Evaluating the Feasibility of Predicting Information Relevance During Sensemaking with Eye Gaze Data

Ibrahim A. Tahmid*[1], Lee Lisle†[1], Kylie Davidson‡[1], Kirsten Whitley[2], Chris North§[1], and Doug A. Bowman¶[1]

[1]Center for Human-Computer Interaction, Department of Computer Science, Virginia Tech
[2]US Department of Defense

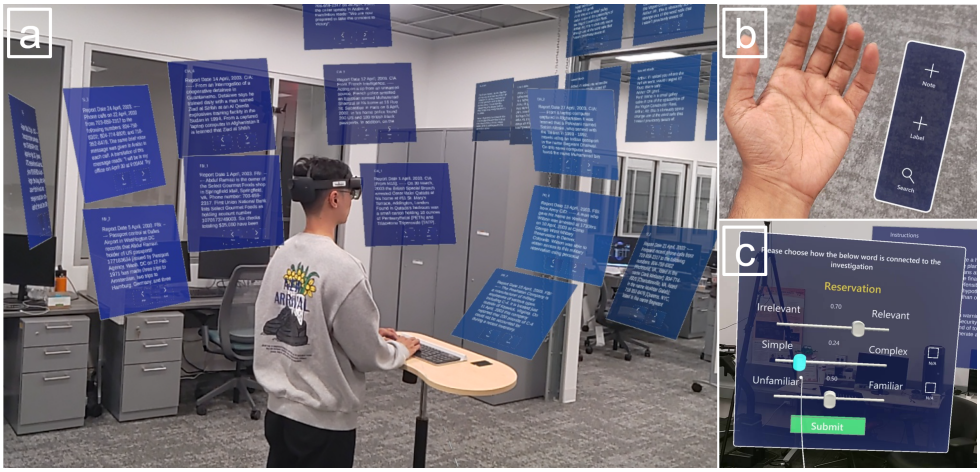

Figure 1: Overview of the EyeST prototype. a) An analyst views and interacts with text documents in Augmented Reality (AR) while eye-tracking data is collected. b) A hand-registered menu allows the analyst to annotate and search the dataset. c) The system retrieves a word that the analyst paid attention to, and the analyst rates it in terms of relevance, complexity, and familiarity.

Abstract

Eye gaze patterns vary based on reading purpose and complexity, and can provide insights into a reader's perception of the content. We hypothesize that during a complex sensemaking task with many text-based documents, we will be able to use eye-tracking data to predict the importance of documents and words, which could be the basis for intelligent suggestions made by the system to an analyst. We introduce a novel eye-gaze metric called 'GazeScore' that predicts an analyst's perception of the relevance of each document and word when they perform a sensemaking task. We conducted a user study to assess the effectiveness of this metric and found strong evidence that documents and words with high GazeScores are perceived as more relevant, while those with low GazeScores were considered less relevant. We explore potential real-time applications of this metric to facilitate immersive sensemaking tasks by offering relevant suggestions.

Index Terms: Immersive Analytics—Sensemaking—Augmented Reality—Human-Computer Interaction; Relevance Perception—Predicted Relevance—Gaze-Based Metric—Multiple Documents

1 Introduction

Visual and immersive analytics tools are often designed to support the complex cognitive task of *sensemaking*. In particular, such tools have focused on sensemaking tasks in which analysts must extract meaningful information from interconnected documents [19, 49]. While these tools have shown great potential [3, 41], analysts must typically read, annotate, and synthesize large amounts of information, making sensemaking with large text-based datasets challenging, no matter how well the tool is designed. We see great promise in combining intelligent analysis aids with immersive analytics tools for sensemaking. Specifically, the availability of eye tracking in many modern immersive display technologies (such as virtual reality (VR) and augmented reality (AR) head-worn displays) provides an opportunity to gather gaze data in real-time during analysis, and to use that data for intelligent assistance.

While previous studies have demonstrated the potential of eye-tracking data in predicting relevance perception during reading tasks [9], its application to sensemaking tasks involving interconnected documents remains largely unexplored. Measures such as fixation duration and count have been effective in predicting relevance judgments for individual words and documents. However, applying these measures to sensemaking tasks poses challenges. One challenge is the influence of word frequency on fixation, as high-frequency words tend to elicit longer total fixations, making the metric unreliable for inferring the analyst's perception of the words. Additionally, sensemaking tasks involve multiple readings of documents at different stages, and fixation duration or count alone cannot fully capture the analyst's perception of the document. Hence, there is a need for a novel eye-gaze metric that can account for the interconnected nature of sensemaking datasets and provide more accurate predictions of analysts' relevance perception.

In this paper, we report on research addressing this need by introducing a novel eye-gaze metric suitable for sensemaking tasks with multiple interconnected documents. This metric considers the

*e-mail: iatahmid@vt.edu
†e-mail: llisle@vt.edu
‡e-mail: kyliedavidson@vt.edu
§e-mail: north@cs.vt.edu
¶e-mail: dbowman@vt.edu

frequency bias in datasets and enables the prediction of analyst-perceived relevance for both documents and individual words. Capturing the interplay between documents, it provides a more comprehensive understanding of analysts' attentional patterns. We evaluated the performance of our eye gaze metric during a typical sensemaking task. Participants were tasked with browsing and extracting information from a collection of interconnected documents, while their eye movements were tracked. After the sensemaking session, participants provided subjective ratings of relevance for both documents and individual words. We compared the predicted relevance based on the eye gaze metric with the analysts' perceived relevance, revealing valuable insights into the metric's effectiveness.

Based on the promising results of our eye gaze metric, we lay the groundwork for the development of an intelligent analysis aid that could be added to immersive analytics tools targeting sensemaking tasks. Such an aid would leverage eye gaze data to provide real-time feedback and suggestions to analysts during the sensemaking process. By highlighting relevant documents and words based on analysts' attentional patterns, the aid aims to enhance analysts' performance and improve the overall efficiency and effectiveness of sensemaking tasks. Our contributions in this work include:

- Introducing an eye gaze metric that is not susceptible to frequency bias in a dataset with multiple documents, and can predict the analyst-perceived relevance of documents and words.
- Evaluating the performance of the metric with a standard sensemaking task, and analyzing the relation between the metric and the analyst-perceived relevance of documents and words.
- Laying the groundwork for developing an intelligent assistive model for sensemaking tasks based on eye gaze data.

2 Related Work

2.1 Sensemaking and Semantic Interaction

Sensemaking is a complex cognitive task that involves browsing documents, extracting meaningful evidence, and synthesizing new insights about one or more topics in a dataset [51]. The task can be divided into two major parts. In the first part, an analyst gathers evidence by browsing the dataset (foraging), while the second part emphasizes organizing and synthesizing information (sensemaking) [63]. Interactive tools from research in visual and immersive analytics support sensemaking by allowing analysts to read, annotate, organize, and synthesize in a visual, spatial setting (e.g., [3, 41]). Andrews et al. [3] showed that large, high-resolution displays can support sensemaking by becoming part of the distributed cognitive process, providing both external memory and a semantic layer. However, for a real-world sensemaking task, the dataset can be quite large which introduces two challenges. First, there needs to be enough space to even visualize all the data at the same time. Second, analysts need assistance to find relevant information without having to resort to exhaustively reading and keeping track of every document. To address the first challenge, Lisle et al. [41] proposed the use of an immersive, three-dimensional space to analyze a large multimedia dataset. Dubbed Immersive Space to Think (IST), this approach allowed analysts to follow different spatial organization strategies [42] during multiple stages in the sensemaking process [16], and improve their overall understanding [41] of the dataset.

Researchers have approached the second challenge from several perspectives. ForceSPIRE [18] used *semantic interaction* to interpret common interactions in spatial analytic processes [3] (such as searching, highlighting, annotating, and repositioning documents) and to update statistical models based on the interactions to suggest documents relevant to the analysts. StarSPIRE [6] built on this idea to propose a visual analytics system that transformed analyst interactions into a combination of a relevance-based foraging model, and a similarity-based synthesis model. The underlying assumption for these models is that if an analyst highlights or searches for a term, it is considered 'relevant' to their strategy, and they would be more interested to explore documents related to that term [50]. Here, 'relevance' relates to user cognition and 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 which holds true for both large 2D displays [6] and immersive spaces [40, 59]. All of these systems rely on explicit user interactions to enhance the visual analytics tool. The inclusion of built-in sensors such as eye-tracking capabilities in VR/AR headsets opens a window of opportunity [44] to use subtler, implicit user interactions to build intelligent models for sensemaking with multiple documents [8].

2.2 Role of Eye Gaze in Reading

Eye gaze provides a wealth of information as it is closely linked to cognitive processing [33, 34, 60] and emotional expression [64]. Finding a negligible lag between eye fixations and cognitive processing, Just and Carpenter [33] went as far as to suggest that what we see is also what we think about. We can infer a lot about a person from their eye gaze behavior. Fixation duration alone is a strong measure to distinguish novice users from experts [1, 28], infer student engagement in extracting and processing information from a set of given sources [26], distinguish reading behaviors of users for different tasks such as comprehension and proofreading [35], and even predict query terms during information processing with high accuracy [14]. Fixation count has also been associated with fixation duration in cognitively processing a word [27, 54], and identifying the relevance of specific areas of interest to readers [38]. More recent studies found the effect of more sophisticated gaze measures such as increased pupil size for novice users compared to experts [1], and gaze velocity being able to predict users' intent to interact [15]. 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 [29] is relatively under-explored. 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 [7]. Thus, in real-world scenarios, the ability to make predictions from eye gaze measures based on reading individual pieces of text decreases [29]. In addition, solely analyzing eye movements does not provide insight into the user's reasoning process for multiple documents as it introduces frequency bias [32]. 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. Hence, we cannot just assume a connection between fixation measures and the user's cognitive process unless we prove it [48]. Therefore, prior to introducing a gaze-based predictor model, it is vital to evaluate the validity of the model by gathering verbal data, such as concurrent thinking aloud or cued retrospective reporting [61], to gain a comprehensive understanding of the users' reasoning behind their gaze behaviors. We aimed to address these gaps by developing a novel gaze measure that is able to infer the users' perception of the relevance of documents and words while reading, and evaluating its ability to infer relevance via explicit confirmation by the user.

2.3 Intelligent Prediction of Text Relevance

It comes as no surprise that systems with intelligent document retrieval features [6, 10, 18, 65, 68] have been studied extensively. They showed that a user's interactions have implicit meanings that help to reveal their information-seeking strategy [67]. For instance, users' interactions with a list of searched documents can provide an understanding of how the searcher's information needs change over time [65]. The searched term itself can help a system to determine which documents the user would be interested in [10]. This may introduce a term-matching problem where the searched term and the

index terms may not match exactly. *Phrasier* [31] tried to address this problem by automatically exploiting predetermined keyphrases from the source documents to create links to similar documents. However, this approach takes some control away from users.

Another approach is to rely more on implicit user actions such as reading time [36, 57], browsing patterns [55], scrolling time [12, 36], or mouse movement [12] to estimate the user-perceived relevance of terms in the documents. Our hypothesis is that the gaze measures could be used in a similar way. Fixation time, for instance, is quite effective in predicting the relevance of individual Web pages [21], and predicting relevant search terms [14, 65] in information-retrieval tasks. The underlying assumption is that the time spent reading a word reflects the user's cognitive processing of that word [27, 53, 54]. This principle has been proven effective in predicting relevance for words [43], paragraphs [5], and documents [22] read by users. McNamara et al. [45] used eye-tracking to measure users' attention to objects of interest, and place labels in an information-rich environment. In this paper, our aim is to focus more on understanding the relationship between gaze measures and cognitive processes to develop intelligent models for immersive sensemaking tasks.

3 Research Questions

Building on the work described in the previous section, our research aimed to answer the following three questions.

RQ1: *How can we design a metric based on eye-tracking data to capture analysts' focus of attention during sensemaking with multiple textual documents?* Prior research has established a link between eye gaze measures, like fixation duration and fixation count, and analysts' attention and cognition. However, it remains uncertain whether these measures are equally reliable when it comes to tasks involving multiple interconnected documents. We propose that while one or a combination of these metrics might indicate analysts' attention at a document level, none of them can be considered dependable for understanding analysts' cognition at the word level due to the influence of frequency bias. Our objective was to develop a gaze metric that aligns with analysts' cognition while not being susceptible to frequency bias.

RQ2: *To what extent does the metric predict the relevance of documents or words?* Our objective was to assess the effectiveness of the metric in determining the analyst-perceived relevance of both documents and words. To achieve this, we conducted an experiment where participants were asked to report on the relevance of documents of words without any knowledge of the metric. We hypothesized that there would be a correlation between the gaze metric values and the relevance ratings provided by analysts.

RQ3: *How can we use the metric in a real-time sensemaking task to provide intelligent suggestions to analysts?* The original motivation of the work was to enable an immersive analytics tool to provide intelligent assistance in the form of suggested documents or words to the analyst. If the metric proves robust and accurate in predicting analyst-perceived relevance, we would then need to understand how to make predictions in real-time, during a sensemaking session, and how to present suggested documents or words to the analyst for effective sensemaking with reduced workload.

4 Designing a Gaze-Based Relevance Metric

Since single eye-gaze measures such as fixation duration or fixation count alone are not reliable to infer cognitive process in real-world scenarios involving multiple interconnected documents [29], we will discuss the possible gaze-based metrics in this section, identify their challenges, and explain how to address them to design a metric for predicting analyst-perceived relevance.

Gaze Duration (GD) refers to the amount of time that an analyst spends fixating (looking for at least 200ms [53]) on a particular document or word before shifting their visual attention elsewhere. For words, we consider the total amount of time spent on a particular word across all documents and term it as GD_w. For documents, since analysts will naturally spend more time gazing at longer documents, we normalize the value by dividing the total amount of time by the number of words in the document and term it as GD_d.

Unique Dwell (UD) refers to the number of unique instances where an analyst fixates on a document or a word.

Z-Score (Z) is a statistical measure that we used to understand how an analyst's attention to a particular document or word differs from the average attention. It indicates how many standard deviations an observed value (GD_x or UD_x) is away from its mean [2] (Eq. 1).

$$Z_{GD_x} = \frac{GD_x - \mu_{GD}}{\sigma_{GD}}; Z_{UD_x} = \frac{UD_x - \mu_{UD}}{\sigma_{UD}} \quad (1)$$

where μ is the mean, σ is the standard deviation, and x represents either document or word. Since both GD and UD contribute to understanding the focus of attention for an analyst [37, 39], we combined them by taking the average of Z_{GD} and Z_{UD}, and assigned them as Z_d for a document and Z_w for a word.

Inverse Document Frequency (IDF) is a statistical measure used in Natural Language Processing (NLP) and information retrieval to evaluate the importance or relevance of a word in a collection of documents [11]. The IDF score of a word is calculated by dividing the total number of documents in a corpus by the number of documents that contain the word and then taking the logarithm of the quotient.

We observed in our initial pilot studies that words with high frequency tended to be ranked higher when we consider GD, UD, or Z_w. IDF helps us reduce that bias by giving higher weight to rare or unique words in the dataset, and lower weight to common words.

GazeScore (GS) is a metric we introduced to rate attention on documents (GS_d) and words (GS_w) based on analysts' eye-tracking data. GS_d is simply Z_d (the average of Z_{GD_d} and Z_{UD_d}, as described above). However, for words, Z_w does not exclusively reflect the analyst's perception of the word. Rather, it also depends on the dataset since the words with high frequencies may receive higher scores. There are two ways we can reduce that bias. First, we can normalize the GD and UD for each word by dividing them by the frequency before taking the average of Z_{GD_w} and Z_{UD_w} (Eq. 1) to get $Z_{w_{norm}}$. Second, we can multiply Z_w with the weighting factor IDF_w (Eq. 2), where $IDF_w = 1$ for the rarest word, and $IDF_w = 0$ for the most common word.

$$GS_w = Z_w * IDF_w \quad (2)$$

Both $Z_{w_{norm}}$ and GS_w are designed to reduce the frequency bias, and there is no trivial way to determine which metric is superior. This led us to run a pilot study to determine which metric best aligns the analyst's eye gaze data with the analyst-perceived relevance of a word during a sensemaking task.

4.1 Evaluation Study to Address Frequency Bias

We recruited two participants (1F) with an average age of 28. Both participants had perfect vision, and tried AR 3-10 times prior to this study. These participants did not take part in the main experiment reported in Section 5.

We followed the same experimental design described in Section 5, but included only the first three phases of the procedure described in Section 5.5. We rated each word by both $Z_{w_{norm}}$ and GS_w, sorted them in descending order, and took the top ten words from each list. We randomized the 20 words, and asked the participant to mark each of them as relevant or irrelevant to the task. For the ten words ranked with $Z_{w_{norm}}$, the participants found three and five words, respectively, to be relevant. In contrast, among the ten words ranked with GS_w, the first participant found nine relevant to the task while the second participant marked all of them as relevant. Thus, we used GS_w as a relevance-predicting metric for the main experiment.

In summary, we answered RQ1 by proposing GS_d and GS_w as metrics to predict the relevance of documents and words. Henceforth, we will refer to both of them as *GazeScores (GS)*.

5 Experiment Design

The goal of our experiment was to evaluate the performance of the *GazeScore* metrics in being able to infer the relevance of documents and words during a sensemaking task. The experiment was approved by the local Institutional Review Board.

5.1 Dataset and Task

We used the Sign of the Crescent dataset [25] which has 41 documents with information on three terrorist plots. However, analyzing all of them would result in a long analysis session, which would make our participants fatigued, compromising their eye-tracking data [24]. This motivated us to reduce the total number of documents to 24, where 20 documents contained information related to the two plots, and four documents were distractors.

We presented the participant with a hypothetical job where they had to play the role of an intelligence analyst. Their task was to analyze the documents and develop a specific hypothesis about any potential terrorist attacks. Their hypothesis needed to identify **who, what, when,** and **where** about the threats or suspicious activities.

5.2 Apparatus

We used a Microsoft HoloLens 2 (HL2) [1] with hand and eye tracking enabled that had a field of view of 54 degrees. The eyes are tracked with two in-device IR cameras at approximately 30 FPS. To avoid common issues while reading in AR [17, 52], we ran multiple tests with humans until we chose the font size of 45 pt for comfortable legibility at up to 2m distance. The participants freely walked around an obstacle-free space of $17'x14'$. The application was implemented using Unity v2020.3.24 with Mixed Reality Toolkit 2.

5.3 System Overview

Participants used hand gestures to interact with the immersive elements (documents, notes, and labels). The only gestures allowed in the system were *pinch* (to grab), *pinch and release* (to press a button from a distance), and *poke* (to press a nearby button). In accordance with previous IST prototypes [41, 42], we allowed the participant to create notes and labels, and to search for words in the dataset (all three features utilized a physical keyboard on a rolling cart; see Figure 1). We developed the prototype in AR so that the participants could still see the keyboard while being immersed in the system. Since our goal was to analyze how the participants' gaze data was related to their perception of the content, we limited the number of notes to just one, forcing the participant to spend more time browsing the dataset rather than writing their findings. We did not have any restriction on the number of labels to help them keep track of the contents in document clusters.

5.4 Participants

We had 12 volunteer participants (3F, 1 Non-Binary) with an average age of 21.08 ($\sigma = 1.24$). All participants had normal or corrected vision (5 with glasses, 2 with contact lenses). Prior work [47] has shown that participants with corrected vision have not had any issues with eye-tracking software. One participant reported having been diagnosed with an eye condition, astigmatism. However, they had no trouble calibrating their eyes with the HL2. Five participants experienced AR prior to this study once or twice, two had tried AR 3-10 times, one participant tried AR more than 10 times, and four participants experienced AR for the first time in this study.

[1] `https://www.microsoft.com/en-us/hololens/hardware`

5.5 Procedure

The study was divided into four phases: Calibration, Tutorial, Main Study, and Post-Study Questionnaire.

Calibration We welcomed the participant to the lab space. The participant signed a consent form and completed a demographic questionnaire. The experimenter started the built-in eye-calibration program for the HL2, and handed over the headset to the participant to complete the calibration process. The phase ended with the successful calibration of the participant's eyes with the HL2.

Tutorial The participants learned how to use the system with a set of 10 documents randomly chosen from the MAVERICK dataset [30]. The participants completed a set of predefined tasks that were designed to help them learn how to a) grab and move documents with their hands, b) create a note, c) create labels, and d) browse longer documents with more than one page. This phase took approximately five minutes to complete.

Main Study We introduced the participants to the 24 documents in the dataset and explained the tasks described in Section 5.1. We gave the participants 40 minutes to complete the tasks and offered to extend up to 10 minutes upon request. The participants took 43.04 minutes on average to complete the tasks. Upon completion, the participants wrote a report on their findings in an IST note, while they were still wearing the HL2.

Post-Study Questionnaire We presented the first part of this questionnaire as a way for participants to *help out a colleague*. **First,** we asked the participant to write down ten keywords that they thought were most relevant to the task. The keywords could be anything the participant wrote down, and we refer to them as 'Free Response Words' henceforth. We kept this part at the beginning as we did not want the participant to be biased by the words scored by the *GazeScore* metric. **Second,** we calculated and sorted the *GazeScore* for all the words that the participant read throughout the sensemaking process, and gave each of them a *GazeRank (GR)* based on their index on the sorted list (inspired by [14]) to have a uniform metric for better comparison across participants. From the list, we extracted 10 words with the highest rank, 10 words with the lowest rank, and 10 words from the middle, with five ranked above the median and five ranked at or below the median. We randomized the 30 words and refer to them as 'Rated Words' henceforth. Our goal was to examine the relationship between *GazeScore* and the relevance of these words. However, the eye-tracking data could also be influenced by external factors such as visual complexity [66] or familiarity [13]. To make sure that the *GazeScore* is not affected by these factors, we asked the participants to rate each word on a scale of 0-1 not only for relevance, but also for complexity and familiarity (see Figure 1) so that we can rule them out as influencers.

Third, we asked the participants to identify four documents that they found relevant to the task. The participants identified the documents by a highlight button on the document. We identify these documents as 'Free Response Documents' henceforth. **Fourth,** we sorted all the documents that the participant read throughout the sensemaking process based on their *GazeScore* and gave them a *GazeRank*. We extracted four documents with the highest rank, four with the lowest rank, and four from the middle, with two ranked above the median and two ranked at or below the median. We refer to these documents as 'Rated Documents' henceforth. We randomized the 12 Rated Documents and asked the participants to rate them in terms of relevance, and complexity.

The second part of the post-study questionnaire had a NASA TLX questionnaire [23] and a semi-structured interview about their strategy, and expected intelligent suggestions in sensemaking tasks.

5.6 Data Collection and Measures

We collected a variety of data to measure participants' actions, their perception of relevance, and the validity of eye gaze data.

During the pre-study phase, we screened for participants with any recent eye-related diagnosis that could invalidate the eye-tracking data. We also asked participants to rate their fatigue on a scale of one to ten [56], to make sure the eye gaze data collected during the study were trustworthy, as prior work [58] identified fatigue to cause changes in some eye gaze measures such as blink rate.

During the main study, we collected the eye gaze origin and direction with 30Hz frequency, which allowed us to analyze eye gaze patterns throughout the sensemaking process. We also automatically logged every time the participant spent time reading a document or a word. We kept logs of every time they interacted with a note, label, or document that helped us to analyze participants' interactions with the dataset. The final report stating the findings by the participant was also saved in a separate file, which enabled us to evaluate the correctness of their sensemaking task.

During the post-study phase, we collected the names of the free-response documents and words reported by each participant. Additionally, we recorded the self-reported ratings of relevance, complexity, and familiarity (word only) on each Rated Word and Rated Document with timestamps. We also collected the NASA TLX measures from each participant at the end of the study which allowed us to measure the task load. The final interview with the participant helped us gain a better perspective of participants' actions, and what they expected from an intelligent system helping in a similar task.

6 Results

In this section, we present the outcomes of the user study where we evaluated the performance of *GazeScore* on both the document level and the word level. We collected the eye-tracking data from the moment the participants started looking at their initial document until they began composing their report. We tested our data for normality and homogeneity of variances before applying ANOVA.

6.1 Effect of *GazeScore* on Relevance

We established a relevance threshold of 0.5. Documents and words with a user-defined relevance score exceeding 0.5 were categorized as relevant, while those below were deemed irrelevant. We conducted a t-test analysis to examine the difference in *GazeScore* between relevant and irrelevant documents (Figure 2). The results revealed a statistically significant difference ($t = 4.83, p \leq 0.001$) with a moderate effect of Cohen's d= 0.61, indicating that the mean *GazeScore* for relevant documents (0.55) was significantly higher than the irrelevant documents (-0.48).

We also conducted a t-test analysis to examine the effect of *GazeScore* on word relevance (Figure 2). The results revealed a statistically significant difference ($t = 4.59, p \leq 0.001$) with a moderate effect of Cohen's d= 0.51, indicating that the mean *GazeScore* was significantly higher for relevant words (2.06) than the irrelevant words (0.31). Interestingly, we observed an increased prevalence of outliers among the irrelevant words. These outliers primarily consisted of words with high *GazeScore* that participants labeled as irrelevant. Upon closer examination, we discovered that many of these words were directly linked to the task's solution, such as the name of the prime suspect. The participants' inclination to deem them irrelevant indicates their inability to solve the task in its entirety which may account for the abundance of outliers in the data.

6.2 Performance of *GazeScore* on Document Level

We present the results on the performance of *GazeScore* in inferring relevance of Free Response Documents, and Rated Documents.

6.2.1 Free Response Documents

The participants selected four relevant documents for their report without any prior knowledge about *GazeScore*. On average, it took them 1.55 minutes to make their choice. 38% of the Free Response Documents were in the Top 4 of the sorted *GazeScore* list, compared to 16.67% probability for a random document to be in the Top 4.

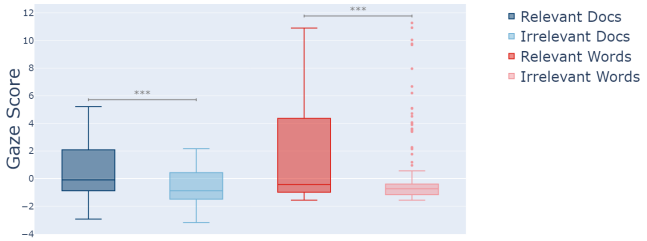

Figure 2: *GazeScore* for relevant content (both documents and words) are higher than irrelevant content.

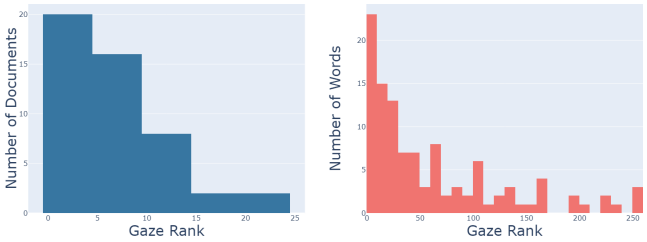

Figure 3: *GazeRank* Distribution for Free Response Documents and Words

We present the distribution of *GazeRank* across all the Free Response Documents in Figure 3. The histogram has a downward slope, which suggests that the majority of the Free Response Documents selected by the participants were high on the *GazeScore* scale. The median *GazeRank* for Free Response Documents is 5.5. We discovered no correlation ($\rho_{spearman} = 0.1$) between documents' selection order and their *GazeRank*. This finding was expected since we did not ask the participants to pick the documents in any order.

6.2.2 Rated Documents

The participants rated 12 documents from three different levels of *GazeScore* (top, middle, bottom) in terms of relevance, and complexity. On average, they took 17.12 seconds to rate each document. There was no correlation between the time to rate and the *GazeScore*. We conducted a one-way ANOVA to examine the effect of the *GazeScore* level on participant-reported relevance. The results showed a significant ($F(2, 139) = 13.07, p \leq 0.0001$) effect (Figure 4). Post-hoc analysis with Bonferroni correction revealed that mean relevance for documents in the top *GazeScore* level (0.8) was rated as significantly more relevant ($p \leq 0.001$) compared to documents in the bottom *GazeScore* level (0.52) with a large effect of Cohen's d=1.07. However, no significant differences in relevance were found between the top and middle *GazeScore* levels ($p = 0.06$) or between the middle and bottom *GazeScore* levels ($p = 0.07$). We found no correlation between *GazeScore* and complexity.

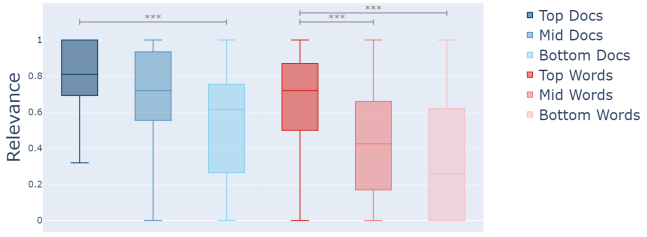

Figure 4: Top-ranked documents are rated as more relevant than bottom-ranked documents. Top-ranked words are rated as more relevant than both mid-ranked and bottom-ranked words.

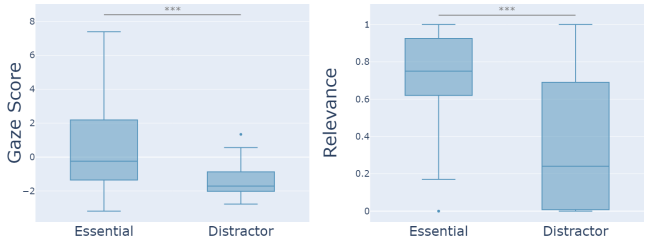

Figure 5: Distractor documents end up with a lower *GazeScore* than documents with actual plots. They are also rated as less relevant by the analysts. The annotations at the top of each boxplot represent the level of significance of the difference with the distractor documents.

We performed additional analysis to explore the influence of document content on *GazeScore*. The documents were categorized into two groups based on ground truth: *Essential* documents containing relevant information and *Distractor* documents unrelated to any relevant information. We conducted a t-test analysis to compare participant-reported relevance between the essential and distractor documents (Figure 5). The results indicated a significant difference ($t = 5.21, p \leq 0.001$) with a large effect size (Cohen's d= 1.48), revealing that the mean relevance for essential documents (0.72) was significantly higher than that of distractor documents (0.34). Similarly, a t-test analysis was performed to compare *GazeScore* between the essential and distractor documents (Figure 5). The results demonstrated a statistically significant difference ($t = 6.68, p \leq 0.001$) with a large effect size (Cohen's d= 0.9), indicating that the mean *GazeScore* for essential documents (0.53) was significantly higher than that of distractor documents (-1.45).

6.3 Performance of *GazeScore* on Word Level

We present the results on the performance of *GazeScore* in inferring relevance of the Free Response Words, and the Rated Words.

6.3.1 Free Response Words

The participants wrote down 10 keywords relevant to their findings without any prior knowledge of the *GazeScore*. It took them an average of 2.95 minutes to report these words. Out of the 120 words reported, 110 were directly related to the dataset. In this section, we only focus on these 110 words while the remaining 10 words will be discussed in Section 7. Out of these 110, 19.17% words ranked within the Top 10 of the sorted *GazeScore* list. Given that the dataset comprised 502 unique words, the probability of a word being randomly included in the Top 10 is 1.9%.

Figure 3 displays the distribution of *GazeRank* among the Free Response Words which has a downward slope, indicating that the participants mostly selected words with a high *GazeScore*. The median *GazeRank* for Free Response Words was 34.5. We found no correlation between the words' selection order and their *GazeRank*.

6.3.2 Rated Words

The participants rated 30 words from three different *GazeScore* levels (top, middle, bottom) in terms of their relevance, complexity, and familiarity. On average, it took them 11.04 seconds to rate each word. We observed no correlation between the time to rate the words and their *GazeScore*. We found a moderate positive correlation between the *GazeScore* and relevance ($\rho_{pearson} = 0.48$). However, we did not find any significant correlation between *GazeScore* and either complexity or familiarity.

A one-way ANOVA to examine the effect of *GazeScore* level on the word relevance revealed a statistically significant ($F(2,275) = 23.78, p \leq 0.0001$) impact. Post-hoc analysis with Bonferroni correction showed (Figure 4) that the mean relevance of words in the top *GazeScore* level (0.66) was significantly higher ($p \leq 0.001$) than the middle (0.42) and bottom *GazeScore* levels (0.35) with moderate (0.75) and large (1.02) effects. We found no significant effect of *GazeScore* levels on the complexity or familiarity of the words.

6.4 Change of *GazeRank* over Time

Our aim was to explore how the *GazeRank* changed over time for relevant and irrelevant documents among all participants. We used the Locally Weighted Scatterplot Smoothing technique to draw trend lines for the median *GazeRank* of relevant and irrelevant documents, the trend line for the p-value showing the difference between the two median *GazeRank*s, and the trend line for the total number of documents read by the participants throughout the task (Figure 6). The timestamps started from around 500 seconds as we did not consider the data from the tutorial phase.

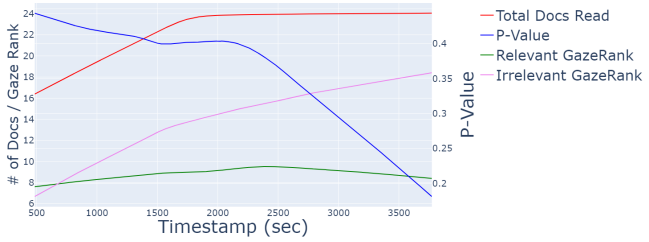

Figure 6: Trend lines to show how Median *GazeRank* changes over time for relevant and irrelevant documents.

Initially, the median *GazeRank* for irrelevant documents was lower than that of relevant documents, but it soon started to increase on an upward slope, and the two lines started diverging. Halfway through the timeline, the p-value started to decline, and the participants had roughly gone through all the documents at least once.

We also examined the trend lines for relevant and irrelevant words to observe any patterns. Figure 7 displays a similar divergence between the two, with significant differences emerging by the end of the session. In comparison to the documents, the trend line for the total number of words read by participants shows a consistent upward trend without reaching a saturation point. Another notable distinction is that the p-value shows an early decrease, suggesting that the distinction between relevant and irrelevant words becomes increasingly evident as the session progresses.

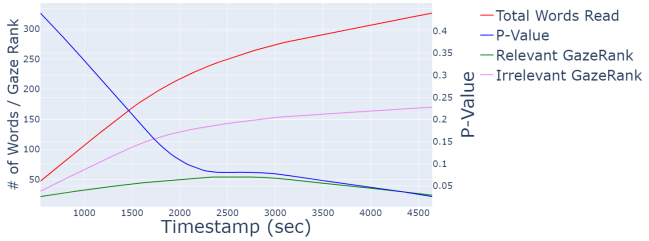

Figure 7: Trend lines to show how Median *GazeRank* changes over time for relevant and irrelevant words.

6.5 *GazeScore* as a Relevance Predictor

Since *GazeScore* varies for different users based on their eye gaze behavior during reading, it is not feasible to apply a universal threshold to this metric. However, we can fashion a predictor from *GazeRank* with the following equation:

$$PredictedRelevance_x = 1 - \frac{GazeRank_x}{GazeRank_{max}} \quad (3)$$

where the value ranges from 0 to 1 for each x with a cutoff threshold of 0.5. Documents or words with values above 0.5 are predicted as relevant, while those below are predicted as irrelevant.

When we applied this rule to the documents, we achieved a precision of 90%, recall of 62.14%, and an F1 score of 0.73. For the

words, applying the same rule yielded a precision of 60.5%, a recall of 50.73%, and an F1 score of 0.55.

7 Discussion

This section presents an interpretation of the results from the user study in the context of answering our research questions.

7.1 Performance of *GazeScore* (RQ2)

Among the numerous measures of attention available from eye-tracking data [15], fixation duration and fixation count are commonly used to interpret users' focus of attention [37, 39]. However, neither measure can accurately capture attention focus at a lower granularity, such as individual words in a document, especially when dealing with multiple documents with overlapping words in different contexts. To address this challenge, we proposed *GazeScore* as a metric for both documents and words.

Our findings showed that *GazeScore* effectively differentiated relevant and irrelevant content at both document and word levels. However, when used as a predictor, *GazeScore* demonstrated higher precision in predicting document relevance compared to word relevance. The lower performance of *GazeScore* for words can be attributed to the participants' inability to complete the entire task. Despite spending more time, on average, than the allotted time, none of the participants were able to identify both major plots in the task. We observed instances where participants paid attention to certain crucial words related to the task but were unable to recognize their relevance, resulting in lower relevance ratings for those words. In contrast, this issue was not prominent for documents, as most of them contained multiple relevant plot points, of which at least one was successfully identified by the participants.

Free Response Documents: We initially hypothesized that the *GazeScore* for the Free Response Documents would be higher than most, resulting in a low *GazeRank*. Our results showed that the median *GazeRank* of the reported documents was 5.5. One possible explanation for this finding is that among the 24 documents, 20 were essential to the task in some way. As a result, *GazeRank* for the documents picked by participants ranged from 0 to 19. Looking back, if we had asked participants to pick documents they spent the most time on, we might have observed a lower median *GazeRank*. However, asking participants to report documents they focused on would be too direct and would not provide insight into the relationship between the participants' eye gaze data and cognitive process, so we stand by our wording for this study.

Upon analysing of the final reports, we discovered some intriguing perspectives. For instance, P14 had picked documents with low *GazeRank*s of 0, 1, 3, and 5 even though they had the lowest score on their report. P14 stated in their interview that they "*felt fatigued and did not have the fullest understanding of the situation.*" We interpret this to mean that due to being in a hurry, P14 did not spend enough time connecting the dots between the documents. Rather than reporting documents relevant to the task, they reported the first documents that came to mind. This may imply that *GazeScore* can be effective in predicting a participant's memory of content. This also aligns with findings from Strien et al. [62] where they found that the participants with prior attitudes towards a topic had significantly different fixation between relevant and irrelevant content, while participants with no prior attitude showed no such behavior.

Rated Documents: Our study did not uncover a direct correlation between the *GazeScore* and the relevance of the documents. This may be attributed to our decision to limit participants' ratings to a subset of the documents in order to prevent the study from becoming excessively lengthy. However, we still found significant relevance differences between documents from the top and bottom *GazeScore* levels. This finding not only reflects the participants' perception of relevance but also sheds light on the nature of the dataset. As previously discussed, 20 out of 24 documents were essential to the sensemaking task, while the rest were distractors. This may be the reason why there was no difference in relevance between the documents from the top and middle *GazeScore* levels. We suggest that the participants were able to identify the distractors successfully, and avoided them during sensemaking, resulting in a lower *GazeScore*, and lower relevance. We examined this assumption and discovered that the distractors were indeed significantly different from the essential documents in terms of both *GazeScore* and relevance.

Free Response Words: We initially hypothesized that the *GazeScore* for the Free Response Words would be higher than most, resulting in a low *GazeRank*. We found the median *GazeRank* for the reported words to be 34.5. Despite this high value, we find the result to be promising in predicting the participants' perception of relevance at the word level, considering that they read over 300 words on average to solve the task. Nonetheless, we must highlight some caveats about using keywords reported by the participants. One thing to look out for is that some participants may consider a phrase as a keyword, leading to ambiguity in selecting a specific word. For example, P05 reported *AMTRAK train* as a free response word. In such cases, we opted for the more specific word in the phrase (*AMTRAK* in this example) and ignored the rest.

Another challenge was that since the participants were recalling words from memory, the reported keywords were not always direct matches from the dataset (*term-matching* problem [10]). For instance, P06 reported *Population Dense Areas* as one of the relevant keywords. While it may make sense to a human, identifying the relevant text in the corpus for such keywords is non-trivial and out of scope for this study. Hence, we excluded such words from our analysis, leaving us with a total of 110 reported words out of 120.

Rated Words: While a document may be tagged as *distractor* or *essential* based on the summary of its content, a word can not be tagged as such since its meaning can vary depending on its context. For instance, the word 'train' in the sentences 'He took train #174 to Atlanta, GA' and 'Train stations are busy' has different meanings. The former contains crucial information that can be relevant to the story, while the latter is more generic and does not require further attention. Therefore, even if the reader focuses on the word 'train' in the first sentence, they would not necessarily do so in the second sentence, reducing the overall attention given to the word. In a dataset where there are many such sentences with connecting evidence, readers tend to focus on the relevant words more than the others as they continue reading more documents. We can see a reflection of this characteristic from our participants, where the words on the top *GazeScore* level were rated as more relevant than the other two levels, even though there were no differences in word relevance between the middle and bottom *GazeScore* levels.

Comparison with Gaze Duration and Unique Dwell: We conducted a comparative analysis involving *GazeScore* along with two established gaze-based metrics: Gaze Duration (GD) and Unique Dwell (UD). We ranked each relevant document and word using *GazeScore*, GD, and UD. Results showed that *GazeScore* outperformed GD and UD for 50% and 54% of the documents, respectively. In the case of words, *GazeScore* achieved a better ranking than GD and UD for 75% and 37% of the words, respectively. Interestingly, words with high UD (top 10) were perceived as notably more complex by participants, aligning with previous research [13, 66]. There was no discernible difference in terms of complexity for the other two metrics. One possible explanation may be related to the design of the sensemaking task. Even if they came across a complex document or an unfamiliar word, they did not focus much on that content unless they found it relevant to the task. Another possible explanation is the simplicity of the dataset itself. The documents had a Flesch–Kincaid readability score [20] of around 60-80, which is equivalent to 8th-grade reading content.

Given that UD performs nearly as effectively as *GazeScore* in gauging relevance and also captures participants' perceptions of

word complexity, one could make a case for favoring UD as a superior metric for inferring the perceived relevance of documents and words during sensemaking. However, it's important to acknowledge potential limitations. With an increase in dataset complexity, UD might start to downplay perceived relevance, and in datasets with frequent word repetition, it might disregard frequency biases since it does not account for them.

7.2 Application of the Eye Gaze Data (RQ3)

Having established the potential of eye gaze data to infer an analyst's perception of content relevance in a sensemaking task, we can leverage this information to develop intelligent sensemaking-assistant models, or contribute additional signals to other models such as ForceSPIRE [18] to improve their accuracy.

This data can be utilized in two key ways. **First**, we can provide real-time visualization of the gaze data that allows analysts to gain insight into their own activities. A few examples would be:

- Display documents with a border whose color or thickness would reflect the attention paid to that document by the analyst.
- After reading a document, visually tag the document with the Top N words that got the most attention. The list of tags could update after each read.
- Upon opening the search bar, display a list of potential search terms based on the analyst's attention.

By providing real-time visualization of gaze data, we may provide analysts with an externalization of their mental concepts, leading to a potential reduction in the cognitive workload associated with sensemaking [4, 59]. In addition, analysts may be more aware of their coverage of the dataset, avoiding the possibility of having their attention captured by only a few documents or concepts.

Second, we can leverage the gaze data to enhance information retrieval that goes beyond the analyst's activities, and may vary depending on the sensemaking stage. During the initial stage of *exploring* evidence files, analysts can benefit from receiving additional information based on their actions, such as search queries [10, 65]. With *GazeScore*, the system can curate search results, prioritizing content that is similar to the analyst's area of interest. As analysts progress in their task and start *organizing* the evidence, they can gain assistance from the system in grouping similar documents together [40, 59]. Applying *GazeScore* on top of a clustering algorithm can provide analyst-perceived relevance of the groups of documents helping the analysts refine their strategy. As the analysts move on to *synthesizing* the collected evidence, the system can aid by searching for documents that are relevant to those they have already reviewed [9, 31]. By predicting words that are relevant to the analyst's strategy the system can retrieve unread documents from the dataset. The search results can be further refined by comparing their similarity with the documents predicted as relevant by the gaze data.

Throughout different stages of the sensemaking process, the **level of system assistance** can vary as well. Striking the appropriate balance between providing helpful suggestions and avoiding excessive interference by the intelligent system is essential to keep the model valuable without overwhelming or frustrating the analyst [46, 59].

While we have insights on *how* to utilize gaze data, we still need to determine *when* our gaze data starts getting effective in predicting relevance. Figure 6 shows that the difference between relevant and irrelevant documents begins to increase after the analysts have reviewed all documents at least once. This implies that the gaze data may not have been sufficiently effective earlier in the sensemaking process. Interestingly, despite supposedly reviewing all documents, the difference in *GazeRank* keeps increasing after the halfway point. This indicates that analysts pay more attention to documents they perceive as relevant after their initial triage, leading to greater divergence in *GazeRank*. This implies that gaze data for documents may not be as useful during the initial stages of sensemaking when there are a limited number of documents. However, it becomes more effective in discerning document relevance over time, proving valuable in sensemaking tasks with larger document sets.

On the other hand, we showed that the distinction between relevant and irrelevant words becomes evident before analysts have read all the available words. This highlights the potential of utilizing gaze data for words even with smaller datasets. We can leverage this finding by integrating gaze data for words at an earlier stage of sensemaking, such as through the implementation of search query suggestions, to enhance analysts' performance from the beginning, leading to improved outcomes throughout the analysis process.

8 Future Work

While this paper presents a novel eye-gaze metric for sensemaking tasks with interconnected documents and demonstrates its effectiveness, there are several avenues for future research in this area.

Refinement of the Eye-Gaze Metric: Further refinement and validation of the *GazeScore* can enhance its accuracy and applicability. Integrating additional factors that influence users' perception of relevance, such as semantic context, can lead to a more comprehensive metric. Additionally, exploring the potential of machine learning techniques to optimize the metric's performance and adapt it to different sensemaking scenarios should be considered.

Dynamic Adaptation of the Assistive Model: By updating the *GazeScore* metric in real-time, the intelligent assistive model can adapt to evolving user needs and provide more personalized and context-aware support. This could involve dynamically adjusting the threshold for relevance prediction, and adapting the feedback mechanism based on user interactions.

Integration with NLP Techniques: Integrating *GazeScore* with advanced NLP techniques can enrich the feedback and support provided by the assistive model. For example, utilizing NLP algorithms for topic modeling, and entity recognition can augment the relevance assessment process enabling more sophisticated assistance, such as extracting meaningful relationships between documents.

9 Conclusion

This paper addresses the need for a novel eye-gaze metric in sensemaking tasks with interconnected documents. Previous research has shown the potential of eye-tracking data in predicting users' perception of relevance, but its application to sensemaking tasks is limited. The newly introduced eye-gaze metric considers word frequency and captures the interconnected nature of the dataset, leading to more accurate predictions of relevance. The metric's performance was evaluated through a standard sensemaking task, comparing predicted relevance with users' subjective ratings. The results demonstrate the effectiveness of the eye-gaze metric and lay the foundation for an intelligent assistive model in sensemaking that can utilize eye-gaze data to provide real-time feedback and enhance analysts' performance, improving the efficiency and effectiveness of sensemaking tasks. Overall, this research contributes valuable insights into intelligent sensemaking with eye tracking, and offers the potential for more effective information extraction and synthesis.

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] S. D. Aljehane, B. Sharif, and J. I. Maletic. Studying developer eye movements to measure cognitive workload and visual effort for expertise assessment. *Proceedings of the ACM on Human-Computer Interaction*, 7(ETRA):1–18, 2023.

[2] E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

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

[4] C. Andrews and C. North. Analyst's workspace: An embodied sensemaking environment for large, high-resolution displays. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 123–131. IEEE, 2012.

[5] M. Barz, O. S. Bhatti, and D. Sonntag. Implicit estimation of paragraph relevance from eye movements. *Frontiers in Computer Science*, 3:808507, 2022.

[6] L. Bradel, C. North, and L. House. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 163–172. IEEE, 2014.

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

[8] M. A. Britt and J.-F. Rouet. c. research challenges in the use of multiple documents. *Information Design Journal*, 19(1):62–68, 2011.

[9] G. Buscher, A. Dengel, R. Biedert, and L. V. Elst. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(2):1–30, 2012.

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

[11] K. Church and W. Gale. Inverse document frequency (idf): A measure of deviations from poisson. *Natural language processing using very large corpora*, pp. 283–295, 1999.

[12] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pp. 33–40, 2001.

[13] L. Copeland and T. Gedeon. The effect of subject familiarity on comprehension and eye movements during reading. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, pp. 285–288, 2013.

[14] M. Davari, D. Hienert, D. Kern, and S. Dietze. The role of word-eye-fixations for query term prediction. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp. 422–426, 2020.

[15] B. David-John, C. Peacock, T. Zhang, T. S. Murdison, H. Benko, and T. R. Jonker. Towards gaze-based prediction of the intent to interact in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–7, 2021.

[16] K. Davidson, L. Lisle, K. Whitley, D. A. Bowman, and C. North. Exploring the evolution of sensemaking strategies in immersive space to think. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[17] M. Dunleavy and C. Dede. Augmented reality teaching and learning. *Handbook of research on educational communications and technology*, pp. 735–745, 2014.

[18] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 473–482, 2012.

[19] K. Fisher, S. Counts, and A. Kittur. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 247–256, 2012.

[20] R. Flesch. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007, 2007.

[21] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve the search experiences. In *Talk presented at SIGIR03 Workshop on Implicit Measures of User Interests and Preferences*, 2003.

[22] J. Gwizdka. Characterizing relevance with eye-tracking measures. In *Proceedings of the 5th information interaction in context symposium*, pp. 58–67, 2014.

[23] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[24] J. F. Hopstaken, D. van der Linden, A. B. Bakker, M. A. Kompier, and Y. K. Leung. Shifts in attention during mental fatigue: Evidence from subjective, behavioral, physiological, and eye-tracking data. *Journal of Experimental Psychology: Human Perception and Performance*, 42(6):878, 2016.

[25] F. Hughes and D. Schum. Discovery-proof-choice, the art and science of the process of intelligence analysis-preparing for the future of intelligence analysis. *Washington, DC: Joint Military Intelligence College*, 2003.

[26] B. Ibrahim and L. Ding. Students' sensemaking of synthesis physics problems: an exploration of their eye fixations. *International Journal of Science Education*, pp. 1–20, 2023.

[27] A. W. Inhoff and R. Radach. Definition and computation of oculomotor measures in the study of cognitive processes. *Eye guidance in reading and scene perception*, pp. 29–53, 1998.

[28] S. Ishimaru, S. S. Bukhari, C. Heisel, J. Kuhn, and A. Dengel. 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*, pp. 1041–1045, 2016.

[29] H. Jarodzka and S. Brand-Gruwel. Tracking the reading eye: Towards a model of real-world reading, 2017.

[30] M. Jenkins, A. Bisantz, J. Llinas, and R. Nagi. Maverick synthetic murder mystery dataset. 2014.

[31] S. Jones and M. S. Staveley. 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*, pp. 160–167, 1999.

[32] B. J. Juhasz and K. Rayner. 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):1312, 2003.

[33] M. A. Just and P. A. Carpenter. Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480, 1976.

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

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

[36] D. Kelly and N. J. Belkin. 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*, pp. 408–409, 2001.

[37] Y. Kim and A. Varshney. Persuading visual attention through geometry. *IEEE Transactions on Visualization and Computer Graphics*, 14(4):772–782, 2008.

[38] K. Kurzhals, B. Fisher, M. Burch, and D. Weiskopf. Evaluating visual analytics with eye tracking. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pp. 61–69, 2014.

[39] K. Kurzhals, M. Höferlin, and D. Weiskopf. Evaluation of attention-guiding video visualization. In *Computer graphics forum*, vol. 32, pp. 51–60. Wiley Online Library, 2013.

[40] J. H. Lee, D. Ma, H. Cho, and S.-H. Bae. 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*, pp. 1–7, 2021.

[41] L. Lisle, X. Chen, J. E. Gitre, C. North, and D. A. Bowman. Evaluating the benefits of the immersive space to think. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 331–337. IEEE, 2020.

[42] L. Lisle, K. Davidson, E. J. Gitre, C. North, and D. A. Bowman. Sensemaking strategies with immersive space to think. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 529–537. IEEE, 2021.

[43] T. D. Loboda, P. Brusilovsky, and J. Brunstein. Inferring word relevance from eye-movements of readers. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pp. 175–184, 2011.

[44] O. H.-M. Lutz, C. Burmeister, L. F. dos Santos, N. Morkisch, C. Dohle, and J. Krüger. Application of head-mounted devices with eye-tracking

in virtual reality therapy. *Current Directions in Biomedical Engineering*, 3(1):53–56, 2017.

[45] A. McNamara, K. Boyd, J. George, W. Jones, S. Oh, and A. Suther. Information placement in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1765–1769. IEEE, 2019.

[46] C. Morrison, E. Cutrell, M. Grayson, A. Thieme, A. Taylor, G. Roumen, C. Longden, S. Tschiatschek, R. Faia Marques, and A. Sellen. Social sensemaking with ai: Designing an open-ended ai experience with a blind child. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.

[47] S. Oney, N. Rodrigues, M. Becher, T. Ertl, G. Reina, M. Sedlmair, and D. Weiskopf. Evaluation of gaze depth estimation from eye tracking in augmented reality. In *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–5, 2020.

[48] J. L. Orquin and K. Holmqvist. Threats to the validity of eye-movement research in psychology. *Behavior research methods*, 50:1645–1656, 2018.

[49] S. A. Paul and M. R. Morris. Cosense: enhancing sensemaking for collaborative web search. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1771–1780, 2009.

[50] P. Pirolli and S. Card. Information foraging. *Psychological review*, 106(4):643, 1999.

[51] P. Pirolli and S. Card. 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, pp. 2–4. McLean, VA, USA, 2005.

[52] P.-L. P. Rau, J. Zheng, Z. Guo, and J. Li. Speed reading on virtual reality and augmented reality. *Computers & Education*, 125:240–245, 2018.

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

[54] E. D. Reichle, A. Pollatsek, D. L. Fisher, and K. Rayner. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125, 1998.

[55] Y.-W. Seo and B.-T. Zhang. Learning user's preferences by analyzing web-browsing behaviors. In *Proceedings of the fourth international conference on Autonomous agents*, pp. 381–387, 2000.

[56] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro. Visual analogue scale to evaluate fatigue severity (vas-f). In *STOP, THAT and one hundred other sleep scales*, pp. 399–402. Springer, 2011.

[57] M. M. Shinoda and Yoichi. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval*, pp. 272–281, 2012.

[58] J. A. Stern, D. Boyer, and D. Schroeder. Blink rate: a possible measure of fatigue. *Human factors*, 36(2):285–297, 1994.

[59] I. A. Tahmid, L. Lisle, K. Davidson, C. North, and D. A. Bowman. Evaluating the benefits of explicit and semi-automated clusters for immersive sensemaking. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 479–488. IEEE, 2022.

[60] G. Underwood, L. Jebbett, and K. Roberts. 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):165–182, 2004.

[61] T. Van Gog, F. Paas, J. J. Van Merriënboer, and P. Witte. Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied*, 11(4):237, 2005.

[62] J. L. van Strien, Y. Kammerer, S. Brand-Gruwel, and H. P. Boshuizen. How attitude strength biases information processing and evaluation on the web. *Computers in Human Behavior*, 60:245–252, 2016.

[63] K. E. Weick, K. M. Sutcliffe, and D. Obstfeld. Organizing and the process of sensemaking. *Organization science*, 16(4):409–421, 2005.

[64] L. J. Wells, S. M. Gillespie, and P. Rotshtein. Identification of emotional facial expressions: Effects of expression, intensity, and sex on eye gaze. *PloS one*, 11(12):e0168307, 2016.

[65] R. W. White, J. M. Jose, and I. Ruthven. An implicit feedback approach for interactive information retrieval. *Information processing & management*, 42(1):166–190, 2006.

[66] S. J. White, M. Hirotani, and S. P. Liversedge. Eye movement behaviour during reading of japanese sentences: Effects of word length and visual complexity. *Reading and Writing*, 25:981–1006, 2012.

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

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