SenseSeek Dataset: Multimodal Sensing to Study Information Seeking Behaviors

KAIXIN JI, RMIT University, Australia
DANULA HETTIACHCHI, RMIT University, Australia
FALK SCHOLER, RMIT University, Australia
FLORA D. SALIM, The University of New South Wales, Australia
DAMIANO SPINA, RMIT University, Australia

Information processing tasks involve complex cognitive mechanisms that are shaped by various factors, including individual goals, prior experience, and system environments. Understanding such behaviors requires a sophisticated and personalized data capture of how one interacts with modern information systems (e.g., web search engines). Passive sensors, such as wearables, capturing physiological and behavioral data, have the potential to provide solutions in this context. This paper presents a novel dataset, *SenseSeek*, designed to evaluate the effectiveness of consumer-grade sensors in a complex information processing scenario: searching via systems (e.g., search engines), one of the common strategies users employ for information seeking. The *SenseSeek* dataset comprises data collected from 20 participants, 235 trials of the stimulated search process, 940 phases of stages in the search process, including the realization of Information Need (IN), Query Formulation (QF), Query Submission by Typing (QS-T) or Speaking (QS-S), and Relevance Judgment by Reading (RJ-R) or Listening (RJ-L). The data includes Electrodermal Activities (EDA), Electroencephalogram (EEG), PUPIL, GAZE, and MOTION data, which were captured using consumer-grade sensors. It also contains 258 features extracted from the sensor data, the gaze-annotated screen recordings, and task responses. We validate the usefulness of the dataset by providing baseline analyses on the impacts of different *cognitive intents* and *interaction modalities* on the sensor data, and *effectiveness* of the data in discriminating the search stages. To our knowledge, *SenseSeek* is the first dataset that characterizes multiple stages involved in information seeking with physiological signals collected from multiple sensors. We hope this dataset can serve as a reference for future research on information-seeking behaviors.

CCS Concepts: • **Information systems → Users and interactive retrieval**; • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: dataset; information seeking; information interaction; physiological signals; user studies

ACM Reference Format:
Kaixin Ji, Danula Hettiachchi, Falk Scholer, Flora D. Salim, and Damiano Spina. 2025. *SenseSeek* Dataset: Multimodal Sensing to Study Information Seeking Behaviors . *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 92 (September 2025), 29 pages. https://doi.org/10.1145/3749501

1 Introduction

Information is ubiquitous [15], and modern information access systems such as web search engines or intelligent assistants have evolved to become increasingly diverse, offering greater mobility and multiple ways of interaction.

Authors' Contact Information: Kaixin Ji, RMIT University, Melbourne, Australia, kaixin.ji@student.rmit.edu.au; Danula Hettiachchi, RMIT University, Melbourne, Australia, danula.hettiachchi@rmit.edu.au; Falk Scholer, RMIT University, Melbourne, Australia, falk.scholer@rmit.edu.au; Flora D. Salim, The University of New South Wales, Sydney, Australia, flora.salim@unsw.edu.au; Damiano Spina, RMIT University, Melbourne, Australia, damiano.spina@rmit.edu.au.

This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2025 Copyright held by the owner/author(s).
ACM 2474-9567/2025/9-ART92
https://doi.org/10.1145/3749501

In this context, wearable sensors that capture physiological and behavioral data have demonstrated remarkable capabilities in enhancing human-computer interaction and improving daily life applications. There is growing research interest in leveraging these wearable sensors to explore novel questions in information interaction, including the detection of user interest [2, 14] and engagement during video browsing [4, 22], recognition of cognitive preference [8, 47] and emotional responses [17] on social media content, and understanding various information needs [43].

Information only becomes meaningful when individuals extract and interpret it [15]. This interpretation process engages complex cognitive mechanisms, such as attention allocation, memory retrieval, reasoning, and decision-making [26, 49], shaped by various factors, including individual goals, prior experience, and system environments [15, 52]. Physiological data offers unique insights as it responds sensitively to different activities and variables, while physiological states themselves can also influence these activities. Given these intricate relationships between physiological data and contextual factors, careful examination of their interactions is essential [53]. Moreover, wearable sensors are becoming smaller and more compact, as seen in new devices like earphone-style EEGs that have fewer sensors [28, 66]. Although these smaller devices are more convenient to wear, researchers need to balance this convenience with how well they can collect accurate and reliable data.

Rationale for a new dataset. In this paper, we focus on a complex information processing scenario: information seeking, an iterative problem-solving process where users identify gaps in their knowledge and employ various strategies to address them [34]. Searching via systems (e.g., search engines) is one of the common strategies users employ, involving multiple downstream tasks (referred to as search stages in this paper) that each carry different cognitive intents. A simplified explanation of this process is as follows, summarized from [26, 34]. The user first recognizes a problem regarding their need for information and then reaches the system; this stage is termed *the realization of Information Need (IN)*. Next, they think about the query, *Query Formulation (QF)*, and submit, *Query Submission (QS)*. When they receive the results from the system, they evaluate both the information contents and the quality of the search session; this stage is termed *Relevance Judgment (RJ)*. Given that these stages engage different cognitive processes, they may be reflected in bodily responses, which leads to our first research question, **RQ1: To what extent do different cognitive intents across search stages influence users' physiological responses?** Moreover, the information systems have been diversified by providing different interaction modalities, i.e., Query Submission by Typing (QS-T) or by Speaking (QS-S) and Relevance Judgment by Reading (RJ-R) or by Listening (RJ-L). As these modalities engage different sensory and motor systems, we propose our second research question, **RQ2: How do different interaction modalities influence users' physiological responses?** On the other side, given the growing trend of employing physiological and behavioral data in information activities, especially in real-world settings (e.g., social media usage [17], everyday workspace [32, 45]), it is essential to understand the baseline performances of these measurements. This leads us to our third research question, **RQ3: How effectively can physiological and behavioral data detect and distinguish between specific search stages in information seeking processes?**

Therefore, we introduce the *SenseSeek* dataset, a comprehensive dataset comprising multimodal sensing data for studying information seeking behaviors. To the best of our knowledge, *SenseSeek* is the first dataset to analyze downstream searching tasks (i.e., search stages) using physiological and behavioral data collected using multiple consumer-grade sensors. It includes data from 20 participants who completed a controlled lab-based study simulating a mock-up information search scenario, divided into 6 search stages with 4 different interaction modalities. The user study content was carefully curated to ensure objectivity by minimizing the influence of subjective factors such as feelings, attitudes, and relevance. The data comprises a range of physiological and behavioral data, including Electrodermal Activity (EDA), Electroencephalogram (EEG), MOTION (including wrist and head motion) from wearable devices, and PUPIL and GAZE from a screen-mounted eye tracker. It includes

both raw and cleaned versions of the data, along with 258 extracted features. To provide contextual richness, it also contains the task materials, self-rated perceptions, and gaze-annotated screen recordings.

The contribution of this paper is 4-fold:

(1) Introducing the *SenseSeek* dataset, consisting of multiple sensor data (raw and cleaned) and extracted features, self-rated perceptions, gaze-annotated screen recordings from 20 participants performing mock-up information searching tasks;
(2) Understanding the physiological responses by the influence of cognitive intents and interaction modalities in a sequence of search stages by a highly controlled lab study;
(3) Assessing the effectiveness of physiological and behavioral data captured by consumer-grade sensors for monitoring information behavior;
(4) Proposing an experimental setup for data collection during information seeking can be replicated across diverse experimental settings, from naturalistic information search to conversational-style interaction, or multi-media information-access interaction.

2 Related Work

This section introduces the background about information seeking and passive sensing, and discusses the necessity of establishing a baseline of the effectiveness of sensor data in this context.

2.1 Human Information Interaction & Information Seeking

Introduced by Fidel [15], Human Information Interaction (HII) investigates how users interact with and make sense of information. HII is a broad and multidisciplinary field. Based on the research problem, HII has often been segmented and incorporated into specialized areas, such as Human-Computer Interaction (HCI), Information Retrieval (IR), or Information Behavior studies.

Information seeking is a sub-area in HII [15]. Information search is a process in which a user interacts with an information access system (e.g., a web search engine) to fulfill their need for information (i.e., information need) [34]. In the field of IR, this process is usually described as an iterative process which consists of a sequence of search stages: the realization of Information Need (IN), Query Formulation (QF), Query Submission (QS) and Relevance Judgment (RJ) [34]. Circumstances, goals, and motivations might drive user to seek particular information to address a gap in their knowledge or understanding (IN). In a lab study, a backstory describing the scenario is usually provided for this purpose. They then plan about what query can help obtain the desired results from the system (QF) and proceed to execute the plan (QS). Finally, they evaluate whether the results meet their information goals (RJ).

To enhance user experience and accurately meet user needs in various contexts, wearable sensors might have great potential to offer a portable and accessible solution [39].

2.2 Understanding Information Behaviors with Passive Sensing

Passive sensing, including the use of behavioral and physiological sensors, offers a unique way to study information behaviors by capturing real-time, naturalistic data. This approach provides deeper insights into how individuals engage with and process information in everyday contexts.

Eye-tracking data, such as eye movements, is commonly used in information behavior studies to visualize attention and assess user judgments on relevance [10], task goals [12], and interests [24]. Additionally, motion data, including wrist and head movements, is valuable for understanding emotional states and engagement. For instance, wrist movements have provided insights into learning engagement [16]. In a study about social media use, Gebhardt et al. [17] found that behavioral features have outperformed the physiological features in emotion detection.

Physiological data, including skin conductance, heart rate, pupil dilation, and brain activity, have been widely used to explore cognitive or affective states in information behavior studies. For example, Gupta et al. [20] used physiological data and eye activity for recognition of emotional memory recalls. Jimenez-Molina et al. [27] measured workload changes during web browsing with multiple physiological data. Their classification models using EEG data outperformed those excluding it. Moshfeghi et al. [50] have utilized brain activity to detect information need realization, while Ye et al. [64] explored EEG responses when individuals identify keywords linked to information needs. Moshfeghi and Jose [48] found that combining physiological data with task dwell time significantly improved the discrimination of search intentions. Wu et al. [62] demonstrated that EDA could predict online shopping satisfaction, while White and Ma [61] showed that heart rate features were closely linked to user interest on search pages. Edwards and Kelly [14] found that skin conductance was associated with interest, while heart rate indicated frustration during search tasks.

2.3 Dynamics of Information Interaction & Sensor Data

Information activity centers on the interaction between users and information [15]. Beyond the inherent characteristics of the information contents, various factors during user interactions with information access systems might also influence sensor data. These influences, which may affect the generalizability of results, remain insufficiently understudied.

First of all, the *cognitive intent*. As previously discussed, during information searching, users progress through the information stages with varying cognitive intents. Research found that these shifts in intent trigger diverse affective and cognitive processes [51], which may, in turn, influence physiological responses.

Secondly, the *interaction modalities*. To accommodate evolving user demands, the information access systems have been renovated into multi-media platforms offering diverse interaction methods. These include not only traditional screen-based interactions but also voice-based interactions and their combinations. Different interaction methods undergo distinct internal processes in humans, evoking varied psychological responses. For example, the single-channel EEG data [58] or physiological responses [56] are used to assess mental workloads across contexts with varying cognitive and motor demands. Iadarola et al. [23] found that EDA captured by the Empatica E4 wristband could differentiate the emotions in pleasant and unpleasant acoustic stimuli. It is yet understudied whether these responses impact the physiological data captured by consumer-level wearable sensors. In this paper, we focus on comparing different interaction modalities in submitting or receiving information. Specifically, submitting the query to the system by typing (QS-T) or by speaking (QS-S), and receiving the search results from the system in text format for reading (RJ-R) or audio format for listening (RJ-L).

Furthermore, physiological and behavioral data have their unique advantages and limitations. Firstly, since different information stages require various physical actions, behavioral data may exhibit more observable differences among search stages compared to physiological data. However, physiological data can provide insights into users' internal states such as stress levels, cognitive load, and emotional responses, which behavioral data alone cannot capture. Secondly, physiological data have varying response times to stimuli [55]. This variability is particularly relevant in the context of information searching, where the duration of search stages can range from a few seconds to over a minute, unlike tasks in other studies with consistent engagement durations, e.g., [17]. There is a growing interest in incorporating these sensing data into advanced analytical applications (e.g., building predictive models), but what is less clear is how these variants, instead of the experimental stimulus, impact model performance.

2.4 Multimodal Sensing Datasets

A body of recent research has established valuable datasets for studying physiological and wearable sensor data across diverse contexts. The DEAP dataset [30] laid the groundwork for emotion analysis using physiological

signals, while newer contributions have expanded into specialized domains. The Emognition dataset Saganowski et al. [54] and Banik et al. [4] focused on continuous emotion annotation during video consumption, and EEG-SVRec [68] for short-video recommendation. Bota et al. [9] focused on the group emotion experiences during movie sessions in a more naturalistic setting. VREED [59] incorporated virtual reality contexts with eye tracking, and LAUREATE [36] targeted memory augmentation applications. WEAR [7] focused on outdoor sports activity recognition. Zhang et al. [67] worked on collaborative cognitive experience using EEG. These datasets demonstrate the evolution toward multimodal sensing approaches in naturalistic settings, with a focus on solving practical problems. It also highlights a persistent challenge in the need for standardized annotation.

3 Data Collection Methodology

This section introduces the details of the data collection procedure.

3.1 Participants

We collected the data from 20 participants (12 male, 8 female). A summary of their demographic attributes is given in Table 1. All participants were required to have at least professional working proficiency in English to ensure a minimum of additional effort was involved when completing the study, which was presented in English.

The study received ethics approval from RMIT University, and the participants provided written consent prior to the experiment. A total of 29 participants were recruited. After the experiment, the participants were contacted again for follow-up consent to release the raw data, and 20 out of 29 participants provided consent.

Table 1. Demographic of participants in the dataset.

Gender		**Age**		**English Proficiency**		**Right-Handed**		**Wear-Glasses**	
Male	12	18–24 years old	5	Professional	4	Y	18	Y	12
Female	8	25–34 years old	14	Full Professional	11	N	2	N	8
		35–44 years old	1	Native English	5				
Total Number of participants		20							

3.2 Material

The topics and the corresponding backstories (i.e., task scenarios) are selected from the TREC2002-InformationNeed dataset [46]. This dataset contains backstories corresponding to different topics from the TREC2002-4 topic set, and categorizes them into three labels in terms of cognitive complexity. We use the topics from the Understanding category, which requires the participants to find information and gain some understanding of the topics. After removing the topics related to crises, wars, conspiracy, or politics, which might trigger subjective feelings, we select 12 topics and their corresponding backstories.

We reviewed and adjusted the backstories and the search results based on Flesch-Kincaid readability scores and word counts. All search results were converted into an audio format with the Google text-to-speech API for the listening tasks. As a result, the backstories have an average of 40±1 words, the search results have an average 148±3 word count with a readability level of 12, and the corresponding audios average 64±4 seconds.

3.3 Search Task

The search task is inspired and modified based on the experiment in Moshfeghi and Pollick [49], which contains the following search stages: Information Need (IN), Query Formulation (QF), Query Submission (QS), and Relevance

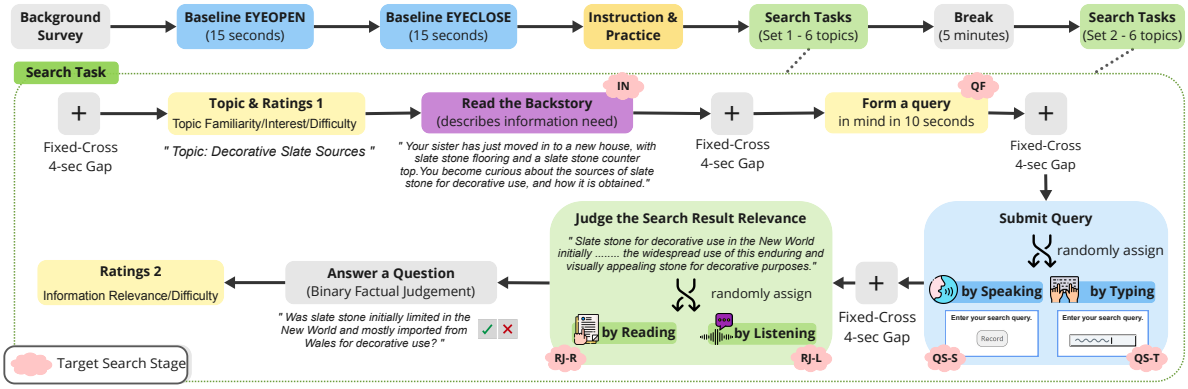


Fig. 1. User study procedure. The top section outlines the overall procedure, while the bottom details the search task procedure. Topic order and interaction modalities are randomized. Example contents for one topic are shown in *italic* font. The target stages are labeled with their abbreviations: Information Need (IN), Query Formulation (QF), Query Submission by Typing (QS-T) or Speaking (QS-S), and Relevance Judgment by Reading (RJ-R) or Listening (RJ-L).

Judgment (RJ). Note that as we are interested in the different responses among search stages rather than particular evaluation criteria, only one pre-defined highly-relevant search result is provided at RJ for each topic.

Each task starts with a 4-second blank, where the participants are asked to look at the fixed cross in the middle of the screen. A topic is shown afterward. And the participants need to rate their interests, familiarity, and perceived difficulty regarding the topic, with a 5-point Likert scale (as listed in Appendix B). After that, a backstory that evokes their information need is presented. The participants are then given 10 seconds to form a search query in mind (QF). Next, they follow an instruction to submit the query by speaking (QS-S) or typing (QS-T). Then the participants receive the search result by either a text paragraph to read (RJ-R) or an audio to listen (RJ-L). In the end, the participants need to answer a binary judgment question to ensure engagement and rate perceived relevance and difficulty in understanding the search result, with a 5-point Likert scale (as listed in Appendix B). Apart from QF, there was no time constraint on completing the task.

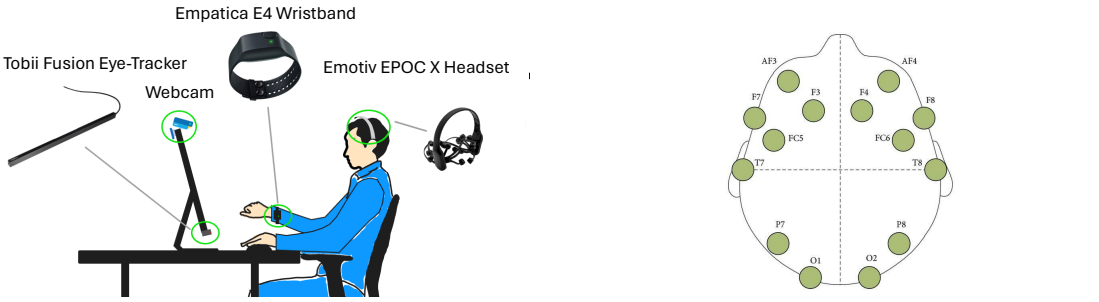
To avoid the physiological responses overlapping between the stages, the 4-second gap is also provided before the three interested stages, i.e., QF, QS, RJ. The sequences of topics and the combination of interaction modalities are randomized.

3.4 Procedure

After calibrating all sensors, the participants need to answer a background survey. The survey asked questions that might impact the data quality, e.g., handiness, caffeine intake, sleep hours, and spectacular. Next, the participants are instructed to have a 15-second eyes-open (EYEOPEN) and 15-second eyes-close (EYECLOSE) section to collect the baseline, followed by a training section containing the instruction and two practice tasks. Then the participants proceed to the main search tasks (12 tasks in total). There is a 5-minute break provided in the middle. After completing the whole experiment, participants are asked some questions regarding their experience during the experiment to reveal some undetected activities. For example, 'do you feel the 10 seconds given for query formulation is unnecessary or useful?'.

The details in the search tasks, e.g, the time given to query formulation and the total number of tasks, were pilot-tested with 4 additional participants.

3.5 Apparatus and Setup



(a) Experimental setup. (b) The EEG electrode locations (adapted from [19]).

Fig. 2. Experimental setup and apparatus.

The list of apparatus and collected data is presented in Table 2, and the setup is presented in Figure 2. The study is conducted in an illuminated room, which includes a desktop PC mounted with an eye-tracker and a web camera. All participants are provided with a standard right-handed computer mouse to complete the tasks. The participant sits in front of the computer and wears the wristband on the left hand. The instructor cleans the electrodes and the participant's skin on the inner and outer wrist with alcohol wipes [3], and helps the participants to wear the headset and adjusts the position of the electrodes. The experiment is created on the Qualtrics[1] platform.

Table 2. List of sensor data collected in the dataset.

Description	Data	Sampling Rate	Channels	Comments	Apparatus
Electrodermal Activity	EDA	4Hz	1	Unit in microsiemens (uS)	Empatica E4 Wristband[2]
Wrist Motion - accelerometer	MOTION	32Hz	3 (x, y, z)	Data in [-2g, 2g] range	
Head Motion - accelerometer	MOTION	128Hz	3 (x, y, z)	Data in [-16g, 16g] range	Emotive EPOC X Headset[3]
- magnetometer		128Hz	3 (x, y, z)	Data in [-2000dps, 2000dps] range	
- quaternions		128Hz	1	Data in [-1g, 1g] range	
Electroencephalogram	EEG	128Hz	14 (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4)	Unit in microvolt (uV)	
Gaze Movement	GAZE	60Hz	2 (x, y) x 2 (left, right)	The coordinate of gaze point in screen	Tobii Fusion Eye-tracker[4]
Pupil Diameter	PUPIL	60Hz	1 x 2 (left, right)	Unit in millimeters (mm)	

4 *SenseSeek* Dataset Description

This section describes the dataset and the data cleaning and processing steps for the validation.

4.1 Overview

This dataset comprises data collected from 20 participants, each completing 12 search tasks, for a total of 240 trials. We excluded 5 trials due to unexpected environmental disturbances and equipment failures, leaving 235 trials in the dataset. Additionally, eye-tracking data for 2 participants were not recorded because of device errors. The summary of the dataset is presented in Table 4.

[1] https://www.qualtrics.com/about/

Table 3. Table of Glossaries and Abbreviations. *Note that these are specialized terminologies used in the field of information behavior studies [29].

Stage Name	Abbreviation	Description
Baseline EYEOPEN	EYEOPEN	A baseline task to relax and look at the screen with eyes open for 15 seconds.
Baseline EYECLOSE	EYECLOSE	A baseline task to relax and close the eyes for 15 seconds.
The realization of Information Need*	IN	Read the given backstory* (i.e., scenario) to understand the context and the need for information.
Query Formulation*	QF	Think about a search query for getting the desired information from the system.
Query Submission* by Typing	QS-T	Type the search query and submit to the system.
Query Submission* by Speaking	QS-S	Speak the search query and submit to the system.
Relevance Judgment* by Reading	RJ-R	Receive a (highly relevant in this study) search result in text format to read, and judge the relevance to the information need.
Relevance Judgment* by Listening	RJ-L	Receive a search result in audio format to listen to, and judge the relevance to the information need.

Table 4. Summary of the *SenseSeek* dataset.

Recorded Contextual Data										
Number of participants		20				**Number of Topics**				12
Number of Gaze-annotated Screen Recordings		20				**Number of Task Responses**				235
Recorded Sensing Data										
Search Stages		**Baseline**		**IN**	**QF**	**QS**		**RJ**		**Total** (+ baseline)
		EYEOPEN	**EYECLOSE**			**QS-T**	**QS-S**	**RJ-R**	**RJ-L**	
Duration in seconds (sd)		15.0 (0)	15.0 (0)	23.4 (14.1)	10.0 (0)	17.7 (10.1)	6.8 (2.5)	49.2 (19.8)	66.0 (4.4)	-
Amount of Data Instance	**EEG**	20	20	235	235	119	116	118	117	**940 (+40)**
	MOTION (head)	20	20	235	235	119	116	118	117	**940 (+40)**
	MOTION (wrist)	20	20	235	235	119	116	118	117	**940 (+40)**
	EDA	20	20	235	235	119	116	118	117	**940 (+40)**
	GAZE	18	NA	211	211	107	101	106	105	**859 (+18)**
	PUPIL	18	NA	211	211	107	101	106	105	**859 (+18)**

4.2 Sensor Synchronization

The event timestamps of the task behaviors were recorded with the JavaScript library on Qualtrics. To synchronize all sensors, we convert the timestamps into ISO 8601 time format with milliseconds. The signals were aligned and segmented according to the event timestamps recorded. An example of apparatus timeline and synchronization is presented in Figure 3.

4.3 Data Pre-Processing

All data was processed with the Python programming language. EEG data were processed using the MNE library [5]. EDA and wrist motion data obtained from the wristband were processed using the NeuroKit2 [40] library. PUPIL data was cleaned with the *PyPlr* [41] library. Data collected during the break section were excluded from pre-processing.

EEG. Following similar procedures to Martínez-Santiago et al. [42], Ye et al. [63], Gwizdka et al. [21], EEG data were first re-referenced with the common average, denoised with a Butterworth filter (1 – 40Hz), and removed the signal mean for the middle 90% percentile. Next, to remove the artifacts (e.g., blinking), the data was further cleaned and interpolated with the Autoreject [25] package and Independent Component Analysis (ICA) combined

[5] https://mne.tools/stable/i

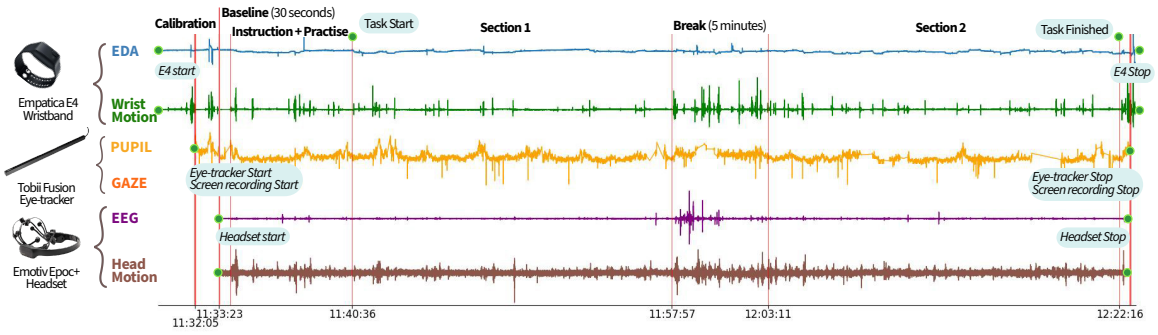


Fig. 3. Example of Apparatus Timeline and Synchronization in the Data Collection.

with ICLabel [38]. The power spectral density (PSD) of each EEG channel was then calculated using Welch's method and hamming window and normalized [33, 37, 58], for 4 frequency bands, *Theta* (4–8Hz), *Alpha* (8–13 Hz), *Beta* (13–30Hz) and *LowGamma* (25–40Hz) [21].

EDA. Following a similar procedure as by Gao et al. [16], Gebhardt et al. [17], EDA was first cleaned with a rolling median filter with a 3-second window (12 data points) [3], and then standardized using the z-score. Next, the convex optimization *cvxEDA* method [18] was applied to decompose the tonic value, i.e., the Skin Conductance Level (SCL), and the phasic value, i.e., the Skin Conductance Response (SCR).

MOTION. No pre-processing was done on the head motion data. For the wrist motion, we followed the procedure of Gao et al. [16]. We first calculated the magnitude from the 3 coordinates (x, y, z), $|a| = \sqrt{x^2 + y^2 + z^2}$, then cleaned with a rolling median filter with 0.2-second window (6 data points).

GAZE & PUPIL. The invalid gaze samples identified by the eye-tracker were removed. We used Tukey's method on the first derivatives to remove the noise caused by blinks from the left and right pupils, respectively. The samples with first derivatives lower than or higher than 3 standard deviations are removed. Then, the cleaned data of both sides were combined by taking the arithmetic mean, and linear interpolation was applied to fill in the blink gaps. Finally, a zero-phase Butterworth filter (4Hz, 3^{rd}) was applied to remove outliers [41]. We also calculated Relative Pupil Dilation (RPD), the relative changes of current pupil diameter compared to a baseline value (EYEOPEN), following the formula by Gwizdka et al. [21]: $RPD_t^i = (P_t - P^i_{baseline})/P^i_{baseline}$ where t is time, i is participant, and baseline is the average pupil diameter during EYEOPEN.

4.4 Features Extraction

We extracted a total of 258 features from both *physiological* (EDA, EEG, and PUPIL) and *behavioral* (MOTION and GAZE) data, following [5, 6, 8, 13, 16, 21]. The features are listed in Table 5. More details on the extraction are described in the Appendix A.1.

Normalization. Same as in Gebhardt et al. [17], the normalization was done by per participant. The physiological features were normalized by subtracting the value in the baseline period (i.e., EYEOPEN). The behavioral features were normalized by subtracting the mean values overall for the whole experiment and then dividing by the standard deviation.

Table 5. List of total 258 extracted features. Skin Conductance Level (SCL), Skin Conductance Response (SCR), standard deviation (sd). Relative Pupil Dilation (RPD). *Four EEG frequency band: *Theta* (4–8 Hz), *Alpha* (8–13 Hz), *Beta*(13–25 Hz) and *LowGamma* (25–40 Hz) bands [21]. **Four channel groups: frontal-left (AF3, F7, F3, FC5), frontal-right (FC6, F4, F8, AF4), back-left (P7, O1), and back-right (P8, O2) [21]. ***statistics (5): five statistical description features include mean, max, min, sd, range.

Type	Group	Category	Features	N. Features
Physiological	**EDA(31)**	EDA	statistics (5), median, variance, skewness, kurtosis statistics (5) of 1st/2nd derivative of EDA	9 10 (5x2)
		SCL	statistics (5), median, mean amplitude	7
		SCR	statistics (5)	5
	EEG(180)	14 channels	mean, sd, skewness, kurtosis, curve length, zero crossings, number of peaks, wavelet entropy	112 (8x14)
		4 frequency bands*	the logarithms of the spectral power	56 (4x14)
		4 bands of 4 channel groups**	Inter-/Intra-hemispheric asymmetry power ratios between/within right-left groups	12 (4+8)
	PUPIL(25)	Pupil Diameter	statistics (5), median LHIPA index [13]	6 1
		RPD	statistics (5), 5 quarterlies (10, 25, 50, 75, 90) statistics (5), 3 quarterlies (25, 50, 75) of 1st derivative of RPD	10 8
Behavioral	**MOTION(13)**	Wrist Motion	statistics (5) of magnitude of acceleration	5
		Head Motion	mean of *x, y, z* coordinates of angular speed/acceleration mean, energy of magnitude of acceleration	6 (3x2) 2
	GAZE(6)	fixation/saccade	number of occurrence, total duration (normalized), mean duration of each occurrence	6 (3x2)

5 Validation

This paper aims to validate the *SenseSeek* dataset by understanding how different *cognitive intents* **(RQ1)** and *interaction modalities* **(RQ2)** across search stages influence physiological responses, and investigating *the effectiveness of using physiological and behavioral data to distinguish search stages* **(RQ3)**. This section presents our validation methodology and findings organized according to these research questions.

Before conducting any analysis, we first reviewed the self-reported task perceptions to ensure the experimental conditions met our anticipation (Section 5.1). To answer RQ1 and RQ2, we conducted the exploratory analysis with statistical tests in a within-subject setting. We first conducted non-parametric Friedman tests to compare across EYEOPEN baseline, and all search stages (IN, QF, QS-T, QS-S, RJ-R, RJ-L) for overall impacts (Section 5.2).

We then performed post-hoc analysis to compare stages with similar characteristics: EYEOPEN, IN, and RJ-R (different cognitive intents, same interaction modalities, RQ1, Section 5.3), and QS-T vs. QS-S, and RJ-R vs. RJ-L (same cognitive intents, different interaction modalities, RQ2, Section 5.4). The multiple comparisons were accounted for with Bonferroni correction. For RQ3, we conducted a classification analysis by developing machine learning models using features from individual and multiple sensor groups, and comparing the performances in discriminating all search stages.

Table 6. Non-parametric Friedman Statistical Testing Results (F-value) comparing all search stages and EYEOPEN for the 4 frequency bands at 14 EEG nodes. $p < .05^{*}$, $< .01^{**}$, $< .001^{***}$.

Band\Node	AF3	AF4	FC5	FC6	F3	F4	F7	F8	T7	T8	P7	P8	O1	O2
Theta	1.850	3.052**	0.737	5.387***	2.366*	7.667***	2.569*	2.648*	4.070**	3.034**	8.366***	6.261***	7.441***	11.601***
Alpha	8.687***	11.116***	15.794***	13.900***	13.870***	29.763***	11.948***	17.068***	23.055***	29.342***	32.476***	23.458***	28.649***	15.345***
Beta	10.223***	10.621***	12.942***	8.465***	7.047***	9.358***	15.278***	14.554***	8.508***	7.991***	12.368***	12.479***	11.850***	9.208***
LowGamma	5.848***	8.949***	9.945***	6.160***	5.062***	4.535***	9.850***	8.254***	2.745*	4.750***	1.411	3.720**	5.116***	10.531***

5.1 Task Completeness & Manipulation Check

Task Perceptions. As the confounding variables, for instance, perceived difficulty [42], interest [2, 8], or search result relevance [63], could impact the physiological responses, the stimulus used in this study has been taken care to minimize the impact from these. In particular, we anticipate that i) all the topics should be easy to understand ($\leq$ 3.0) and not familiar ($\leq$ 3.0), ii) all the search results (in RJ) provided are easy to understand ($\leq$ 3.0), and relatively relevant ($\geq$ 3.0) to their submitted query.

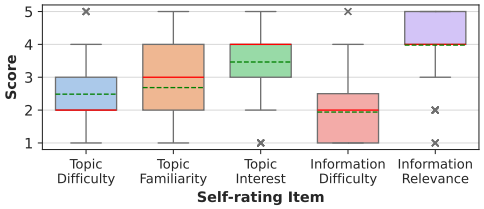


Fig. 4. Distribution of Self-rated Task Perceptions on Perceived Difficulty, Familiarity, Interest in the Topic, and Perceived Difficulty and Relevance on the received Information.

As summarized in Figure 4, the participants have rated Topic Difficulty as 2.5 (SD: 1.1), Topic Familiarity as 2.7 (SD: 1.2), Topic Interest is 3.5 (SD: 1.1). For the search results, the participants have rated Information Difficulty as 1.9 (SD: 1.0) and Information Relevance as 4.0 (SD: 1.1). Overall, the task perceptions have met our anticipation, all data have been processed further for analysis.

Task Durations. As reported in Table 4, the participants take an average of 23.4 (SD: 14.1) seconds to complete reading the backstories at IN, and 10 seconds (as pre-defined) to think about a query at QF. The interaction modality affects completion time. QS-S takes an average of 6.8 (SD: 2.5) seconds, faster than QS-T, which takes an average of 17.7 (SD: 10.1) seconds. RJ-L takes an average of 66.0 (SD: 4.4) seconds, longer than RJ-R, which takes an average of 49.2 (SD: 19.8) seconds.

5.2 Understanding Physiological Responses at Different Stages of Information Search

Brain Activity Responses. As presented in Figure 5, the *Alpha* bands show the most prominent differences, with the highest F-values concentrated in the left-parietal, i.e., $P7$ ($F[6, 112] = 32.476, p < .001$), left-occipital regions, i.e., $O1$ ($F[6, 112] = 28.649, p < .001$), and right-frontal regions, i.e., $F4$ ($F[6, 112] = 29.763, p < .001$). The differences observed in *Alpha* band likely reflect task-related variations in attention allocation and spatial processing [44, 47], e.g., for filtering out the irrelevant information[63], across different information stages. The *Beta* bands also show significant differences across all regions. As *Beta* activity is commonly associated with active thinking and problem solving [35, 65], the results suggest that these processes might be similarly and significantly engaged across all conditions. Lastly, the *Theta* and *LowGamma* bands have generally lower

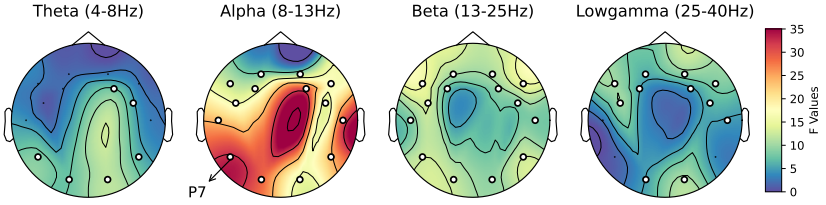


Fig. 5. Topography which shows the significance of difference (F-value from Friedman Test) between brain response in the baseline, EYEOPEN and the 6 search stages (IN, QF, QS-T, QS-S, RJ-R, RJ-L). Highlighted channels indicate the differences are significant at $p < .001$. Greenhouse-Geisser correction is applied to adjust for the violations of sphericity.

Table 7. Pairwise Wilcoxon Test Results for the EEG P7 Alpha bands. $p < .01^{**}$, $< .001^{***}$. One-step Bonferroni Correction applied for multiple comparisons.

A	**EYEOPEN**						**IN**				
B	**IN**	**QF**	**QS-S**	**QS-T**	**RJ-R**	**RJ-L**	**QF**	**QS-S**	**QS-T**	**RJ-R**	**RJ-L**
W	45	3	39	65	32	4	7	14	8	65	9
p	.503	**.000*****	.254	1	.102	**.000*****	**.001****	**.004****	**.001****	1	**.001****

A	**QF**				**QS-S**			**QS-T**		**RJ-R**
B	**QS-S**	**QS-T**	**RJ-R**	**RJ-L**	**QS-T**	**RJ-R**	**RJ-L**	**RJ-R**	**RJ-L**	**RJ-L**
W	0	1	16	103	65	10	0	10	1	16
p	**.000*****	**.000*****	**.007****	1	1	**.002****	**.000*****	**.002****	**.000*****	**.007****

F-values. This might indicate that both *Theta* and *LowGamma* are less sensitive to the specific differences among the search stages compared to the other bands.

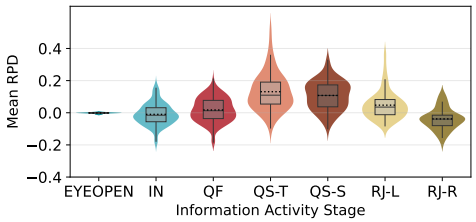


Fig. 6. Violin plot represents the distribution of mean Relative Pupil Dilation (RPD) across the baseline EYEOPEN and the 6 search stages (IN, QF, QS-T, QS-S, RJ-R, RJ-L).

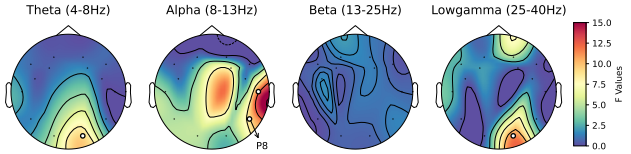
Pupil Dilation. In a constantly illuminated condition, Relative Pupil Dilation (RPD) can indicate attention and exerted cognitive load [21]. As demonstrated in Figure 6, both modalities for QS show larger RPD compared to the other stages. The Friedman test revealed significant difference ($F[6, 100] = 37.962, p < .001$) among search stages (refer to Table 8). And the post-hoc test reveals that both QS-S and QS-T are significantly different from EYEOPEN and all other stages. In particular, $W(\text{EYEOPEN, QS-S}) = 4, p < .01$, $W(\text{EYEOPEN, QS-T}) = 1, p < .001$, $W(\text{IN, QS-S}) = 0, p < .001$, $W(\text{IN, QS-T}) = 0, p < .001$, $W(\text{QF, QS-S}) = 0, p < .001$, $W(\text{QF, QS-T}) = 3, p < .01$, $W(\text{RJ-R, QS-S}) = 0, p < .001$, $W(\text{RJ-R, QS-T}) = 0, p < .001$, $W(\text{RJ-L, QS-T}) = 19, p < .05$. This result might indicate the demanding efforts for formulating and expressing queries in text or speech to the external system [57].

Table 8. Friedman Statistical Testing Results comparing among all search stages and EYEOPEN. $p < .001^{***}$.

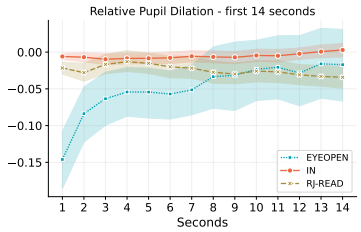
Data Type	Feature	F	p
EDA [6,112]	Mean Mixed EDA	1.321	0.254
	Mean SCL	1.244	0.290
	Mean SCR	0.259	**0.000*****
PUPIL [6, 100]	Mean RPD	37.962	**0.000*****

Peripheral Responses. Statistical testing results in Table 8 found no significant difference in the mixed or tonic EDA values (SCL) . Yet, a significant difference is revealed in phasic values, SCR ($F[6, 112] = 5.259, p < .001$).

5.3 The influence of Cognitive Intents: Comparing Eye-Driven Activities



(a) Friedman test results of EEG *Alpha* band. Highlighted channels indicate the differences are significant at $p < .001$.



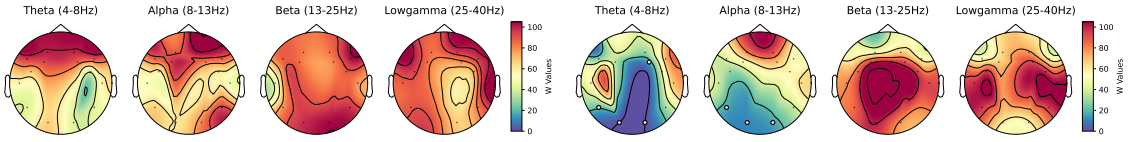
(b) Changes in Relative Pupil Diameter (RPD) during the first 15 seconds following stage onset. The values are subtracted from the value at the first second.

Fig. 7. The physiological responses over 3 eye-driven activities with different intentions, EYEOPEN, IN, and RJ-R.

The stages, EYEOPEN, IN, and RJ-R, require visual engagement but vary in cognitive demands and intentionality. As demonstrated in Figure 7a, the Friedman test reveals significant differences in the temporal-parietal-occipital regions on the right hemisphere among the 3 stages. This suggests that cognitive demands or the focus of attention differ primarily in these areas. In particular, the post-hoc Wilcoxon test reveals that the *Alpha* band during EYEOPEN is significantly different than during IN at T8, $W(\text{EYEOPEN, IN}) = 9, p < .01$. When comparing to RJ-R, the difference found at T8, $W(\text{EYEOPEN, RJ-R}) = 11, p < .01$, and P8, $W(\text{EYEOPEN, RJ-R}) = 26, p < .05$. Besides, the *Theta* bands also found significant differences at occipital regions. More specifically, *Theta* at O1 differs between EYEOPEN and RJ-R, $W(\text{EYEOPEN, RJ-R}) = 27, p < .05$, and at O2 differs between IN and RJ-R, $W(\text{IN, RJ-R}) = 9, p < .01$.

The mean RPD at EYEOPEN does not significantly differ from IN, $W(\text{EYEOPEN, IN}) = 63, p > .05$, or RJ-R $W(\text{EYEOPEN, RJ-R}) = 35, p > .05$. But it is significantly different between IN and RJ-R, $W(\text{IN, RJ-R}) = 1, p < .001$. Figure 7b illustrates the changes in pupillary responses during the first 15 seconds (align with the shortest stage, EYEOPEN), normalized to the average value of the onset second. At EYEOPEN, as participants stared at the screen, their pupils dilated overall, reflecting heightened attention and cognitive engagement. During IN, the pupillary responses remained relatively stable with a slight increase. This suggests the participants might sustain attention or consistent cognitive processing to comprehend and recognize their task. In contrast, at RJ-R, the pupil diameters generally contracted with a convex response at the beginning. This might imply a brief period of heightened attention or arousal, possibly due to curiosity or interest when the search result first appeared, followed by a gradual decline.

5.4 The Influence of Interaction Modalities



(a) QS-S v.s. QS-T (b) RJ-R v.s. RJ-L

Fig. 8. Topography which shows the significance of difference (W-value) between brain responses between interaction modalities. Bonferroni Correction applied. Highlighted channels indicate the differences are significant at $p < .05$.

As presented in Figure 8a, no significant difference is found between QS-T and QS-S for query submission. And the RPD shows no significant difference, $W(\text{QS-S}, \text{QS-T}) = 21, p > .05$.

Comparing between RJ-R and RJ-L, the differences are observed mainly at the parietal-occipital regions (i.e., P7, O1, O2) for both *Theta* and *Alpha* bands (refer to Figure 8b): $W(\text{P7}, \text{Theta}) = 22, p < .05$, $W(\text{P7}, \text{Alpha}) = 16, p < .01$, $W(\text{O1}, \text{Theta}) = 0, p < .001$, $W(\text{O1}, \text{Alpha}) = 11, p < .01$, $W(\text{O2}, \text{Theta}) = 0, p < .001$, $W(\text{O2}, \text{Alpha}) = 16, p < .01$. There is also a significant difference observed at the frontal region for *Theta* band, $W(\text{F4}) = 15, p < .01$. Besides, the RPD shows a significant difference, $W(\text{RJ-R}, \text{RJ-L}) = 1, p < .001$.

5.5 Classification Performance of the Sensor Data

5.5.1 Experimental Setting. The objective is to classify all search stages in considering interaction modalities, IN, QF, QS-S, QS-T, RJ-R, and RJ-L. We categorized the sensor data into 2 types – physiological and behavioral data – and 5 groups – including EDA, EEG, PUPIL, GAZE, and MOTION. MOTION combines the wrist and head motion.

Classifiers. The classification experiment was developed with the Scikit-learn Python Library. The support vector machine (SVM) with an RBF kernel was used as the base model [8, 58]. We have built two types of models, one with only features from every single type of sensor data (results refer to Section 5.5.2), the other was a late-decision-fusion model which combines two or more trained single-sensor models (results refer to Section 5.5.3). We ensemble the model with a soft voting classifier with equal weights assigned to the base models [30]. The soft voting classifier calculates the argmax of the sums of the predicted probabilities computed by the base models.

Dimension Reduction & Hyper-Parameter Tuning. Studies involving multiple physiological data often result in a large number of extracted features [21, 30], necessitating dimension reduction for the subsequent process, e.g., model training. We implemented with GridSearchCV from the Scikit-learn library, and tuned the model with different dimension reduction techniques, including Linear Discriminant Analysis, Neighborhood Components Analysis, and Select-K-Best (using mutual information or F-statistic), and parameters including the regularization parameter (C), kernel coefficient (gamma), and polynomial degree.

Evaluation. We conducted Leave-One-participant-Out (LOO) cross-validation and reported accuracy and F1-macro metrics for the evaluation. LOO cross-validation assesses the performance of a model by iteratively training on data from all participants except one, which is reserved for testing. In this project, which involves data from 20 participants, each model was evaluated 20 times, with the data from 19 participants used for training and the data from 1 participant used for testing each time[6]. Accuracy reflects the overall correctness of a model's predictions, while the F1-macro score addresses class imbalance by averaging performance equally across all classes, offering

[6]Note that as 2 participants are missing the eye-tracking data, the models involving PUPIL or GAZE were evaluated 18 times.

Table 9. Classification Performance on SVM with RBF kernel of single sensor groups using features extracted from different time windows. The evaluation metrics are Accuracy and F1-macro (with standard deviation).

Sensor Group	**EDA**		**EEG**		**PUPIL**		**MOTION**		**GAZE**	
Feature Window	Accuracy	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro
1s–0s overlap	0.273 (0.065)	0.170 (0.067)	0.324 (0.092)	0.246 (0.116)	0.549 (0.145)	0.530 (0.153)	0.452 (0.114)	0.409 (0.098)	0.295 (0.124)	0.191 (0.139)
2s–1s overlap	0.273 (0.065)	0.170 (0.067)	0.324 (0.092)	0.246 (0.116)	0.549 (0.145)	0.530 (0.153)	0.452 (0.114)	0.409 (0.098)	0.295 (0.124)	0.191 (0.139)
4s–2s overlap	0.282 (0.067)	0.146 (0.067)	0.328 (0.118)	0.252 (0.128)	0.585 (0.151)	0.574 (0.152)	0.506 (0.077)	0.424 (0.071)	0.403 (0.104)	0.299 (0.126)

a balanced evaluation of model effectiveness. Higher scores for both metrics indicate better model performance. We employed the random guess baseline, which is around 17%.

5.5.2 Single-Sensor Performance. As presented in Table 9 and Figure 9, PUPIL consistently demonstrated the best performance, while EDA was the only model to perform below baseline on F1 score. When comparing window feature settings, we observed that 1–0 and 2–1 windows produced identical performance metrics across all sensor groups, suggesting minimal impact of this particular window size variation. However, implementing the 4–2 window features yielded variable effects across sensor groups. GAZE showed the most substantial improvement with a 10.8% increase in accuracy and 10.0% increase in F1. PUPIL and MOTION demonstrated moderate enhancements (4.4% increase in both metrics for PUPIL; 5.4% accuracy and 2.5% F1 increase for MOTION), while EEG showed minimal improvement (1.1% increase in accuracy, 0.6% increase in F1). Interestingly, EDA exhibited a decrease of 3.6% in F1 score.

Figure 10 reveals distinct classification patterns across models. The EDA model showed poor discriminative capability with predictions skewed toward IN and QS-T. While EEG provided more balanced predictions, it still favored QS-T and RJ-L. Only the PUPIL model produced well-balanced and accurate predictions across all classes, while GAZE and MOTION models exhibited biases similar to EDA. Notably, when comparing the eye-driven activities IN and RJ-R (which involve different cognitive intentions), only MOTION and PUPIL models successfully differentiated between them. This suggests that similar gaze movement patterns in reading activities may confound the GAZE model, while the wristband may lack sufficient sensitivity to capture subtle EDA variations during these activities.

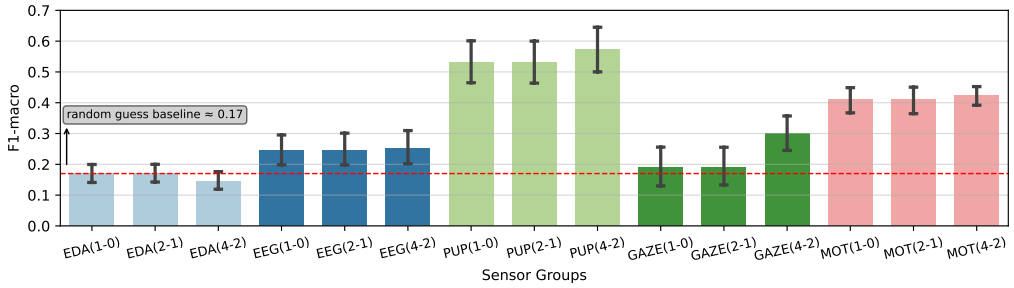


Fig. 9. The performance in F1-macro metrics of the SVM models for single sensor groups and different time-window features. The sensor groups are EDA, EEG, PUPIL, GAZE, and MOTION (from left to right). Error bars present 95% confidence intervals.

5.5.3 Decision-Fusion Performance. To build the fusion models, we used the features extracted from 4-second windows with 2-second overlap.

As shown in Table 10, all fusion models surpassed the baseline performance of 17%. Surprisingly, the highest performance was not achieved by incorporating all sensor groups. While the comprehensive model including

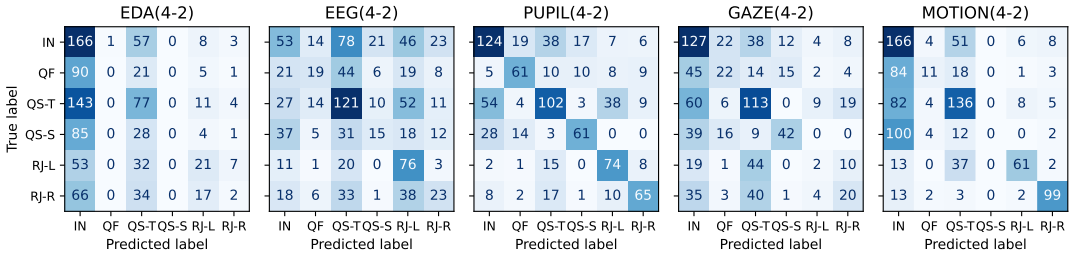


Fig. 10. Confusion Matrix of single sensor groups with 4 seconds with 2-second overlap time window features.

Table 10. Classification Performance of SVM with RBF kernel of fusion models and 4-2 window features. The highest and lowest in terms of F1 scores are highlighted in bold text. The evaluation metrics are Accuracy and F1-macro (with standard deviation).

	EEG	PUPIL	MOTION	GAZE	Accuracy	F1
EDA	✓	✓	✓	✓	0.637 (0.143)	0.609 (0.158)
	✓	✓	✓	–	0.570 (0.115)	0.525 (0.129)
	✓	✓	–	✓	0.552 (0.159)	0.519 (0.159)
	✓	–	✓	✓	0.569 (0.170)	0.537 (0.184)
	–	✓	✓	✓	0.647 (0.147)	0.614 (0.145)
	✓	✓	–	–	0.535 (0.141)	0.499 (0.147)
	✓	–	✓	–	0.475 (0.108)	0.405 (0.107)
	✓	–	–	✓	0.525 (0.135)	0.458 (0.148)
	–	✓	✓	–	0.598 (0.154)	0.580 (0.139)
	–	✓	–	✓	0.591 (0.146)	0.550 (0.150)
	–	–	✓	✓	0.543 (0.114)	0.486 (0.124)
	✓	–	–	–	**0.329 (0.090)**	**0.234 (0.099)**
	–	✓	–	–	0.520 (0.104)	0.502 (0.106)
	–	–	✓	–	0.484 (0.075)	0.471 (0.076)
	–	–	–	✓	0.425 (0.103)	0.328 (0.116)

	PUPIL	MOTION	GAZE	Accuracy	F1
EEG	✓	✓	✓	**0.671 (0.147)**	**0.645 (0.161)**
	✓	✓	–	0.551 (0.146)	0.508 (0.155)
	✓	–	✓	0.570 (0.154)	0.534 (0.164)
	–	✓	✓	0.556 (0.165)	0.531 (0.164)
	✓	–	–	0.555 (0.133)	0.533 (0.137)
	–	✓	–	0.452 (0.114)	0.396 (0.121)
	–	–	✓	0.490 (0.124)	0.430 (0.123)

	MOTION	GAZE	Accuracy	F1
PUPIL	✓	✓	0.645 (0.153)	0.621 (0.146)
	✓	–	0.597 (0.111)	0.590 (0.119)
	–	✓	0.549 (0.192)	0.507 (0.184)

	GAZE	Accuracy	F1
MOTION	✓	0.529 (0.116)	0.484 (0.121)

all sensors yielded an F1 score of 60.9% (SD: 15.8%) and accuracy of 63.7% (SD: 14.3%), excluding EDA produced better results with an F1 score of 64.5% (SD: 16.1%) and accuracy of 67.1% (SD: 14.7%). Combinations including GAZE with other sensor groups demonstrated consistently strong performance. The MOTION+PUPIL+GAZE configuration achieved the second-highest F1 score of 62.1% (SD: 14.6%), while adding EDA to this combination slightly reduced performance to an F1 score of 61.4% (SD: 14.5%). EEG+GAZE+PUPIL and EEG+GAZE+MOTION combinations yielded comparable results with F1 scores of 53.4% (SD: 16.4%) and 53.1% (SD: 16.4%), respectively.

The weakest performing combinations involved EDA paired with either EEG or GAZE, resulting in F1 scores of 23.4% (SD: 9.9%) and 32.8% (SD: 11.6%) with corresponding accuracies of 32.9% (SD: 9%) and 42.5% (SD: 10.3%), respectively. These combinations proved substantially less effective for the classification task.

Notably, the model using only physiological groups (EEG+EDA+PUPIL) achieved comparable performance to the model using only behavioral groups (GAZE+MOTION), with F1 scores of 49.9% (SD: 14.7%) and 48.4% (SD: 12.1%), and accuracies of 53.5% (SD: 14.1%) and 52.9% (SD: 11.6%), respectively. This suggests that physiological and behavioral groups contribute useful but distinct information to the classification task.

6 Discussion & Limitation

This section broadly discusses our results and the implications and opportunities that it hints towards. It also identifies the limitations and discusses the ecological validity of the released dataset.

6.1 RQ1. The Influence of Cognitive Intents on the Physiological Responses

Our first research question investigates whether varying cognitive intents during different search stages influence the physiological responses. To address this, we performed statistical analyses comparing all search stages (refer to Section 5.2), and further analyses focusing on comparing EYEOPEN, IN, and RJ-R (refer to Section 5.3) – each representing eye-driven activities but with distinct intents.

Our results showed that when comparing all stages, the EEG *Alpha* band activity differs across search stages at all regions, with particularly the highest statistical values observed at the left-parietal (P7), right-frontal (F4) and left-occipital (O1) regions, reflecting shifts in attention due to the visual, audio or spatial processing demands, and in decision-making processing. The EEG *Beta* also revealed differences across all regions, suggesting sustained cognitive engagement. The statistically significant differences in phasic EDA indicate arousal changes.

Our findings might suggest that during stages of information searching, the participants were continuously engaged in active thinking and problem-solving cognitive processing to navigate and evaluate information. The statistically significant differences imply the influence of both interaction modalities and cognitive intents.

To examine this more closely, we then focused on the eye-driven stages: EYEOPEN, IN, and RJ-R, which have the same interaction modality but different intents. Our results that EEG *Alpha* differed in temporal-parietal regions suggest that attentional focus and cognitive effort vary depending on the intent behind gaze movements.

Considering the possible cognitive activities participants might engage in during these stages, at EYEOPEN, participants only need to focus on the screen and relax. At IN, participants may engage in cognitive processes such as comprehending information, retrieving memories, assessing personal interest, and updating knowledge gaps, which drive subsequent actions [43, 49]. At RJ-R, beyond comprehension, participants also evaluated information based on factors such as usefulness and topical relevance, ultimately deciding whether to accept it [1, 26, 49]. Taken together, and in alignment with existing findings, our results indicated the lack of significant difference between IN and RJ-R implies these two tasks engage similar neural mechanisms, while EYEOPEN appears to invoke distinct attentional shifts.

6.2 RQ2. The Influence of Interaction Modalities on the Physiological Responses

Our second research question explores whether different interaction modalities, at search stages, influence the physiological responses. To address this, we performed pairwise comparisons between search stages that share the same cognitive intents but differ in interaction method, specifically, QS-S and QS-T, and RJ-R and RJ-L (refer to Section 5.4).

As mentioned earlier, our results showed that EEG *Alpha* activity at the left-parietal (P7) and left-occipital (O1) regions exhibited the most prominent statistical values across all search stages, indicating dynamics in visual, auditory, or spatial processing demands. This might suggest that when the intents are the same, the ways of interaction influence how the brain processes information, mainly related to parts that control the sensory interactions. The pairwise analysis revealed statistical differences in the EEG *Theta* and *Alpha* activities at the parietal-occipital regions (P7, O1, O2), and also the right-frontal regions (F4) between RJ-R and RJ-L. This might suggest that receiving the search results by reading or by listening primarily affects the cognitive processes related to memory, visuospatial or auditory processing, and temporal integration.

In this dataset, the lengths of the queries submitted by participants between QS-T and QS-S were similar. The text queries had an average of 8 words, ranging from 2 to 19 words, and the voice queries had an average of 7.8 words, ranging from 2 to 20 words. However, the interaction durations differed. Typing required more time (average: 17.7 seconds) compared to speaking (average: 6.8 seconds). These differences in interaction duration and underlying cognitive processing might manifest in changing physiological data [23]. Interestingly, no significant differences were observed in any brain regions or peripheral physiological data between QS-T and QS-S. This

suggests that the modality used to express the query – typed or spoken – may not significantly affect cognitive load or neural engagement captured by the physiological data.

For future research, depending on the research goals, we suggest that wearable EEG headsets should prioritize sensor placement at parietal-occipital sites to effectively track modality-related cognitive shifts. For real-time attention and workload estimation, adding frontal electrodes (e.g., F4) can help detect mental effort differences in information reception tasks, or integrating eye-tracking and EDA sensors in smart wearable devices. The wearable systems (e.g., AR glasses, brain-sensing headbands) [31] should adjust content delivery, optimizing input/output modalities for cognitive efficiency.

6.3 RQ3. Classification Performance of Physiological and Behavioral Data

Our third research question RQ3 examines the ability of physiological and behavioral data to discriminate between search stages. To address this, we develop machine-learning models and compare their performance in classifying search stages using both individual and combined sensor groups (refer to Section 5.5).

Overall, the physiological data yielded varying classification performances. The model using only EDA data demonstrated low performance, with the average F1 scores only marginally better than the baseline. Then, the fusion models that incorporate EDA also showed reduced performance. These suggest that cognitive processes during search stages do not strongly influence wristband-captured EDA in this controlled setup. EEG data demonstrated limited capacity in discriminating the search stages. The minimal impacts of feature window size might suggest that similar neural responses occur within short time frames and remain consistent. This aligns with EEG characteristic of capturing rapid neural events, where extending the time window provides little additional information. PUPIL data exhibited the best performance, with the single model and majority of the fusion models achieving average F1 scores exceeding 50%. Pupillary response is usually used to indicate attention and cognitive load [60]. This high performance might imply varying inherent demands of cognitive load across search stages. Therefore, researchers should be cautious when using pupillary responses to investigate experimental factors in different contexts to ensure validity.

Behavioral data, including GAZE and MOTION, also discriminated certain search stages with moderate performance. But the skewed predictions shown in Figure 10 reflect that they mainly captured the variants in interaction modalities rather than cognitive intent. The fusion models combining these with physiological data led to minor improvements. This further suggests they are useful for context detection but limited for detecting subtle activities.

Besides, the improvements in GAZE, PUPIL, and MOTION models with longer feature windows suggest these modalities benefit from capturing complete behavioral sequences. Ocular and motor responses likely unfold over longer periods and contain meaningful temporal patterns that shorter windows fail to detect. For instance, complete gaze patterns, pupillary reactions to cognitive load, and motion sequences associated with different tasks may require several seconds to fully manifest their distinctive characteristics.

Overall, the differential performance across sensor groups further validates their varying effectiveness in this classification task, and underscores the importance of sensor selection based on the specific cognitive processes under investigation. And the impacts of feature window sizes suggest that optimal feature extraction windows should be tailored to each sensor's specific temporal dynamics rather than applying uniform parameters across all modalities. For real-time applications, this may necessitate modality-specific processing pipelines with different temporal configurations to maximize classification performance.

6.4 Limitations & Ecological Validity

Our dataset serves as a baseline for understanding and assessing consumer-grade sensor data during information searching in a controlled lab setting. However, we note a number of limitations that researchers should be aware

of when using the dataset. The size of the *SenseSeek* is relatively small compared to other sensor datasets, with only 20 participants, where young adults are over-represented. While tasks were presented in English and all participants had professional working proficiency in English, the majority of participants are not native speakers. The dataset only contains eye-tracking data from 18 participants due to a device error. In addition, tasks like speaking, listening, and typing did not necessarily require participants to focus on the screen. This naturally resulted in missing eye-tracking data, which reflects inherent limitations of applying eye-tracking to such tasks.

The data was collected in a controlled lab environment with simulated searches, which may not fully represent real-world distractions and variability. In a natural setting, sensing data can be affected by factors such as contents, visual elements, subjective perceptions, and external environments. We focused specifically on information searching scenarios while controlling variables such as the type of information need (search goal), topic difficulty, familiarity, information complexity, and relevance using a simplified interaction system. While this approach may not perfectly reflect real-life user behaviors, it provides clearer insights into how sensing data responds during information activities. This methodology helps analyze distinct search stages, though it might not capture how these stages seamlessly integrate in everyday use.

Despite our controls, some intricate activities retained variability in their natural state, revealing the dynamic nature of human-information interaction. For example, we separated query formulation from submission to isolate underlying processes. This acknowledges differences between voice interactions (where users may pause before speaking) and text-based interactions (where thinking and typing occur simultaneously). In our study, some participants reported mind-wandering during query formulation, while others used the time for focused thinking, and a few needed to recall their information need before forming queries. As a result, cognitive processes and corresponding physiological responses still overlapped across search stages.

Moreover, search stage durations varied considerably, from 6.8 to 66.0 seconds on average, highlighting the dynamic nature of information searching and the importance of context-specific factors. This variability likely affected physiological responses, as longer stages may elicit more complex responses than shorter ones. In contrast to most existing physiological datasets, which typically involve longer engagement periods, this duration variability could benefit studies featuring rapid cognitive transitions of short-term interactions that characterize real-world search interactions.

Lastly, our study employed only consumer-grade sensors that offer lower accuracy and sampling rates than medical-grade equipment; hence, there is a possibility of artifacts or missed responses that may have impacted the results. As a trade-off, these sensors are closer to the actual technologies users encounter in everyday settings. Future research could bridge this gap by implementing a dual-monitoring approach, using both consumer- and medical-grade or more advanced equipment simultaneously to validate findings while maintaining ecological validity. Additionally, larger sample sizes would strengthen statistical analysis and help distinguish true physiological patterns. Advantages of consumer-grade sensors, such as accessibility, comfort, and ease of use, allow for a scalable version of this dataset.

A future extension of the *SenseSeek* dataset could focus on featuring participants 1) of an older age, 2) with their native language, and 3) with neuro-diversity features, or featuring the tasks with 4) complex information needs, 5) personalized search results (e.g., conversational interaction), and 6) multi-turn interactions.

7 Conclusion and Future Work

We presented SenseSeek, a comprehensive dataset examining the ramifications of passive sensing data and information activity, particularly focusing on the information search process. The dataset comprises physiological and behavioral data from 20 participants across 235 trials, 940 search stages in a controlled lab environment, where each participant completed simulated search tasks. Both the raw and cleaned data are made publicly available, along with 258 extracted features. To provide contextual richness, it also contains the task materials, self-rated

perceptions, and gaze-annotated screen recordings. We validate the usefulness of the dataset by investigating how cognitive intents and interaction modalities affect information searching stages and evaluating the effectiveness of sensor data in distinguishing these stages. Statistically significant differences were observed in EEG and PUPIL data, particularly influenced by cognitive intents and interaction modality (reading vs. listening). Search stages classification was performed using features from individual or multiple sensor groups. Our analysis showed that PUPIL features performed best, and EDA features performed worst. The varying effectiveness of sensor groups underscores the need for sensor selection based on specific cognitive processes and modality-specific processing configurations to optimize classification performance in multi-modality models.

Future work can further explore the relationship between the physiological signals and the variables involved in search stages. In particular, the SenseSeek dataset can be used to advance knowledge on characterizing the perception of topic familiarity, topic difficulty, and relevance with multiple physiological signals.

By making this dataset publicly available and deepening our understanding of these variations, we hope to contribute to the development of passive sensors and multimedia information access systems. For instance, designing adaptive search interfaces that respond to users' cognitive needs in real-time. This dataset lays a foundation for future research exploring more complex search activities and provides a rich source of data for developing algorithms that monitor user experiences such as attention, workload, and stress during information interaction. Researchers can leverage this baseline data to train computational models before applying them in naturalistic environments, ensuring more reliable transitions from controlled to real-world applications.

8 Dataset and Code Availability

The use of the *SenseSeek* dataset is limited to academic research purposes. The *SenseSeek* dataset is publicly available and can be accessed on the OSF platform at https://osf.io/waunb/?view_only=94756f9d2c7a49e094ae42d494c9516a. Additional descriptions of the dataset are included in the appendices. The code used for data processing, analysis, and model training is also available at https://github.com/ADMSCentre/SenseSeek-Dataset-code. The repository contains the instructions for the dataset, several Jupyter Notebook files with data analysis and visualizations, and Python source files for model training. All required packages are listed in the `requirements.txt` file.

Acknowledgments

This research is partially supported by the Australian Research Council and the Australian Research Council Centre of Excellence for Automated Decision-Making and Society (DE200100064, CE200100005).

References

[1] Marco Allegretti, Yashar Moshfeghi, Maria Hadjigeorgieva, Frank E. Pollick, Joemon M. Jose, and Gabriella Pasi. 2015. When Relevance Judgement is Happening? An EEG-based Study. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 719–722. doi:10.1145/2766462.2767811

[2] Ioannis Arapakis, Miguel Barreda-Ángeles, and Alexandre Pereda-Baños. 2019. Interest as a Proxy of Engagement in News Reading: Spectral and Entropy Analyses of EEG Activity Patterns. *IEEE Transactions on Affective Computing* 10, 1 (2019), 100–114. doi:10.1109/TAFFC.2017.2682089

[3] Ebrahim Babaei, Benjamin Tag, Tilman Dingler, and Eduardo Velloso. 2021. A Critique of Electrodermal Activity Practices at CHI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 177, 14 pages. doi:10.1145/3411764.3445370

[4] Swarnali Banik, Sougata Sen, Snehanshu Saha, and Surjya Ghosh. 2024. Towards Reducing Continuous Emotion Annotation Effort During Video Consumption: A Physiological Response Profiling Approach. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 91 (Sept. 2024), 32 pages. doi:10.1145/3678569

[5] Oswald Barral, Manuel J.A. Eugster, Tuukka Ruotsalo, Michiel M. Spapé, Ilkka Kosunen, Niklas Ravaja, Samuel Kaski, and Giulio Jacucci. 2015. Exploring Peripheral Physiology as a Predictor of Perceived Relevance in Information Retrieval. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 389–399. doi:10.1145/2678025.2701389

[6] Jordan J. Bird, Luis J. Manso, Eduardo P. Ribeiro, Anikó Ekárt, and Diego R. Faria. 2018. A Study on Mental State Classification using EEG-based Brain-Machine Interface. In *2018 International Conference on Intelligent Systems (IS)* (Funchal, Portugal). IEEE, 795–800. doi:10.1109/IS.2018.8710576

[7] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. 2024. WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 175 (Nov. 2024), 21 pages. doi:10.1145/3699776

[8] Nattapat Boonprakong, Xiuge Chen, Catherine Davey, Benjamin Tag, and Tilman Dingler. 2023. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases When Facing Different Opinions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 19 pages. doi:10.1145/3544548.3580917

[9] Patrícia Bota, Joana Brito, Ana Fred, Pablo Cesar, and Hugo Silva. 2024. A real-world dataset of group emotion experiences based on physiological data. *Scientific Data* 11, 1 (2024), 116.

[10] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V Elst. 2012. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 2 (2012), 1–30.

[11] Luis Cabañero-Gomez, Ramon Hervas, Ivan Gonzalez, and Luis Rodriguez-Benitez. 2021. eeglib: A Python module for EEG feature extraction. *SoftwareX* 15 (2021), 100745. doi:10.1016/j.softx.2021.100745

[12] Michael J. Cole, Chathra Hendahewa, Nicholas J. Belkin, and Chirag Shah. 2014. Discrimination between Tasks with User Activity Patterns during Information Search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 567–576. doi:10.1145/2600428.2609591

[13] Andrew T. Duchowski, Krzysztof Krejtz, Nina A. Gehrer, Tanya Bafna, and Per Bækgaard. 2020. The Low/High Index of Pupillary Activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376394

[14] Ashlee Edwards and Diane Kelly. 2017. Engaged or Frustrated? Disambiguating Emotional State in Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 125–134. doi:10.1145/3077136.3080818

[15] Raya Fidel. 2012. *Human Information Interaction: An Ecological Approach to Information Behavior.* Mit Press.

[16] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D. Salim. 2020. n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 79 (sep 2020), 26 pages. doi:10.1145/3411813

[17] Christoph Gebhardt, Andreas Brombach, Tiffany Luong, Otmar Hilliges, and Christian Holz. 2024. Detecting Users' Emotional States during Passive Social Media Use. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 77 (May 2024), 30 pages. doi:10.1145/3659606

[18] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2016. cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. *IEEE Transactions on Biomedical Engineering* 63, 4 (2016), 797–804. doi:10.1109/TBME.2015.2474131

[19] Akash Gupta, Harsh Sahu, Nihal Nanecha, Pradeep Kumar, Partha Pratim Roy, and Victor Chang. 2019. Enhancing Text Using Emotion Detected from EEG Signals. *Journal of Grid Computing* 17 (2019), 325–340.

[20] Kunal Gupta, Sam W. T. Chan, Yun Suen Pai, Nicholas Strachan, John Su, Alexander Sumich, Suranga Nanayakkara, and Mark Billinghurst. 2022. Total VREcall: Using Biosignals to Recognize Emotional Autobiographical Memory in Virtual Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 55 (July 2022), 21 pages. doi:10.1145/3534615

[21] Jacek Gwizdka, Rahilsadat Hosseini, Michael Cole, and Shouyi Wang. 2017. Temporal Dynamics of Eye-Tracking and EEG during Reading and Relevance Decisions. *J. Assoc. Inf. Sci. Technol.* 68, 10 (oct 2017), 2299–2312.

[22] Zhiyu He, Shaorun Zhang, Peijie Sun, Jiayu Li, Xiaohui Xie, Min Zhang, and Yiqun Liu. 2023. Understanding User Immersion in Online Short Video Interaction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 731–740. doi:10.1145/3583780.3615099

[23] Grazia Iadarola, Angelica Poli, and Susanna Spinsante. 2021. Analysis of Galvanic Skin Response to Acoustic Stimuli by Wearable Devices. In *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (Lausanne, Switzerland). IEEE, 1–6. doi:10.1109/MeMeA52024.2021.9478673

[24] Soumy Jacob, Shoya Ishimaru, Syed Saqib Bukhari, and Andreas Dengel. 2018. Gaze-Based Interest Detection on Newspaper Articles. In *Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction* (Warsaw, Poland) *(PETMEI '18)*. Association for Computing Machinery, New York, NY, USA, Article 4, 7 pages. doi:10.1145/3208031.3208034

[25] Mainak Jas, Denis A. Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. 2017. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage* 159 (2017), 417–429. doi:10.1016/j.neuroimage.2017.06.030

[26] Kaixin Ji, Danula Hettiachchi, Flora D. Salim, Falk Scholer, and Damiano Spina. 2024. Characterizing Information Seeking Processes with Multiple Physiological Signals. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington, DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 12 pages. doi:10.1145/3626772.3657793

[27] Angel Jimenez-Molina, Cristian Retamal, and Hernan Lira. 2018. Using psychophysiological sensors to assess mental workload during web browsing. *Sensors* 18, 2 (2018), 458.

[28] Ryan Kaveh, Carolyn Schwendeman, Leslie Pu, Ana C. Arias, and Rikky Muller. 2024. Wireless Ear EEG to Monitor Drowsiness. *Nature Communications* 15, 1 (Aug. 2024), 6520. doi:10.1038/s41467-024-48682-7

[29] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224. doi:10.1561/1500000012

[30] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[31] Nataliya Kosmyna and Eugene Hauptmann. 2024. Are you still watching? Assessing internal and external attention in AR using wearable brain sensing glasses. In *Optical Architectures for Displays and Sensing in Augmented, Virtual, and Mixed Reality (AR, VR, MR) V*, Naamah Argaman, Hong Hua, and Daniel K. Nikolov (Eds.), Vol. 12913. International Society for Optics and Photonics, SPIE, 1291310. doi:10.1117/12.3023316

[32] Nataliya Kosmyna and Pattie Maes. 2019. AttentivU: a Biofeedback Device to Monitor and Improve Engagement in the Workplace. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1702–1708. doi:10.1109/EMBC.2019.8857177

[33] Vladimir Kosonogov, Danila Shelepenkov, and Nikita Rudenkiy. 2023. EEG and peripheral markers of viewer ratings: a study of short films. *Frontiers in Neuroscience* 17 (2023), 1148205. doi:10.3389/fnins.2023.1148205

[34] Carol Collier Kuhlthau. 2005. Information Search Process. *CITE Seminar: Information Literacy and Pre-service Programs, Hong Kong, China* 7 (2005), 226.

[35] J. Satheesh Kumar and P. Bhuvaneswari. 2012. Analysis of Electroencephalography (EEG) Signals and Its Categorization–A Study. *Procedia Engineering* 38 (2012), 2525–2536. doi:10.1016/j.proeng.2012.06.298 INTERNATIONAL CONFERENCE ON MODELLING OPTIMIZATION AND COMPUTING.

[36] Matias Laporte, Martin Gjoreski, and Marc Langheinrich. 2023. LAUREATE: A Dataset for Supporting Research in Affective Computing and Human Memory Augmentation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 106 (sep 2023), 41 pages. doi:10.1145/3610892

[37] Minji Lee, Gi-Hwan Shin, and Seong-Whan Lee. 2020. Frontal EEG Asymmetry of Emotion for the Same Auditory Stimulus. *IEEE Access* 8 (2020), 107200–107213. doi:10.1109/ACCESS.2020.3000788

[38] Adam Li, Jacob Feitelberg, Anand Prakash Saini, Richard Höchenberger, and Mathieu Scheltienne. 2022. MNE-ICALabel: Automatically annotating ICA components with ICLabel in Python. *Journal of Open Source Software* 7, 76 (2022), 4484. doi:10.21105/joss.04484

[39] Irene Lopatovska and Ioannis Arapakis. 2011. Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. *Inf. Process. Manage.* 47, 4 (jul 2011), 575–592. doi:10.1016/j.ipm.2010.09.001

[40] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. Neurokit2: A Python Toolbox for Neurophysiological Signal Processing. *Behavior Research Methods* 53, 4 (Aug. 2021), 1689–1696. doi:10.3758/s13428-020-01516-y

[41] Joel T Martin, Joana Pinto, Daniel P Bulte, and Manuel Spitschan. 2021. PyPlr: A versatile, integrated system of hardware and software for researching the human pupillary light reflex. *Behavior Research Methods* (2021), 2720–2739. arXiv:https://link.springer.com/content/pdf/10.3758/s13428-021-01759-3.pdf doi:10.3758/s13428-021-01759-3

[42] Fernando Martínez-Santiago, Alejandro A Torres-García, Arturo Montejo-Ráez, and Nicolás Gutiérrez-Palma. 2023. The impact of reading fluency level on interactive information retrieval. *Universal Access in the Information Society* 22, 1 (2023), 51–67.

[43] Dominika Michalkova, Mario Parra-Rodriguez, and Yashar Moshfeghi. 2022. Information Need Awareness: An EEG Study. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 610–621. doi:10.1145/3477495.3531999

[44] Randall K. Minas, Robert F. Potter, Alan R. Dennis, Valerie Bartelt, and Soyoung Bae. 2014. Putting on the Thinking Cap: Using NeuroIS to Understand Information Processing Biases in Virtual Teams. *Journal of Management Information Systems* 30, 4 (2014), 49–82. doi:10.2753/MIS0742-1222300403

[45] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K. Dey, Sidney K. D'Mello, Ge Gao, Julie M. Gregg, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Qiang Liu, Gloria Mark, Gonzalo J. Martinez, Stephen M. Mattingly, Edward Moskal, Raghu Mulukutla, Subigya Nepal, Kari Nies, Manikanta D. Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, and Aaron Striegel. 2019. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (June 2019), 1–24. doi:10.1145/3328908

[46] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2014. Assessing the Cognitive Complexity of Information Needs. In *Proceedings of the 2014 Australasian Document Computing Symposium* (Melbourne, VIC, Australia) *(ADCS '14)*. ACM, New York, NY, USA, Article 97, 4 pages. doi:10.1145/2682862.2682874

[47] Patricia L. Moravec, Randall K. Minas, and Alan R. Dennis. 2019. Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense At All. *MIS Quarterly* 43, 4 (2019), pp. 1343–1360, A1–A13. https://www.jstor.org/stable/26848107

[48] Yashar Moshfeghi and Joemon M. Jose. 2013. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 133–142. doi:10.1145/2484028.2484074

[49] Yashar Moshfeghi and Frank E. Pollick. 2018. Search Process as Transitions Between Neural States. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1683–1692. doi:10.1145/3178876.3186080

[50] Yashar Moshfeghi, Peter Triantafillou, and Frank Pollick. 2019. Towards Predicting a Realisation of an Information Need based on Brain Signals. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1300–1309. doi:10.1145/3308558.3313671

[51] Yashar Moshfeghi, Peter Triantafillou, and Frank E Pollick. 2016. Understanding Information Need: An fMRI Study. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 335–344. doi:10.1145/2911451.2911534

[52] Diane Nahl. 2007. Social–biological information technology: An integrated conceptual framework. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 2021–2046.

[53] René Riedl, Fred Davis, and Alan Hevner. 2014. Towards a NeuroIS Research Methodology: Intensifying the Discussion on Methods, Tools, and Measurement. *Journal of the Association for Information Systems* 15, 10 (2014), 4. doi:10.17705/1jais.00377

[54] Stanisław Saganowski, Joanna Komoszyńska, Maciej Behnke, Bartosz Perz, Dominika Kunc, Bartłomiej Klich, Łukasz D Kaczmarek, and Przemysław Kazienko. 2022. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific data* 9, 1 (2022), 158.

[55] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. 2019. Wearable-Based Affect Recognition—A Review. *Sensors* 19, 19 (Jan. 2019), 4079. doi:10.3390/s19194079 Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

[56] Kshitij Sharma, Evangelos Niforatos, Michail Giannakos, and Vassilis Kostakos. 2020. Assessing Cognitive Performance Using Physiological and Facial Features: Generalizing across Contexts. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 95 (Sept. 2020), 41 pages. doi:10.1145/3411811

[57] Md. Hedayetul Islam Shovon, D (Nanda) Nandagopal, Jia Tina Du, Ramasamy Vijayalakshmi, and Bernadine Cocks. 2015. Cognitive Activity during Web Search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 967–970. doi:10.1145/2766462.2767784

[58] Winnie K. Y. So, Savio W. H. Wong, Joseph N. Mak, and Rosa H. M. Chan. 2017. An evaluation of mental workload with frontal EEG. *PLOS ONE* 12, 4 (04 2017), 1–17. doi:10.1371/journal.pone.0174949

[59] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. 2022. VREED: Virtual Reality Emotion Recognition Dataset Using Eye Tracking & Physiological Measures. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 178 (Dec. 2022), 20 pages. doi:10.1145/3495002

[60] Pauline van der Wel and Henk Van Steenbergen. 2018. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychon Bull & Review* 25 (2018), 2005–2015. doi:10.3758/s13423-018-1432-y

[61] Ryen W. White and Ryan Ma. 2017. Improving Search Engines via Large-Scale Physiological Sensing. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 881–884. doi:10.1145/3077136.3080669

[62] Yingying Wu, Yiqun Liu, Ning Su, Shaoping Ma, and Wenwu Ou. 2017. Predicting Online Shopping Search Satisfaction and User Behaviors with Electrodermal Activity. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 855–856. doi:10.1145/3041021.3054226

[63] Ziyi Ye, Xiaohui Xie, Qingyao Ai, Yiqun Liu, Zhihong Wang, Weihang Su, and Min Zhang. 2024. Relevance Feedback with Brain Signals. *ACM Trans. Inf. Syst.* 42, 4, Article 93 (feb 2024), 37 pages. doi:10.1145/3637874

[64] Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022. Towards a Better Understanding of Human Reading Comprehension with Brain Signals. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 380–391. doi:10.1145/3485447.3511966

[65] Heeseung Yu and Eunkyoung Han. 2024. People see what they want to see: an EEG study. *Cognitive Neurodynamics* 18, 3 (2024), 1167–1181.

[66] Mahmoud Zeydabadinezhad, Jon Jowers, Derek Buhl, Brian Cabaniss, and Babak Mahmoudi. 2024. A personalized earbud for non-invasive long-term EEG monitoring. *Journal of Neural Engineering* 21, 2 (apr 2024), 026026. doi:10.1088/1741-2552/ad33af

[67] Baiqiao Zhang, Xiangxian Li, Yunfan Zhou, Juan Liu, Chao Zhou, Weiying Liu, and Yulong Bian. 2024. Are We in the Zone? Exploring the Features and Method of Detecting Simultaneous Flow Experiences Based on EEG Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 152 (Nov. 2024), 42 pages. doi:10.1145/3699774

[68] Shaorun Zhang, Zhiyu He, Ziyi Ye, Peijie Sun, Qingyao Ai, Min Zhang, and Yiqun Liu. 2024. EEG-SVRec: An EEG Dataset with User Multidimensional Affective Engagement Labels in Short Video Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 698–708. doi:10.1145/3626772.3657890

A Additional Description of the *SenseSeek* dataset

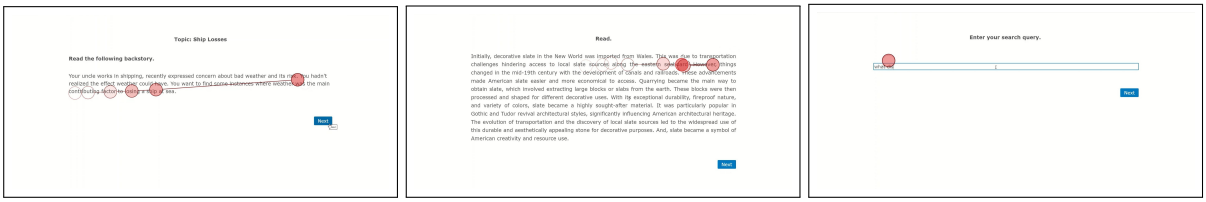


Fig. 11. Examples of Gaze-annotated screen recordings.

Table 11. Demographic information & pre-study questionnaire per participant.

PID	Age (years old)	Gender	English Proficiency Level	Right-handed	Sleep (hours)	Caffeine Intake (cups, in the last two hours)	Wear Glasses
PA5	18–24	Female	Native English	Y	7–9	N	N
PA6	25–34	Male	Native English	Y	<7	Y–1	N
PA8	18–24	Female	Professional Work English	Y	<7	N	Y
PA9	25–34	Female	Professional Work English	N	7–9	Y – 1	N
PA11	25–34	Male	Professional Work English	Y	<7	N	Y
PA12	25–34	Male	Professional Work English	N	7–9	N	N
PA13	18–24	Male	Full Professional English	Y	<7	Y – 1	N
PA17	18–24	Male	Professional Work English	Y	7–9	N	Y
PA18	25–34	Female	Professional Work English	Y	7–9	Y ≥ 2	Y
PA19	25–34	Male	Professional Work English	Y	7–9	N	N
PA20	25–34	Male	Professional Work English	Y	<7	N	N
PA21	25–34	Female	Professional Work English	Y	7–9	N	Y
PA22	25–34	Male	Professional Work English	Y	7–9	N	Y
PA27	25–34	Male	Full Professional English	Y	7–9	N	Y
PA29	35–44	Male	Professional Work English	Y	7–9	N	N
PA30	25–34	Male	Full Professional English	Y	7–9	N	Y
PA31	25–34	Female	Native English	Y	7–9	N	N
PA32	25–34	Male	Native English	Y	7–9	N	N
PA33	25–34	Female	Native English	Y	7–9	N	N

A.1 Details of Feature Extraction

EDA. A total of 31 EDA features were extracted, following Gao et al. [16], Barral et al. [5], and Boonprakong et al. [8]. We extracted the 5 descriptive statistical features, including mean, minimum, maximum, standard deviation, and range, from the mixed EDA values and their first and second derivatives, the Skin Conductance Level (SCL) and the Skin Conductance Response (SCR). Apart from that, we also extracted more statistical features, including median, variance, skewness, and kurtosis, from the mixed EDA values, median, and mean amplitude for the SCL and SCR.

EEG. A total of 180 EEG features were extracted, following Gwizdka et al. [21] and Bird et al. [6]. We used the *eeglib* library [11] for feature extraction.

For each of the 14 channels, we computed the mean, standard deviation, kurtosis, curve length, zero crossings, number of peaks, and wavelet entropy, which are the common features in most EEG experiments [21, 30]. We also calculated the powers of the 4 frequency bands for each of the 14 EEG channels [6], *Theta* (4–8 Hz), *Alpha*

(8–13 Hz), *Beta*(13–25 Hz) and *LowGamma* (25–40 Hz) bands. The powers were normalized by dividing by the total powers.

Additionally, we derived features for left-right hemispheric asymmetry in each frequency band. Following Gwizdka et al. [21], we divided the channels into four groups: frontal-left (AF3, F7, F3, FC5), frontal-right (FC6, F4, F8, AF4), back-left (P7, O1), and back-right (P8, O2). Inter-hemispheric asymmetry was calculated by dividing the mean power of right hemisphere groups (frontal-right and back-right) by the mean power of left hemisphere groups (frontal-left and back-left). The intra-hemispheric asymmetry was calculated by dividing the mean power of the frontal group by the mean power of the back group for each hemisphere.

PUPIL. A total of 25 PUPIL features extracted, following Gwizdka et al. [21]. We extracted the 5 descriptive statistical features, including mean, minimum, maximum, standard deviation and range from pupil diameter, Relative Pupil Dilation (RPD), and the first derivatives of RPD [21]. We also extracted the 5 quartiles ($10th$, $25th$, $50th$, $75th$, $90th$) from the RPD and the 3 quartiles ($25th$, $50th$, $75th$) from the first derivatives of RPD [21]. Apart from that, the LHIPA index was also extracted. It is a frequency-based approach to measure cognitive load from pupil diameter, purposed by Duchowski et al. [13].

MOTION. A total of 13 MOTION features were extracted from the wrist motion and head motion. The 5 descriptive statistical features of wrist motion were extracted using magnitude [16]. For head motion features, we extracted the mean of the x, y, z coordinates on both angular speed and acceleration. Besides, we also extracted the mean and energy from the magnitude of the acceleration.

GAZE. A total of 6 GAZE features were extracted, following Gwizdka et al. [21]. The GAZE features include the number of occurrences, total and mean duration of each type of eye movement, i.e., fixation and saccade.

B Additional Materials used in Data Collection

B.1 Preparation of the Topics & Backstories & Search Results

The topics and the corresponding backstories (i.e., task scenarios) are selected from the TREC2002-InformationNeed dataset [46]. This dataset contains backstories corresponding to different topics from the TREC2002-4 topic set, and categorizes them into three labels in terms of cognitive complexity. We use the topics from the Understanding category, which requires the participants to find information and gain some understanding of the topics. After removing the topics related to crises, wars, conspiracy, or politics, which might trigger subjective feelings, we select 12 topics.

We use GPT-3.5 to slightly edit the backstories to ensure they all have similar word counts. As a result, the backstories have an average of 40±1 words.

For each backstory, we manually select 2 – 3 articles from TREC document collections, which are relevant to the information need. GPT-3.0 is used to generate the search result based on the provided articles and a binary factual judgment question. The judgment question is to ensure the participants have engaged in the task. We further manually examine the search results and check with the Flesch-Kincaid Grade readability level, and edit with GPT-3.5 if it is too easy or too hard. All of them are converted into an audio format with the Google text-to-speech API for the listening tasks.

Table 12. List of Topics and Backstories used. For the search results and more information, please find them in the dataset file. *Text REtrieval Conference (TREC) Data: https://pages.nist.gov/trec-browser/

Topic ID	TREC* ID	Topic	Backstory (revised with GPT3.5)	Backstory word count	Backstory ID (from [46]))	Search Result word count	Search Result (Flesch-Kincaid Readability)
314	R03.314	Marine Vegetation	You recently heard a commercial about the health benefits of eating algae, seaweed and kelp. This made you interested in finding out about the positive uses of marine vegetation, both as a source of food, and as a potentially useful drug.	41	IN14.002	143	12.6
320	R03.320	Undersea Fiber Optic Cable	The FLAG (Fiber optic link around the globe) system will be the world's longest undersea fiber optic cable. You are interested in finding out more about the project, including which companies are involved, and what technology is needed for such an endeavor.	42	IN14.164	146	10.5
353	R03.353	Antarctica Exploration	On the T.V. news last night, you saw footage of scientists in Antarctica. There seemed a surprising number of people there. This got you wondering what scientific expeditions or projects are under way in Antarctica, and what is planned.	39	IN14.018	147	10.8
355	R03.355	Ocean Remote Sensing	A friend at a university is excited about a chance to work with satellite data for ocean remote sensing. You've become interested in this and you'd like to find out what's being developed in this field and how it's being used.	41	IN14.102	148	11.5
416	R03.416	Three Gorges Project	A friend is traveling to China and plans to cruise the Yangtze River. You're not sure whether that's still possible, as it might now be dammed. You'd like to find out the current status of the Three Gorges Dam project.	40	IN14.077	145	12.8
419	R03.419	Recycle, Automobile Tires	You need to buy new tires for your car, and the local dealer has offered to take the old ones for recycling. You didn't know tires could be recycled and you wonder what new uses they are being put to.	40	IN14.084	147	12.6
433	R03.433	Greek, Philosophy, Stoicism	You visited a museum recently, and heard about the Greek philosophy of stoicism. You start wondering if there is any contemporary interest in this philosophy, and whether Greek stoic plays and artistic productions are taking place or being planned.	39	IN14.146	149	11
448	R03.448	Ship Losses	Your uncle works in shipping, recently expressed concern about bad weather and its risk. You hadn't realized the effect weather could have. You want to find some instances where weather was the main contributing factor to losing a ship at sea.	41	IN14.035	152	12.3
708	T04.708	Decorative Slate Sources	Your sister has just moved in to a new house, with slate stone flooring and a slate stone counter top. You are now curious to explore the sources of slate stone for decorative use, and how it is obtained.	39	IN14.033	150	12.6
743	T04.743	Freighter Ship Registration	You are thinking of going into the shipping business, and decide that you want to learn more about the regulations framework. Specifically, you aim to explore rules and considerations related to the registration of a freight ship in a country.	40	IN14.110	146	13.4
711	T04.711	Train Station Security Measures	There has been a great deal of publicity given to security at airports. But many people also travel by train. You are interested in understanding what security measures have been employed at train stations due to increased security concerns.	39	IN14.142	146	11.8
725	T04.725	Low White Blood Cell Count	You've recently been to your doctor and learned you have a low white blood cell count, but it's not clear yet why this is. You want to find out what disease or condition might have caused this symptom.	38	IN14.059	155	11.3

Table 13. List of self-rating items.

	Self-rating Item	Text	Values
Set 1 (on the Topic)	*Topic Difficulty*	I am familiar with this topic.	1: Strongly disagree 2: Somewhat disagree 3: Neither agree nor disagree 4: Somewhat agree 5: Strongly agree
	Topic Interest	I am interested in this topic.	
	Topic Familiarity	I feel difficult to understand this topic.	
Set 2 (on the Search Result)	*Information Relevance*	This information is relevant to my information need.	
	Information Difficulty	I feel difficult to understand this information.	

C Self-Rated Task Perceptions per Topic/Participant

Table 14. Average scores of self-ratings for 12 topics.

ID	**314**	**320**	**353**	**355**	**416**	**419**	**433**	**448**	**708**	**743**	**711**	**725**
Topic	**Marine Vegetation**	**Undersea Fiber Optic Cable**	**Antarctica exploration**	**Ocean Remote Sensing**	**Three Gorges Project**	**Recycle Automobile Tires**	**Greek Philosophy Stoicism**	**Ship Losses**	**Decorative Slate Sources**	**Freighter ship registration**	**Train Station Security Measures**	**Low White Blood Cell Count**
topic_difficulty	2.2	2.4	2.4	2.4	2.7	2.2	2.8	2.3	2.7	2.2	2.6	2.8
topic_familiarity	2.7	3.6	3.1	3.0	1.7	3.1	2.8	2.7	1.5	3.2	3.0	2.0
topic_interest	3.4	4.0	3.7	3.8	3.0	3.6	3.6	3.4	2.4	3.8	3.9	3.0
info_difficulty	1.8	2.2	1.7	2.0	1.5	2.0	2.3	1.7	2.0	1.4	2.2	2.5
info_relevance	3.4	4.2	4.2	4.2	3.7	4.3	3.4	4.2	3.8	4.6	4.4	3.2
attention_check	100%	75%	100%	95%	90%	85%	55%	95%	75%	75%	70%	55%

Table 15. Average scores of self-ratings for each participant.

PID	topic_familiarity	topic_interest	topic_difficulty	info_revelance	info_difficulty	attention_check
PA5	1.8	3.7	4.0	4.6	1.2	83%
PA6	2.8	3.8	1.2	4.1	1.2	100%
PA8	3.1	3.8	2.5	4.4	1.6	67%
PA9	3.2	3.2	2.0	4.6	1.5	83%
PA11	1.7	3.2	3.3	4.5	1.7	67%
PA12	3.6	3.4	2.4	3.4	2.4	75%
PA13	2.3	3.9	2.5	3.6	1.9	83%
PA17	2.6	3.6	2.8	3.8	3.3	92%
PA18	2.9	3.6	2.0	4.2	2.2	25%
PA19	3.4	4.0	1.8	4.3	1.5	83%
PA20	2.9	3.2	3.0	3.8	2.4	67%
PA21	1.3	1.7	3.2	3.7	2.4	67%
PA22	3.3	3.9	2.1	3.3	1.2	83%
PA26	1.7	1.8	4.2	3.9	3.2	83%
PA27	2.9	3.9	2.3	3.7	1.9	92%
PA29	2.8	4.0	2.2	3.8	2.1	100%
PA30	3.2	4.0	1.0	3.6	1.1	100%
PA31	2.2	3.3	2.4	4.5	1.8	92%
PA32	2.8	3.8	2.9	3.6	2.8	75%
PA33	2.9	3.1	2.2	4.2	1.7	100%