyst should remain in control of the sensemaking session, with recommendations tailored to their individual needs. To validate the individuality of EyeST participants, we investigated how a document's perceived interest evolved for each participant. We found that the global interest in the starter documents kept increasing throughout the session. This can be explained by the experiment design. In the beginning, participants were told that the starter documents were important to the plot. Since they began with these and knew they were important, the starter documents became their reference point for gathering evidence for sensemaking. Another piece of evidence for personalized gaze data was the fluctuating global interest in non-starter documents. Since their relevance to the plot was not concrete, different participants perceived the non-starter documents differently, allowing each of them to explore topics of their own interest.

6.2 Effect of Recommendation Cues on Sensemaking

We established that the recommendations were personalized for each participant, while being closely aligned with the ground truth. The next question (RQ4) is how they affected participants' foraging and synthesis abilities.

In terms of foraging, EyeST participants were highly efficient, with 95% of the documents they read being essential, compared to 77% for Freestyle users. Although EyeST participants spent more time reading essential documents, they did not read a greater number of them overall. This suggests that the recommendation cues encouraged a more depth-first approach rather than a breadth-first approach. Additionally, EyeST participants were not derailed by red herrings, unlike Freestyle participants who spent significantly more time reading distractor documents.

When it came to synthesis, both groups struggled with information overload and largely failed to complete the task. This led to a ceiling effect on the NASA TLX, showing no difference between the two conditions. However, since EyeST participants were not distracted by distractors, they stayed on track throughout the session. In contrast, Freestyle users often felt lost after exhausting their strategies, which mainly involved searching by keywords. Many were unsure of their next steps midway through the session.

Curiously, EyeST participants reported lower levels of focused attention compared to Freestyle participants. This hints at a possible downside to human-AI collaboration. While the AI's recommendation cues helped guide participants to the right information, they paid less attention to the content itself, losing some immersion in the process. In other words, since the AI was handling much of the task, participants felt less need to put in extra effort. In addition, participants often found the local recommendations unclear and unexplainable, leading to lack of trust on the EyeST performance.

Interestingly, this phenomenon is not new. In human-AI collaboration, Lee and See recommended caution when providing additional or potentially conflicting information, as it can lead to over-reliance on the AI [46], potentially causing reduced skill even in highly capable participants [57, 61]. In such collaborative scenarios, the relationship between performance and the amount of information typically follows an inverted U-curve [25, 30]. Hence, what the AI shares and how it shares, can both be crucial for establishing an effective teaming paradigm. We can address this issue by clearly defining the roles of the analyst and the machine within an intelligent visual analytics tool, coupled with appropriate feedback and a balanced amount of information [80]. In the following sections, we take a closer look at how the participants interacted with the recommendation cues and discuss ways to improve them for more effective human-AI collaboration.

6.3 Evaluating the Recommendation Cues

Overall, participants reported the global recommendation cues to be more helpful, especially the color coding. The color gradient on the documents clearly highlighted their perceived interest and helped participants narrow their focus, especially when they were overwhelmed with too much information. They relied on color-coding even when browsing through search results, whether revisiting older documents they had already read or exploring new ones. The other global cue was the reordering of the unread document list by their perceived interest. The participants relied on this sorted list most while looking for new information. These two cues, both representing global interest, effectively supported participants in filtering essential information from a complex web of interconnected documents. Both of these global cues involved a higher level of automation, where the AI already made a decision from the participant's implicit gaze data, and was offering suggestions without any extra explicit input from the human. Hence, the human does not grapple with the dilemma of whether or not to trust the AI.

On the contrary, participants had trouble interacting with the local recommendation cues, and had mixed feelings about their usefulness. To reduce visual clutter, local recommendations were presented with overview tabs that displayed global interest (using color), local interest (the three most relevant words), and read status (indicated by an external border for read documents). Participants could decide whether to explore the recommendation further based on this overview. The tab would then disappear, connecting the recommendation with a thin thread. While participants appreciated using the thread to identify connections between documents, they struggled with interpreting the color and the common words in the overview. The color was not very helpful since it did not vary much between the four recommendations. It was also difficult to comprehend the content of the local recommendations from just three disjointed words.

In short, with lower levels of automation, local recommendations required participants to make decisions based on unclear and unexplainable information, often while they were already overwhelmed by the documents themselves. On the other hand, lacking recommendation cues, Freestyle participants used the search feature almost exclusively, which relied on a simple term-matching algorithm. So their foraging steps were straightforward, trustworthy, and predictable. This resulted in an interesting response from the participants. EyeST participants felt they did not explore the dataset very well. This highlights the need to improve the visual cues for local recommendations.

6.4 Design Guidelines: Improving Recommendation Cues for Eye-Enhanced Immersive Analytics

We developed the recommendation cues for EyeST based on design guidelines for human-centered AI interactions [3, 15, 76]. For instance, the emphasis on displaying contextually relevant information guided us in designing the local recommendation cues. However, participants found it challenging to interpret these cues, suggesting that existing guidelines need refinement for domain-specific tasks such as intelligence analysis and sensemaking. Drawing from our study's results and user feedback, we propose new guidelines to help future researchers optimize gaze-based recommendation cues to enhance immersive sensemaking tools.

6.4.1 Adding Syntactic Context. The EyeST participants did not find the local recommendations useful even though they were guided towards the ground truth solution. To enhance participants' understanding of these recommendations and build trust in the AI, we propose adding more contextual information.

Instead of providing N disjointed words, we suggest offering a fully formed, syntactically correct sentence that conveys additional details about the recommended document. This can be accomplished by using a curated prompt for a large language model (LLM), such as: *"Generate a summary of document A in less than 50 characters, including the words: wordA, wordB, wordC."* This approach will provide more comprehensive information about the recommended document while emphasizing the participant's personalized interests.

6.4.2 Diversified Recommendations. One observation about the recommendations was them being closely related to the starter documents, causing EyeST to often keep recommending the documents that were already read by participants. This limited the participants' possibility of expanding their knowledge base, although they still appreciated the ability to create connections among previously read documents.

To encourage the use of local recommendations for foraging new documents in addition to connecting old ones, we suggest separating the recommendations into two categories: old (for synthesis) and new (for foraging). Also, instead of being always-on, we suggest allowing the participant to decide when they want a new document, and when they want to synthesize relationships with older documents. This could be achieved by a button attached to the document. This approach will ensure that for each document, participants will always have the option to move their sensemaking process forward with a mix of foraging new information and synthesizing older information.

6.4.3 User Adaptability. In this study, we only focused on the participants' *implicit* gaze data to generate personalized recommendations. However, there were instances where participants did not find certain recommendations useful, and wished to explore the dataset from a different perspective. To address this, we propose adding an additional layer of human feedback to the recommendation system, allowing participants to approve or decline specific recommendations. This explicit feedback would then update the parameters of the recommendation model, similar to the Star-SPIRE system [9]. By implementing this approach, the recommendation model would learn from both implicit and explicit participant feedback, thereby refining the recommendations over time.

6.4.4 Increasing Transparency. To maintain consistency between the two conditions and prevent participants from explicitly manipulating the recommendations, we chose not to disclose how the recommendation cues were generated until the end of the session. This approach may have led to confusion and distrust toward EyeST.

In applications built for real-world scenarios where such restrictions are unnecessary, we suggest providing detailed information about the recommendations and their functionality. This transparency will help build trust in the system and foster a more effective collaboration between humans and AI. However, that would allow the humans to intentionally manipulate the recommendations which might interfere with the recommendation model parameters. Hence, the researchers should proceed with a balanced approach, ensuring that the intelligent models can address such scenarios.

6.4.5 Better Eye-Tracking Technology. Our gaze-driven recommendation system leverages the headset's ability to capture participants' gaze with word-level precision. The height of a word viewed from a comfortable distance ranges from 0.45 to 0.55°, while the margin of error for the state-of-the-art headset's eye tracking is 0.84°.

Therefore, we recommend that researchers implement noise reduction techniques, such as smoothing and filtering, to minimize fluctuations and achieve more stable and accurate gaze data.

7 Limitations and Future Work

We recognize that the local recommendation visualization cues require significant improvements to enhance their effectiveness. Future efforts will focus on refining these cues based on the proposed design guidelines. One limitation of the current system is that the gaze data used for generating recommendations was not entirely accurate, which may have affected the quality of the recommendations. To address this, we plan to integrate more advanced eye-tracking technology in future iterations. We also acknowledge the low number of participants in this study, and plan to address it in future studies.

Additionally, our concept of rich semantic interaction extends beyond just eye-tracking data. By leveraging rich sensor data, future iterations of intelligent IA tools can offer more nuanced interactions, better aligning with the user's cognitive and physical behaviors. We believe it would be valuable to explore how our design guidelines can evolve to incorporate these additional data streams, enabling future IA tools to interpret analyst intent more holistically.

8 Conclusion

In this paper, we examined the development and evaluation of gaze-driven recommendation cues aimed at enhancing sensemaking tasks within immersive analytic tools. Our findings indicate that the recommendation cues effectively assist analysts in navigating complex document sets by facilitating access to relevant information while minimizing exposure to irrelevant content. However, participants reported lower performance and expressed reduced attentiveness due to the lack of clarity for some of the recommendation cues. Feedback from participants enabled us to identify essential requirements for addressing these challenges. We synthesized these insights into a set of design guidelines for future researchers. The proposed guidelines emphasize the importance of adding contextual information, incorporating user feedback mechanisms, and improving transparency regarding recommendation cues. Ultimately, these guidelines aim to foster a more effective human-AI collaboration in immersive analytic environments.

Acknowledgments

This work was supported in part by NSF I/UCRC CNS-1822080 via the NSF Center for Space, High-performance, and Resilient Computing (SHREC).

References

[1] Seoyoung Ahn, Conor Kelton, Aruna Balasubramanian, and Greg Zelinsky. 2020. Towards Predicting Reading Comprehension From Gaze Behavior. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Short Papers)*. Association for Computing Machinery, New York, NY, USA, Article 32, 5 pages. https://doi.org/10.1145/3379156.3391335

[2] Salwa D Aljehane, Bonita Sharif, and Jonathan I Maletic. 2023. Studying Developer Eye Movements to Measure Cognitive Workload and Visual Effort for Expertise Assessment. *Proceedings of the ACM on Human-Computer Interaction* 7, ETRA (2023), 1–18.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[4] Christopher Andrews, Alex Endert, and Chris North. 2010. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 55–64. https://doi.org/10.1145/1753326.1753336

[5] Samantha Aziz, Dillon J Lohr, Lee Friedman, and Oleg Komogortsev. 2024. Evaluation of Eye Tracking Signal Quality for Virtual Reality Applications: A Case Study in the Meta Quest Pro. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications* (Glasgow, United Kingdom) *(ETRA '24)*. Association for Computing Machinery, New York, NY, USA, Article 7, 8 pages. https://doi.org/10.1145/3649902.3653347

[6] Tirthankar Bandyopadhyay, Kok Sung Won, Emilio Frazzoli, David Hsu, Wee Sun Lee, and Daniela Rus. 2013. Intention-Aware Motion Planning. In *Algorithmic Foundations of Robotics X*, Emilio Frazzoli, Tomas Lozano-Perez, Nicholas Roy, and Daniela Rus (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 475–491. https://doi.org/10.1007/978-3-642-36279-8_29

[7] Michael Barz, Omair Shahzad Bhatti, and Daniel Sonntag. 2022. Implicit Estimation of Paragraph Relevance From Eye Movements. *Frontiers in Computer Science* 3 (2022), 808507.

[8] Doug A Bowman and Ryan P McMahan. 2007. Virtual reality: how much immersion is enough? *Computer* 40, 7 (2007), 36–43.

[9] Lauren Bradel, Chris North, Leanna House, and Scotland Leman. 2014. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Paris,France, 163–172. https://doi.org/10.1109/VAST.2014.7042492

[10] Saskia Brand-Gruwel, Iwan Wopereis, and Yvonne Vermetten. 2005. Information problem solving by experts and novices: Analysis of a complex cognitive skill. *Computers in Human Behavior* 21, 3 (2005), 487–508.

[11] M Anne Britt and Jean-François Rouet. 2011. c. Research challenges in the use of multiple documents. *Information Design Journal* 19, 1 (2011), 62–68.

[12] Tom Chandler, Maxime Cordeil, Tobias Czauderna, Tim Dwyer, Jaroslaw Glowacki, Cagatay Goncu, Matthias Klapperstueck, Karsten Klein, Kim Marriott, Falk Schreiber, and Elliot Wilson. 2015. Immersive Analytics. In *2015 Big Data Visual Analytics (BDVA)*. IEEE, Hobart, TAS, Australia, 1–8. https://doi.org/10.1109/BDVA.2015.7314296

[13] Ricardo Chavarriaga and José del R Millán. 2010. Learning from EEG error-related potentials in noninvasive brain-computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering* 18, 4 (2010), 381–388.

[14] Hsinchun Chen and Vasant Dhar. 1991. Cognitive process as a basis for intelligent retrieval systems design. *Information Processing & Management* 27, 5 (1991), 405–432.

[15] Haomin Chen, Catalina Gomez, Chien Ming Huang, and Mathias Unberath. 2022. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. Issue 1. https://doi.org/10.1038/s41746-022-00699-2

[16] Kenneth Church and William Gale. 1999. Inverse document frequency (idf): A measure of deviations from poisson. *Natural language processing using very large corpora* 11 (1999), 283–295. https://doi.org/10.1007/978-94-017-2390-9_18

[17] Mark Claypool, Phong Le, Makoto Wased, and David Brown. 2001. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*. Association for Computing Machinery, Santa Fe, New Mexico, USA, 33–40.

[18] Masoud Davari, Daniel Hienert, Dagmar Kern, and Stefan Dietze. 2020. The Role of Word-Eye-Fixations for Query Term Prediction. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 422–426. https://doi.org/10.1145/3343413.3378010

[19] Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. 2021. Towards gaze-based prediction of the intent to interact in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Short Papers)*. Association for Computing Machinery, New York, NY, USA, Article 2, 7 pages. https://doi.org/10.1145/3448018.3458008

[20] Kylie Davidson, Lee Lisle, Ibrahim A Tahmid, Kirsten Whitley, Chris North, and Doug A Bowman. 2023. Uncovering Best Practices in Immersive Space to Think. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, IEEE, Sydney, Australia, 1094–1103. https://doi.org/10.1109/ISMAR59233.2023.00126

[21] Kylie Davidson, Lee Lisle, Kirsten Whitley, Doug A. Bowman, and Chris North. 2023. Exploring the Evolution of Sensemaking Strategies in Immersive Space to Think. *IEEE Transactions on Visualization and Computer Graphics* 29, 12 (2023), 5294–5307. https://doi.org/10.1109/TVCG.2022.3207357

[22] Jan-Peter De Ruiter and Chris Cummins. 2012. A model of intentional communication: AIRBUS (Asymmetric Intention Recognition with Bayesian Updating of

Signals). *Proceedings of SemDial 2012* (2012), 149–50.

[23] Gautier Drusch, JC Bastien, and Stéfane Paris. 2014. Analysing eye-tracking data: From scanpaths and heatmaps to the dynamic visualisation of areas of interest. *Advances in science, technology, higher education and society in the conceptual age: STHESCA* 20, 205 (2014), 25.

[24] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 473–482. https://doi.org/10.1145/2207676.2207741

[25] Martin J. Eppler and Jeanne Mengis. 2004. The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society* 20, 5 (2004), 325–344. https://doi.org/10.1080/01972240490507974

[26] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.* 23, 2 (April 2005), 147–168. https://doi.org/10.1145/1059981.1059982

[27] Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, Gregory J Mills, Ronnie Cann, Wilfried Meyer-Viol, and Patrick GT Healey. 2011. Incrementality and intention-recognition in utterance processing. *Dialogue & Discourse* 2, 1 (2011), 199–233.

[28] Jacek Gwizdka. 2014. Characterizing relevance with eye-tracking measures. In *Proceedings of the 5th Information Interaction in Context Symposium* (Regensburg, Germany) *(IIiX '14)*. Association for Computing Machinery, New York, NY, USA, 58–67. https://doi.org/10.1145/2637002.2637011

[29] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, Amsterdam, Netherlands, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[30] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. In *Pacific Asia Conference on Information Systems*. The Association for Information Systems (AIS), Dubai, 78. https://api.semanticscholar.org/CorpusID:237357256

[31] FJ Hughes. 2005. Discovery, proof, choice: The art and science of the process of intelligence analysis, Case Study 6, 'All Fall Down'. *Unpublished Report* (2005).

[32] Bashirah Ibrahim and Lin Ding. 2023. Students' sensemaking of synthesis physics problems: an exploration of their eye fixations. *International Journal of Science Education* 45, 9 (2023), 734–753. https://doi.org/10.1080/09500693.2023.2175183 arXiv:https://doi.org/10.1080/09500693.2023.2175183

[33] Albrecht Werner Inhoff and Ralph Radach. 1998. Chapter 2 - Definition and Computation of Oculomotor Measures in the Study of Cognitive Processes. In *Eye Guidance in Reading and Scene Perception*, Geoffrey Underwood (Ed.). Elsevier Science Ltd, Amsterdam, 29–53. https://doi.org/10.1016/B978-008043361-5/50003-1

[34] Shoya Ishimaru, Syed Saqib Bukhari, Carina Heisel, Jochen Kuhn, and Andreas Dengel. 2016. Towards an intelligent textbook: eye gaze based attention extraction on materials for learning and instruction in physics. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 1041–1045. https://doi.org/10.1145/2968219.2968566

[35] Halszka Jarodzka and Saskia Brand-Gruwel. 2017. Tracking the reading eye: Towards a model of real-world reading. , 193–201 pages.

[36] Steve Jones and Mark S. Staveley. 1999. Phrasier: a system for interactive document retrieval using keyphrases. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Berkeley, California, USA) *(SIGIR '99)*. Association for Computing Machinery, New York, NY, USA, 160–167. https://doi.org/10.1145/312624.312671

[37] Barbara J Juhasz and Keith Rayner. 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of experimental psychology: Learning, memory, and cognition* 29, 6 (2003), 1312.

[38] Marcel Adam Just and Patricia A Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive psychology* 8, 4 (1976), 441–480.

[39] Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.

[40] Johanna K Kaakinen and Jukka Hyönä. 2010. Task effects on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36, 6 (2010), 1561.

[41] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. *Visual analytics: Definition, process, and challenges*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_7

[42] Diane Kelly and Nicholas J. Belkin. 2001. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 408–409. https://doi.org/10.1145/383952.384045

[43] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, 13–30.

[44] Eric Krokos, Catherine Plaisant, and Amitabh Varshney. 2019. Virtual memory palaces: immersion aids recall. *Virtual reality* 23 (2019), 1–15.

[45] Kuno Kurzhals, Brian Fisher, Michael Burch, and Daniel Weiskopf. 2014. Evaluating visual analytics with eye tracking. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization* (Paris, France) *(BELIV '14)*. Association for Computing Machinery, New York, NY, USA, 61–69. https://doi.org/10.1145/2669557.2669560

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

[47] Joon Hyub Lee, Donghyeok Ma, Haena Cho, and Seok-Hyung Bae. 2021. Post-Post-it: A Spatial Ideation System in VR for Overcoming Limitations of Physical Post-it Notes. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 300, 7 pages. https://doi.org/10.1145/3411763.3451786

[48] Songpo Li and Xiaoli Zhang. 2017. Implicit intention communication in human–robot interaction through visual behavior studies. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 437–448.

[49] Lee Lisle, Xiaoyu Chen, J.K. Edward Gitre, Chris North, and Doug A. Bowman. 2020. Evaluating the Benefits of the Immersive Space to Think. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Atlanta, GA, USA, 331–337. https://doi.org/10.1109/VRW50115.2020.00073

[50] Lee Lisle, Kylie Davidson, Edward JK Gitre, Chris North, and Doug A Bowman. 2021. Sensemaking Strategies with Immersive Space to Think. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, IEEE, Lisboa, Portugal, 529–537.

[51] Tomasz D. Loboda, Peter Brusilovsky, and Jöerg Brunstein. 2011. Inferring word relevance from eye-movements of readers. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (Palo Alto, CA, USA) *(IUI '11)*. Association for Computing Machinery, New York, NY, USA, 175–184. https://doi.org/10.1145/1943403.1943431

[52] Otto Hans-Martin Lutz, Charlotte Burmeister, Luara Ferreira dos Santos, Nadine Morkisch, Christian Dohle, and Jörg Krüger. 2017. Application of head-mounted devices with eye-tracking in virtual reality therapy. *Current Directions in Biomedical Engineering* 3, 1 (2017), 53–56.

[53] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the sweet spot of human-computer configurations: A case study in information extraction. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[54] Kim Marriott, Falk Schreiber, Tim Dwyer, Karsten Klein, Nathalie Henry Riche, Takayuki Itoh, Wolfgang Stuerzlinger, and Bruce H Thomas. 2018. *Immersive analytics*. Vol. 11190. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-030-01388-2

[55] Ann McNamara, Katherine Boyd, Joanne George, Weston Jones, Somyung Oh, and Annie Suther. 2019. Information placement in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Osaka, Japan, 1765–1769.

[56] Masahiro Morita and Yoichi Shinoda. 1994. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '94)*. Springer-Verlag, Berlin, Heidelberg, 272–281.

[57] Sue Newell and Marco Marabelli. 2015. Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems* 24, 1 (2015), 3–14.

[58] Stephanie Ortigue, James C Thompson, Raja Parasuraman, and Scott T Grafton. 2009. Spatio-temporal dynamics of human intention understanding in temporo-parietal cortex: a combined EEG/fMRI repetition suppression paradigm. *PloS one* 4, 9 (2009), e6962.

[59] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.

[60] Raja Parasuraman and Christopher D. Wickens. 2008. Humans: Still Vital After All These Years of Automation. *Human Factors* 50, 3 (2008), 511–520. https://doi.org/10.1518/001872008X312198 PMID: 18689061.

[61] Lu Peng, Dailin Li, Zhaotong Zhang, Tingru Zhang, Anqi Huang, Shaohui Yang, and Yu Hu. 2024. Human-AI collaboration: Unraveling the effects of user proficiency and AI agent capability in intelligent decision support systems. *International Journal of Industrial Ergonomics* 103 (2024), 103629.

[62] Bastian Pfleging, Drea K. Fekety, Albrecht Schmidt, and Andrew L. Kun. 2016. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*

(San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5776–5788. https://doi.org/10.1145/2858036.2858117

[63] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[64] 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. Association for Computational Linguistics (ACL), McLean, VA, USA, 2–4.

[65] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.

[66] Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review* 105, 1 (1998), 125.

[67] Gerard Salton, Edward A Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Commun. ACM* 26, 11 (1983), 1022–1036.

[68] Luisa Sartori, Cristina Becchio, and Umberto Castiello. 2011. Cues to intention: the role of movement information. *Cognition* 119, 2 (2011), 242–252.

[69] Young-Woo Seo and Byoung-Tak Zhang. 2000. Learning user's preferences by analyzing Web-browsing behaviors. In *Proceedings of the Fourth International Conference on Autonomous Agents* (Barcelona, Spain) *(AGENTS '00)*. Association for Computing Machinery, New York, NY, USA, 381–387. https://doi.org/10.1145/336595.337546

[70] Thomas B Sheridan, William L Verplank, and TL Brooks. 1978. Human/computer control of undersea teleoperators. In *NASA. Ames Res. Center The 14th Ann. Conf. on Manual Control*. Massachusetts Institute of Technology, Man-Machine Systems Laboratory, Cambridge, MA, USA, 15 pages.

[71] Richard Skarbez, Nicholas F Polys, J Todd Ogle, Chris North, and Doug A Bowman. 2019. Immersive analytics: Theory and research agenda. *Frontiers in Robotics and AI* 6 (2019), 82.

[72] Ibrahim A. Tahmid, Lee Lisle, Kylie Davidson, Chris North, and Doug A. Bowman. 2022. Evaluating the Benefits of Explicit and Semi-Automated Clusters for Immersive Sensemaking . In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Computer Society, Los Alamitos, CA, USA, 479–488. https://doi.org/10.1109/ISMAR55827.2022.00064

[73] Ibrahim A Tahmid, Lee Lisle, Kylie Davidson, Kirsten Whitley, Chris North, and Doug A Bowman. 2023. Evaluating the Feasibility of Predicting Information Relevance During Sensemaking with Eye Gaze Data. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Sydney, Australia, 713–722. https://doi.org/10.1109/ISMAR59233.2023.00086

[74] J.J. Thomas and K.A. Cook. 2006. A visual analytics agenda. *IEEE Computer Graphics and Applications* 26, 1 (2006), 10–13. https://doi.org/10.1109/MCG.2006.5

[75] Geoffrey Underwood, Lorraine Jebbett, and Katharine Roberts. 2004. Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *Quarterly Journal of Experimental Psychology Section A* 57, 1 (2004), 165–182.

[76] Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. 2022. Designing Transparency for Effective Human-AI Collaboration. *Information Systems Frontiers* 24 (6 2022), 877–895. Issue 3. https://doi.org/10.1007/s10796-022-10284-3

[77] Yao Wang, Yue Jiang, Zhiming Hu, Constantin Ruhdorfer, Mihai Bâce, and Andreas Bulling. 2024. VisRecall++: Analysing and Predicting Visualisation Recallability from Gaze Behaviour. *Proceedings of the ACM on Human-Computer Interaction* 8, ETRA (2024), 1–18.

[78] Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the process of sensemaking. *Organization science* 16, 4 (2005), 409–421.

[79] Laura Jean Wells, Steven Mark Gillespie, and Pia Rotshtein. 2016. Identification of emotional facial expressions: Effects of expression, intensity, and sex on eye gaze. *PloS one* 11, 12 (2016), e0168307.

[80] John Wenskovitch, Corey Fallon, Kate Miller, and Aritra Dasgupta. 2021. Beyond visual analytics: Human-machine teaming for ai-driven data sensemaking. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE, New Orleans, LA, USA, 40–44.

[81] Ryen W White, Joemon M Jose, and Ian Ruthven. 2006. An implicit feedback approach for interactive information retrieval. *Information processing & management* 42, 1 (2006), 166–190.

[82] Heather M Wojton, Daniel Porter, Stephanie T. Lane, Chad Bieber, and Poornima Madhavan. 2020. Initial validation of the trust of automated systems test (TOAST). *The Journal of social psychology* 160, 6 (2020), 735–750.

[83] Hong Xie. 2002. Patterns between interactive intentions and information-seeking strategies. *Information processing and Management* 38, 1 (2002), 55–77.

[84] Poonam Yadav and RP Singh. 2012. An ontology-based intelligent information retrieval method for document retrieval. *International Journal of Engineering Science and Technology* 4, 9 (2012), 3970–3974.

[85] Raquib Bin Yousuf, Nicholas Defelice, Mandar Sharma, Shengzhe Xu, and Naren Ramakrishnan. 2024. LLM Augmentations to support Analytical Reasoning over Multiple Documents . In *2024 IEEE International Conference on Big Data (BigData)*. IEEE Computer Society, Washington DC, USA.