IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 January 2024, accepted 25 February 2024, date of publication 4 March 2024, date of current version 4 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3373542

RESEARCH ARTICLE

A Comprehensive Analysis of Cognitive CAPTCHAs Through Eye Tracking

NGHIA DINH[1], LIDIA DOMINIKA OGIELA[2], KIET TRAN-TRUNG[3], TUAN LE-VIET[3], AND VINH TRUONG HOANG[3], (Graduate Student Member, IEEE)

[1]Faculty of Computer Science, VSB—Technical University of Ostrava, 708-33 Ostrava-Poruba, Czech Republic
[2]Faculty of Computer Science, AGH University of Krakow, 30-059 Kraków, Poland
[3]Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh 722000, Vietnam

Corresponding author: Vinh Truong Hoang (vinh.th@ou.edu.vn)

ABSTRACT CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) has long been employed to combat automated bots. It accomplishes this by utilizing distortion techniques and cognitive characteristics. When it comes to countering security attacks, cognitive CAPTCHA methods have proven to be more effective than other approaches. The advancement of eye-tracking technology has greatly improved human-computer interaction (HCI), enabling users to engage with computers without physical contact. This technology is widely used for studying attention, cognitive processes, and performance. In this specific research, we conducted eye-tracking experiments on participants to investigate how their visual behavior changes as the complexity of cognitive CAPTCHAs varies. By analyzing the distribution of eye gaze on each level of CAPTCHA, we can assess users' visual behavior based on eye movement performance and process metrics. The data collected is then employed in Machine Learning (ML) algorithms to categorize and examine the relative importance of these factors in predicting performance. This study highlights the potential to enhance any cognitive CAPTCHA model by gaining insights into the underlying cognitive processes.

INDEX TERMS Cognitive, security, CAPTCHA, eye tracking, machine learning.

I. INTRODUCTION

CAPTCHA [1] (Completely Automated Public Turing test to tell Computers and Humans Apart) is a widely used method in human-machine interaction systems to distinguish human users from malicious attacks. Despite the popularity of text and picture CAPTCHAs, these designs are still susceptible to automated attacks. Consequently, cognitive CAPTCHAs [2] have emerged as a more secure alternative, employing unique combinations of neurobiological and psychological approaches. However, the use of complex cognitive CAPTCHAs can negatively impact user usability and comprehension. Therefore, it is crucial to investigate the correlation between cognitive CAPTCHA models and users' attention and visualization literacy [3]. Visualization literacy refers to the ability to effectively interpret and extract information from data visualizations, while mental attention [4] pertains to the capacity to maintain and manage information within one's mind.

Eye tracking is currently being used as an effective method to study attention and other cognitive processes [5]. This technique provides us with practical measures in various aspects that can be interpreted to understand how individuals process information. Eye-tracking systems, as depicted in Figure 1, utilize sensors to track the position of the pupil and record different eye movements, including fixations, saccades, and their distinct features, which have been associated with cognitive events. The metrics derived from eye tracking are indicative of users' cognitive abilities, such as mental attention capacity, perceptual speed, and visual working memory. It has been observed that individuals with lower cognitive capacities tend to perform poorly on cognitive tests, both in terms of accuracy and completion time. In recent times, eye movement research has started to leverage machine learning techniques to classify and analyze the significance of these characteristics in predicting

The associate editor coordinating the review of this manuscript and approving it for publication was Jad Nasreddine.

an individual's performance in tasks like literacy [6], [7]. In these studies, classification algorithms were employed to predict whether individuals would exhibit poor or high performance in cognitive activities.

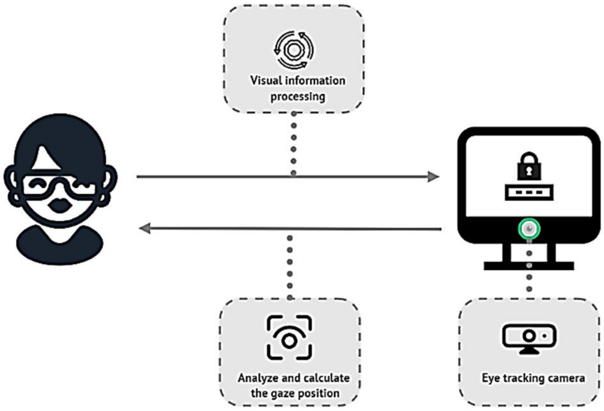
Visual information processing
Analyze and calculate the gaze position
Eye tracking camera

FIGURE 1. Eye tracker procedure.

By utilizing eye-tracking technology, we have conducted a study to explore the correlation between cognitive CAPTCHA models and users' attention and visualization literacy. We aimed to examine how users' comprehension of cognitive CAPTCHAs is influenced by changes in CAPTCHA complexity. To accomplish this, the proposed cognitive CAPTCHA models require users to engage in tasks that involve attention, visual or spatial processing, short-term memory retention, natural language understanding, executive processes, fine-grained motor capabilities, and common-sense reasoning.

To ensure the effectiveness of our study, we carefully selected abstract images without any textures, colors, or closed outlines. Despite the absence of these visual cues, humans are still able to recognize these images due to a combination of perception, visual processing, and past knowledge processing. Additionally, common sense reasoning [8], which involves drawing conclusions based on information obtained from past experiences, plays a crucial role in the cognitive process. As a result, the cognitive CAPTCHA models presented in this research specifically focus on these aforementioned aspects:

- Story completion: selects an ending object to complete a story.
- Object association: makes proper object associations in a semantic context.
- Feature identification: selects objects sharing similar features or having different features from the rest.
- Object composition: combines multiple objects to match a target object.

The selection process for cognitive CAPTCHA was carried out following the guidelines of the Visualization Literacy Assessment Test (VLAT) [3], which adheres to the commonly adopted research approach in the fields of Psychological and Educational Measurement. During the development of the test, cognitive CAPTCHAs that demonstrated the highest Content Validity Ratio (CVR) [9] were carefully chosen. The CVR serves as a measure of the importance of each item in the test, categorizing them as either "essential", "helpful but not essential", or "not necessary". Furthermore, the CAPTCHAs with the highest item discrimination index were selected. This index evaluates how effectively an item can differentiate between individuals who score low and high on the test. The complexity level of each CAPTCHA was determined by its item difficulty index, which indicates the percentage of test-takers who answered the CAPTCHA correctly. In order to examine the relationship between individuals' attention and performance, statistical techniques such as ANOVA (Analysis of Variance) and T-test were utilized to validate our proposed attention theories. To evaluate individual performance, machine learning techniques were employed to classify and determine the relative significance of eye-tracking data, including fixations, saccades, and other variables.

The following sections provide a comprehensive overview of the content of this paper. Section II presents a compilation of works closely associated with the subject matter. In Section III, we delve into the approach adopted for this investigation. Part 4 delves into the experimental findings and subsequent discussions. Lastly, Section V comprises the conclusion and limitations of this study.

II. RELATED WORKS

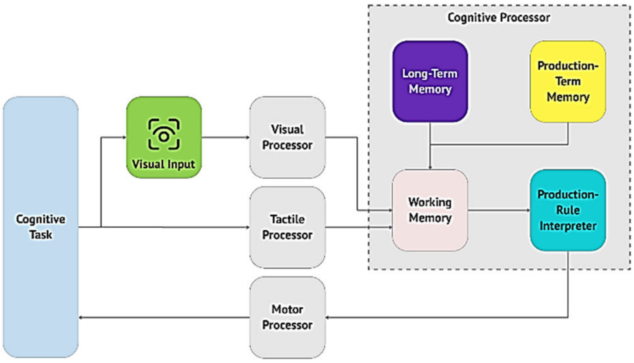

FIGURE 2. Cognitive architecture.

Figure 2 depicts the process of stimulus-processing-response in the cognitive mechanism [10] during visual literacy. The visual pathway, specifically the retina-fovea, is responsible for transmitting information to the central nervous system, where it undergoes processing at both subcortical and cortical levels. This processing leads to the formation of a response to sensory stimulation. Saccades, which are transitions from sensory to motor, occur in response to a stimulus and involve redirecting the fovea from one point of interest (POI) to another. Fixation is also employed to maintain alignment between the fovea and the target throughout subsequent stimulus processing. By utilizing eye-tracking devices to track the retina-fovea, we can gain insights into cognitive processes by analyzing various indices associated with cognitive processes, such as fixations, saccades, and their characteristics.

Eye-tracking technologies, including Tobii [11], EyeTribe [12], and EyeLink [13], among others, have significantly advanced our understanding of human cognition in fields such as psychology, biology, cognitive neuroscience, medical advancements, therapeutic intervention, educational practice, and computer vision. In this context, eye-tracking refers to the process of determining where and when a user's gaze is focused, commonly known as the gazing point or pupil size. By utilizing an eye tracker, one can monitor the position, movement, and other characteristics of a person's eyes, which ultimately helps in comprehending their visual attention. This sensor technology uses a camera to measure observable changes in eye characteristics, like blink frequency, pupil diameter, and light source reflection. The main objective of eye-tracking technology is to identify and interpret eye movements as patterns of movement. Modern eye-tracking systems heavily rely on sensors to detect the position of the pupil and record eye movements, allowing for the detection of various indicators related to cognitive processes, such as fixations, saccades, and their specific characteristics. The data obtained from eye-tracking provides valuable insights into the quantity and quality of information processing during the search phase, depending on the given task. Moreover, when combined with traditional inferential information, eye-tracking data has been used to predict success in a wide range of complex cognitive activities, such as the conceptualization of physics [15]. Typically, studies [16] have demonstrated that shorter saccades and longer fixations are indicative of accuracy across multiple modalities.

Eye movement research has recently incorporated machine learning approaches in various fields related to vision and recognition science. Machine learning is a data analysis technique that enables the automatic and rapid identification of patterns in large datasets. Classification and regression are common problems addressed by machine learning algorithms. In cognitive science, machine learning algorithms have been developed and successfully applied to identify cognitive task performance using eye movement data. In a recent study [17], participants underwent mental attentional capacity exercises with six levels of difficulty, and prediction models were created based on metrics such as reaction time, activity difficulty, and eye movements. The results demonstrated that machine learning algorithms can accurately predict performance, with difficulty level and response time being reliable indicators. Additionally, SVM (Support Vector Machine) [18] has been employed to assess competency levels, literacy levels, and perceived work difficulty based on metrics such as first-pass rereading time, fixation time, second-pass fixation time, and dwell time. The bagged tree classifier and the K-Nearest Neighbors algorithm have also been utilized to identify top performers in the Ruff Figural Fluency test [6], which challenges participants to generate as many meaningful figures as possible from a set of dot combinations.

In recent times, there has been limited exploration into CAPTCHAs, specifically focusing on cognitive CAPTCHAs, utilizing the eye-tracking method. A preliminary examination conducted by Al-Khalifa [19] analyzed how participants tackled CAPTCHAs using eye-tracking. However, this investigation solely focused on traditional CAPTCHAs and only presented empirical data concerning eye-tracking measurements, such as the quantity and duration of fixations. Additionally, there are several studies that do not specifically emphasize cognitive CAPTCHA models but still employ the eye-tracking approach to gauge cognitive workload and forecast individuals' performance in cognitive activities.

Bachurina et al. [17] conducted a comprehensive investigation into the complexities associated with mental attention capacity in adults. The study encompassed six distinct levels of difficulty across various tasks. The researchers discovered a strong correlation between task intensity and an increase in response time, as well as observed variations in reaction time and eye-tracking measurements. To forecast accuracy scores, the team employed advanced machine learning techniques, considering metrics related to task difficulty, reaction time, and eye movements. However, a notable weakness of this research lies in the fact that the testing tasks were not based on any established development standards. The tasks primarily focused on assessing color memorizing, neglecting other cognitive abilities, which makes it challenging to evaluate individuals' overall cognitive capabilities. Additionally, it is evident that there is a linear relationship between task difficulty and response time, with higher difficulty levels resulting in longer response times. Furthermore, the study did not provide a clear analysis of how eye-tracking metrics relate to the difficulties of the tasks.

Ogiela [20] introduced a novel method for contactless cognitive CAPTCHA authentication using eye tracking technology. This innovative approach facilitates contactless authentication by identifying and selecting the appropriate CAPTCHA elements based on the presented questions or semantic requirements. Remarkably, it can be employed in contactless fashion, even in fast-moving transportation systems, ensuring reliable user verification. However, their proposed method primarily focuses on developing a cognitive authentication protocol that leverages eye tracking devices to authenticate users possessing advanced expertise or exceptional perceptual skills. The analysis of eye tracking data, while relevant, falls beyond the scope of this research paper.

In the research conducted by Ktistakis et al. [21], they explored visual search tasks of varying complexities and durations. The objective of the study was to assess the participants' cognitive workload levels using the subjective NASA-TLX test. Through comprehensive data analysis, the researchers extracted eye and gaze features from essential eye-recording metrics. Subsequently, they evaluated and tested different machine-learning models to estimate the cognitive workload level. The findings showed promising results, indicating that machine learning analysis could effectively differentiate between different levels of cognitive workload based solely on eye-tracking characteristics. However, it is important to note that the testing tasks primarily focused on locating a specific item within a square in a given picture. These tasks were categorized based on tasking (multi-tasks or single tasks) and time constraints. The relatively small size of the challenge items could impact the time required for searching. Moreover, while the researchers attempted to distinguish individual cognitive workload levels, they did not provide any basis for improving cognitive workload or testing images.

Our approach adopts a fresh perspective in investigating the interplay among cognitive CAPTCHA models, user attention, visualization literacy, and performance. We developed multiple cognitive CAPTCHA models according to the guidelines set forth by the Visualization Literacy Assessment Test (VLAT) [3], which follows the widely accepted research approach in the fields of Psychological and Educational Measurement. Our aim is to understand how users' visual behavior towards cognitive CAPTCHAs evolves as the complexity of the CAPTCHA increases. This study emphasizes the potential to enhance any cognitive CAPTCHA model by gaining insights into the underlying cognitive processes. In addition to utilizing eye-tracking technology to analyze visual attention and behavior effectively, we employ statistical techniques such as ANOVA (Analysis of Variance) and T-test to validate our attention theories of the correlation between individuals' attention and performance. Furthermore, we employ machine learning techniques to classify and assess the significance of the collected eye-tracking data, which includes fixations, saccades, and other variables, in predicting individual performance.

III. THE METHOD

A. TEST DEVELOPMENT PROCEDURE

Figure 3 illustrates our creation of a visual literacy evaluation exam [3], employing the widely accepted approach found in the fields of Psychological and Educational Measurement. The development of the test encompasses six distinct stages: blueprint formation, item generation, assessment of validity, conducting test trials, analysis and selection, and evaluation of reliability.

1) TEST BLUEPRINT CONSTRUCTION

A test blueprint outlines the key elements of a test, encompassing two primary aspects: (1) the essential subjects that the test will cover, and (2) the cognitive activities associated with those subjects. This blueprint serves as a valuable resource in assessing the content validity of a test. A group comprising five experts specialized in information visualization and cognitive analytics was selected to construct the test blueprint. The average age of the specialists was 35 years. Each individual possessed professional experience ranging from 7 to 15 years (M = 10) in this particular field. Two hailed from the industry, while the remaining three were esteemed academics. These experts collectively put forth various cognitive CAPTCHA models and visualization tasks, which are detailed in Table 1.

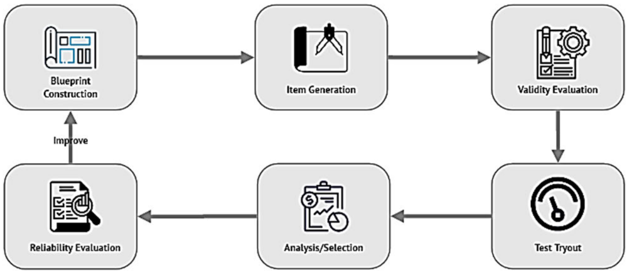

FIGURE 3. Test development procedure.

2) ITEM GENERATION

The test blueprint was used to produce the test CAPTCHAs. During this phase, we will face and must answer the following questions: (1) What kinds of CAPTCHA models would be used? (2) How many CAPTCHAs would be produced?

We polled testers after completing 19 test CAPTCHAs to obtain more feedback for improvement. Initially, 21 participants were recruited, none of them were color-limited vision. Just 19 persons remained due to objective factors. The remaining participants were 7 men and 12 women aged 18 to 35 (M = 22). Everyone had a university education or higher: 16% had a master's or doctoral degree, and 26% had a bachelor's degree. During the survey, we randomly distributed cognitive CAPTCHAs and related tasks, along with task descriptions based on the test design. We asked the participants the following question: "When executing the task, you may obtain your improvement information. Please describe the facts in your own words." Participants were instructed to write down what they had learned. The poll results were reviewed and used as reference materials to improve design quality and usability.

3) CONTENT VALIDITY EVALUATION

Independent domain experts should evaluate the test items to verify that the exam contains relevant tasks by calculating the content validity ratio (CVR) [9]. CVR goes from -1.0 to 1.0 and shows expert consensus on how a certain item is necessary in the test. To acquire CVR for each item, we presented the developed test items with the graphics and tasks one by one and asked the following questions: Is the item and related task "essential", "useful but not essential", or "not necessary" for visualization literacy? We calculated CVR for each item based on the number of experts who indicated "essential". A good CVR value is considered to be one in which more than half of the experts rate this item as "essential". As a result, we kept 10 things with CVR > 0 and removed 9 items with CVR < 0.

TABLE 1. The test blueprint of Cognitive CAPTCHA.

#	Cognitive model	Visualization task	Description
1	Story completion	Select an ending object to complete a story.	The CAPTCHA presents users with a set of instructions where they must choose an image to complete the task. These instructions consist of three abstract images and three corresponding solution images. To successfully solve the CAPTCHA, users are expected to select the accurate solution image from the provided options. This requires the application of common sense and logical reasoning.
2	Object association	Make proper object associations in a semantic context.	The CAPTCHA presents a set of instructions that prompt users to establish accurate connections between challenge images and solution images. The CAPTCHA comprises three abstract challenge images and three corresponding abstract solution images. Users must successfully associate the challenge images with the solution images that carry identical semantic meanings. This CAPTCHA necessitates the utilization of background knowledge processing and abstraction to overcome it.
3	Feature identification	Select objects sharing similar features or having different features from the rest.	The CAPTCHA presents users with a set of instructions, asking them to choose images that exhibit either similar or distinct characteristics from the others. This CAPTCHA comprises six abstract images that serve as challenges. Users must carefully follow the instructions to identify and select the challenge images that possess the desired similarities or differences. Accomplishing this task necessitates the utilization of natural language processing, background knowledge processing, abstraction, and pattern recognition skills.
4	Object composition	Combine multiple objects to match a target object.	The CAPTCHA presents a set of guidelines where users must choose specific challenge images that form a composition resembling the target image. Within the CAPTCHA, there are six abstract challenge images alongside a single abstract target image. To complete the CAPTCHA, users must carefully select challenge images that come together harmoniously to replicate the target image. Resolving this particular CAPTCHA model necessitates a considerable level of abstraction.

4) TEST TRYOUT

The items that have been evaluated are tested on a group of testers. The collected responses are assessed to get quality evidence. Inappropriate items are removed or changed based on the item analysis results.

At the beginning, 41 people were recruited. We followed the same guidelines as the participants in the early stages. There were eventually 29 contestants left. These included 17 women and 12 men ranging in age from 18 to 41 (M = 27). 17% of participants held a master's or doctorate degree, while 28% held a bachelor's degree. We conducted the test, which included the specified ten test items. Then, we gave the participants test instructions. The goal of the exam was explained to the participants in the instructions, and they were instructed to pick the best response to each item within a time restriction. Due to the standardized assessment test, participants were provided with limited time to complete the test. As a result, we gave participants a maximum of 25 seconds to react to a question, and the test should be completed in no more than 5 minutes.

5) ITEM ANALYSIS AND SELECTION

We used classical test theory (CTT) [22] to conduct an item analysis, which included basic statistics, item discrimination index, and item difficulty index.

For basic statistics, we examined the testers' scores. The maximum possible score on the test was 10. The testers' scores ranged from 1 to 10 (M = 6.32, SD = 1.41). We also used the Shapiro-Wilk test to ensure that the scores were normally distributed. The test scores were found to be normally distributed (W = 0.99, p = 0.32). We also recorded the test completion time. The testers' average test completion time was 3 minutes and 10 seconds (SD = 37 seconds). It stated that the time constraint (25 seconds per item) was reasonable for the testers to finish all of the test items.

The item difficulty index [23] is the percentage of the testers who correctly answered the item. The index value is calculated using the following formula and ranges from 0 to 1.0:

$$\boldsymbol{P_i} = \frac{\boldsymbol{N_c}}{\boldsymbol{N}} \tag{1}$$

where P_i is the item difficulty index of the item i, N is the total number of testers, and N_c is the number of testers who answered item i correctly. Using Equation 1, we determined the item difficulty indices of 10 test items. According to the Office of Educational Assessment at the University of Washington [24], each question is classed as easy if the value is greater than 0.85, moderate if the value is between 0.5 and 0.85, and difficult if the value is less than 0.5. The item difficulty indices varied from 0.15 to 1.0 (M = 0.64). There were three simple items, four intermediate items, and three difficult ones among the ten.

The item discrimination index [23] measures how well a test item differentiates between low and high-scoring testers.

The index value ranges from -1.0 to 1.0 and is calculated as follows:

$$D_i = \frac{N_U - N_L}{N} \quad (2)$$

where D_i is the item discrimination index of the item i, N is the total number of testers, N_L is the number of testers who answered item i correctly in the lower group, and N_U is the number of testers who answered item i correctly in the upper group. Using Equation 2, we determined the item discrimination indices of the test's ten items. Each item has a high discriminating value if it is larger than 0.3, a medium discriminating value if it is between 0.1 and 0.3, and a low discriminating value if it is less than 0.1. The indices varied from -0.04 to 0.66 (M = 0.28). As a result, there were 3 medium-discriminating items, 3 low-discriminating items, and 4 high-discriminating items.

We thoroughly analyzed all the items, taking into account their complexity and discrimination. In general, difficult items have high discrimination whereas easy items have low discrimination. To improve test quality, hard items with low discrimination or negative values were deleted. As a result, we selected the five items with the greatest CVRs, valid difficulty, and discrimination, as shown in Table 2.

B. PARTICIPANTS

We recruited 29 individuals (17 men and 12 females) ranging in age from 18 to 35 (M = 22, SD = 2.17). 17% of participants held a master's or doctorate degree, while 31% held a bachelor's degree. They were all skilled computer users who had encountered visual CAPTCHAs before to participating in this trial. None of them had light or color blindness. They also had no issue reading on a computer screen or solving picture CAPTCHAs.

C. APPARATUS

In this study, we utilized the Tobii T60 eye tracker, which was operated through a laptop operating on the Windows 10 platform. The eye tracker had a data acquisition rate of 60 Hz and an angular resolution of 0.5 degrees. To conduct the tests, CAPTCHAs were displayed on a 24-inch screen with a resolution of 1280 × 720. Figure 9 illustrates that we followed the recommended settings provided by the manufacturer to ensure optimal performance of the eye tracker:

- The screen was approximately 80 cm away from the person.
- The CAPTCHAs were presented in the center of the screen.
- The chair's height could be adjusted such that a participant's eyes were horizontally parallel with the center of the screen.
- The brightness of each picture may be affected by changes in pupil size. As a result, the lighting settings in the room were set to be photopic, guaranteeing that the influence of brightness shifts in the photos was minimized.

TABLE 2. Selection test items (P: lower value is more difficult, D: higher value is more discriminative).

#	Cognitive CAPTCHA	Task	Description	CVR	P	D
1	Story completion	Select an ending object to complete a story	As Figure 4, this challenge requires a user to select the next images following the provided rule	1	**0.25**	**0.67**
2	Feature identification	Select objects having different features from the rest	As Figure 5, this challenge requires a user to choose all shapes that differ from the provided rule	1	0.39	0.52
3	Feature identification	Select objects sharing similar features	As Figure 6, this challenge requires a user to choose all animals belonging to reptiles	1	0.56	0.39
4	Object composition	Combine multi objects to match a target object.	As Figure 7, this challenge requires a user to choose all component images that make the provided scene	1	0.75	0.36
5	Object association	Make proper object associations in a semantic context	As Figure 8, this challenge requires a user to make connections between images having the same semantic context	1	0.83	0.24

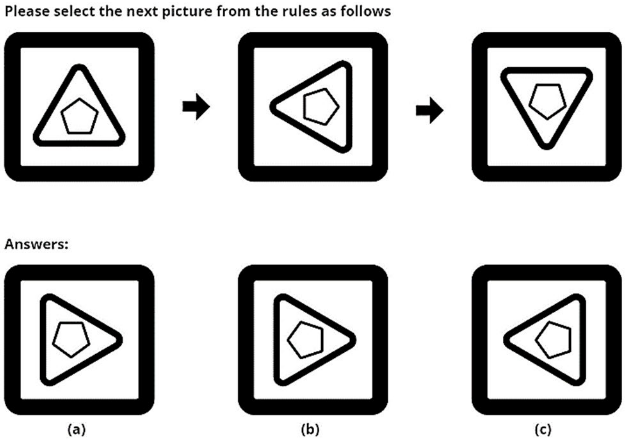

FIGURE 4. Story completion.

The eye-tracking data was recorded using Tobii Studio Pro 4.0, an analysis program provided by the manufacturer. Before the test session, each participant underwent a

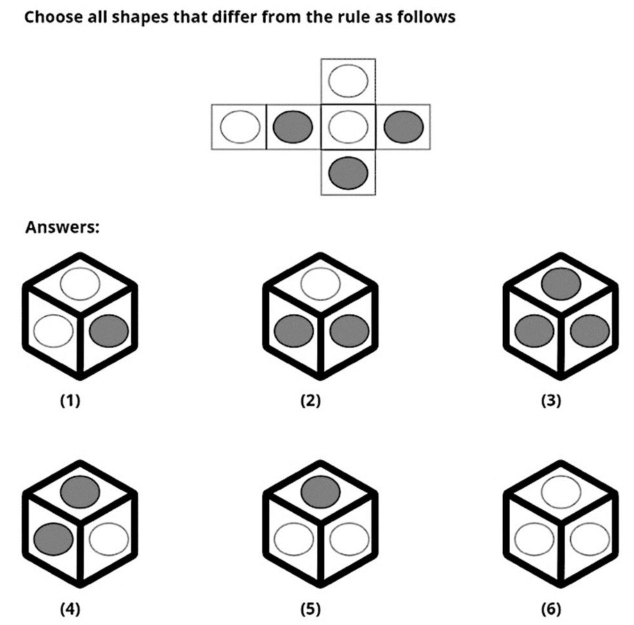

FIGURE 5. Different feature identification.

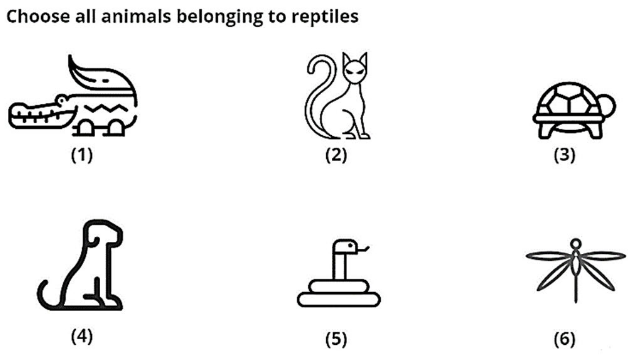

FIGURE 6. Similar feature identification.

calibration session on Tobii Studio Pro 4.0 to ensure the accuracy of the eye-tracking data. The recorded data includes the path of the gaze on the screen and the duration of fixation on specific areas of the screen. The test CAPTCHAs were generated on a remote server and then downloaded as webpages onto the local browser, Firefox. To solve the test CAPTCHAs, participants were provided with conventional input devices such as a keyboard and mouse. The solution time and accuracy rate for each CAPTCHA design were stored on a remote server for further investigation.

D. EXPERIMENTAL PROCEDURE

Initially, every participant underwent the process of carefully reading and signing a DNA (Non-Disclosure Agreement) Information Consent Form. Subsequently, they were accompanied to the designated experiment room, where necessary precautions were implemented to ensure the safety of both the participants and the research team, considering the prevailing SARS-CoV-2 epidemic and the potential transmission of the coronavirus.

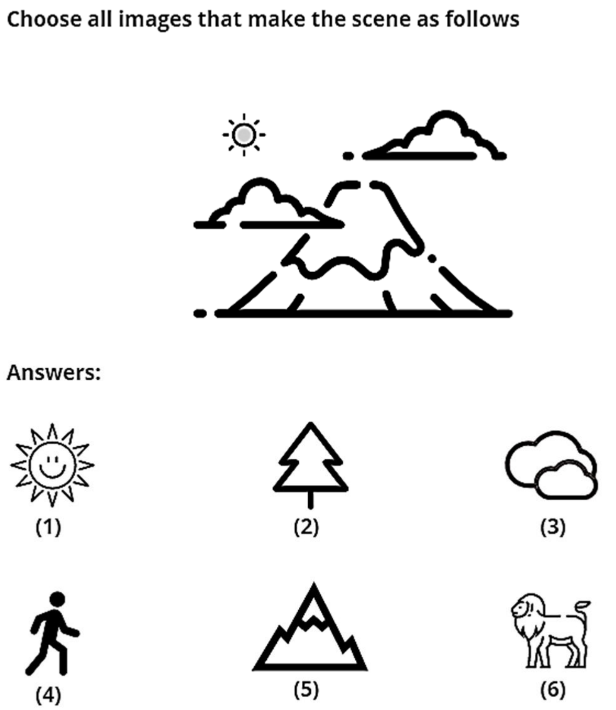

FIGURE 7. Object composition.

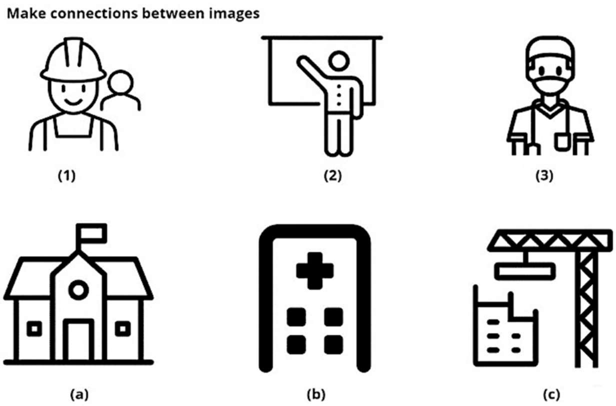

FIGURE 8. Object association.

Following the safety measures, participants were directed to a computer screen, where they were required to provide basic demographic information encompassing age, gender, education level, and job status. To familiarize themselves with the procedure, a preliminary test involving a random picture was conducted. Once this introductory test was completed, the main phase of the research commenced. During the primary phase, participants were given a limited timeframe of 25 seconds to respond to each test item. These test items were presented in a randomized order to ensure fairness. A brief interval of 15 seconds was provided between each test item, allowing participants a momentary break. Upon completion of the examination, participants were asked to fill out a questionnaire to gauge their sentiments and opinions regarding the test.

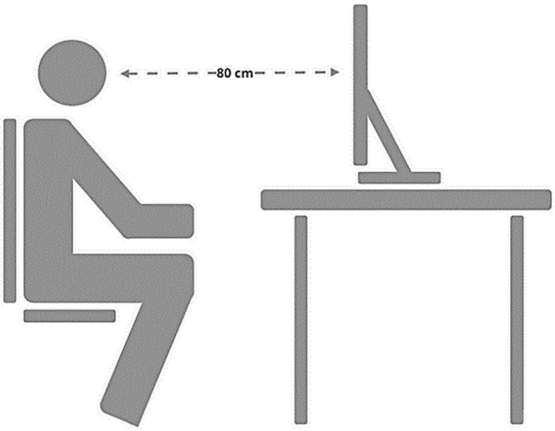

FIGURE 9. Experimental setup.

Throughout the entire process, the research team members diligently monitored the output on a separate screen, providing additional support as needed. It is important to note that participants retained the right to request discontinuation of the procedure and deletion of their data at any point during the research endeavor.

E. QUESTIONNAIRES

The satisfaction questionnaire associated with each cognitive CAPTCHA test focuses on three primary aspects:

- (Q1) Visual comfort - is it pleasing to the eye?
- (Q2) Ease of use - is it easy to identify and resolve?
- (Q3) Applicability - is it suitable for the eye-tracking purposes?

Participants are asked to score each design feature in terms of (Q1) visual comfort, (Q2) convenience of use, and (Q3) applicability. The ratings were determined using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

IV. RESULT ANALYSIS

A. METHODOLOGY

1) DATA PREPROCESSING

The provided information undergoes a pre-processing phase to eliminate any unwanted disturbances such as noise, missing data, and irrelevant details. Additionally, adjustments are made to account for variations in individual pupil sizes. Unwanted volatility in the eye movement data is removed through the implementation of denoising and filtering techniques [22]. To mitigate velocity noise, a five-tap finite impulse response (FIR) velocity filter [23] is utilized, which responds to a predetermined velocity peak value during a saccade. The identification of saccades and fixations is performed using the Velocity-Threshold Identification (I-VT) method [24] due to its superiority in sample-by-sample comparisons. In this method, the velocity threshold for saccade detection is set to 45 degrees per second. Furthermore, a minimum fixation duration threshold of 55 milliseconds is established.

2) FEATURE SELECTION

Fixation-based measurements [25] are frequently utilized in the analysis of eye gaze data as they provide insights into the amount of information processed by the user. Saccadic characteristics [26] are also employed to uncover patterns of consumer attention. Pupil dilation is chosen due to its ability to provide information regarding decision-making [27] and the relevance of search results [28]. Saccadic velocity is also employed to gain an understanding of the complexity of tasks and fluctuations in mental workload [29]. These metrics are computed using basic statistical measures such as absolute value, mean, and standard deviation (SD). A comprehensive list of these features can be found in Table 3. To address potential issues, we applied the MinMaxScaler function to normalize each feature within a range of 0 to 1 before fitting the model. Furthermore, we constructed a correlation matrix to examine the correlation coefficients between variables for more advanced analysis.

3) ANALYSIS TECHNIQUES

a: HYPOTHESIS TESTING ANALYSIS

In this research, we investigate our proposed theories related to individuals' focus, as outlined in Table 4. To assess these theories, we utilize statistical techniques such as ANOVA (Analysis of Variance) and T-test. These methods allow us to analyze the experimental data effectively, with the assistance of the statistical software package SPSS v.26.

b: MACHINE LEARNING ANALYSIS

In order to predict performance, we have chosen a diverse array of well-known and extensively utilized classifiers. These classifiers include Gaussian Naive Bayes (GNB), Logistic Regression (LR), and Support Vector Machine (SVM). Gaussian Naive Bayes (GNB) is commonly employed for classification tasks, particularly when working with continuous data and assuming a normal distribution for the features is a reasonable assumption. This classifier is highly regarded for its simplicity, computational efficiency, and its ability to produce satisfactory results in practical applications. Logistic regression (LR) is widely utilized across various fields, such as medicine, social sciences, and machine learning. It is a robust and easily interpretable algorithm, especially when the relationship between the features and the outcome is believed to be approximately linear. SVMs are extensively used in diverse applications, including text classification, image recognition, and bioinformatics. They are particularly effective when dealing with high-dimensional data and situations where a distinct margin exists between classes. Furthermore, in order to evaluate the importance of attributes in forecasting performance, we utilize the Random Forest Classifier (RFC). This particular classifier is renowned for its resilience, exceptional precision, and ability to avoid overfitting. Unlike individual decision trees, Random Forests are less susceptible to overfitting due to their ensemble methodology and the introduction of randomness during

TABLE 3. Selected features.

#	Measure	Feature	Description
1	Fixation	Fixation frequency Fixation duration	When a tester repeatedly and frequently focuses on a particular stimulus, it could indicate their lack of familiarity with the task or their struggle to differentiate between relevant and irrelevant information. The duration of these fixations reveals the tester's reaction time and level of interest. Longer fixation durations are linked to more extensive cognitive processing and greater effort. Additionally, for more challenging tasks, users tend to exhibit longer average fixation durations.
2	Saccades	Saccade frequency Saccade duration Saccade velocity Saccade amplitude	A higher count of saccades signifies a larger variety of search strategies employed. The magnitude of the saccade is directly influenced by the reduced cognitive effort exerted. This can also be linked to challenges in understanding information. Additionally, the velocity of the saccade corresponds to the speed at which information is processed when transitioning between different elements within a stimulus.
3	Pupil	Pupil diameter	The dilation of a pupil is a reflection of the importance of search results and provides valuable information about decision-making. When pupils are large, it indicates improved performance in detecting relevant information.
4	Blink	Blink frequency Blink duration	Blinking is linked to the processing of information that occurs when we are exposed to stimuli, which then triggers subsequent actions. Individuals who possess adept information analysis skills tend to experience shorter and less frequent blinks. However, it is worth noting that this phenomenon can also manifest in individuals with attention deficit disorders.
5	Scan path		Scan paths are comprised of the sequences of fixations and saccades formed by the trajectory of eye movements over a specific duration. Additionally, the extent of the scan path provides insights into the duration it takes to complete a task.

training. Random Forests find extensive application in diverse fields such as image classification, remote sensing, and bioinformatics.

To optimize overall accuracy and performance, we conducted a random search through the training data 1000 times to identify the best combination of parameters. We assessed the models using accuracy metrics and validated them through 5-fold cross-validation.

In the realm of time series classification, we have employed two highly effective deep learning classifiers called ResNet and MLSTM-FCN. The MLSTM-FCN classifier, initially designed for text neural machine translation, introduces an attention technique that enhances the LSTM's ability to understand long-term relationships by considering both current and previously observed data in context. A context vector, denoted as τ, is formed by adjusting the weights based on the correlation between the elements within the sequence. By evaluating the strength of correlation between the elements of unseen data and those in τ, this vector is utilized to predict unseen data. The combination of MLSTM and FCN in MLSTM-FCN implies a model architecture that benefits from both LSTM's sequential modeling capabilities and FCN's preservation of spatial information. Such an architecture proves valuable in tasks where both temporal and spatial dependencies are crucial, such as certain types of time series analysis, video analysis, or other sequential data tasks.

TABLE 4. Testing hypotheses.

#	Hypothesis	Abbr	Description
1	CAPTCHAs with greater complexity result in longer fixation durations on key Areas of Interest (AOIs) compared to CAPTCHAs with lower complexity	H1	The hypothesis pertains to the level of attention associated with each CAPTCHA model. It suggests that the complexity of a CAPTCHA model directly influences the attention it receives. In other words, the more complex a CAPTCHA model is, the more attention it garners
2	Individuals who possess higher fixation duration on key Areas of Interest (AOIs) tend to exhibit equal or even greater accuracy rates in their responses compared to individuals with lower fixation durations	H2	The hypothesis highlights how the accuracy rate is influenced by an individual's level of attention. It suggests that as attention increases, the accuracy rate also improves.

ResNet [35] versions, initially intended for computer vision, have proven to be effective for time series classification as well. ResNet has also shown remarkable results in activity recognition [36]. ResNet structures have been extensively embraced in multiple computer vision assignments, including image classification, object detection, and segmentation. The incorporation of skip connections and the concept of residual learning have served as a source of inspiration for subsequent architectures, establishing them as a fundamental aspect in the development of deep neural networks for a wide array of applications.

A summary of the common hyperparameters for both time series classification models is presented in Table 5. Both models were implemented with identical hyperparameters as described in the original works. The experimentation was conducted on an NVIDIA TESLA T4 GPU equipped with 16 GB of memory. To prevent overfitting, a 5-fold cross-validation approach was employed, and training was terminated after 100 epochs if there was no improvement in the validation loss. The models were trained to predict whether a user would correctly answer a specific question. The efficiency of the models was evaluated based on the inference time per sample, measured in milliseconds (ms).

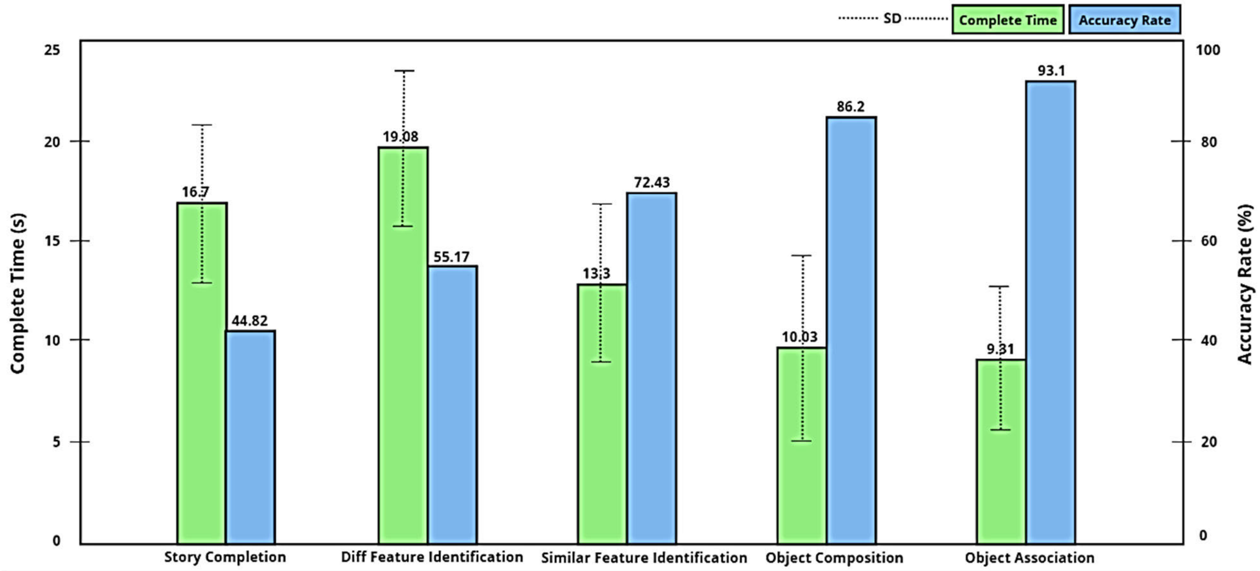

FIGURE 10. Task Performance (Tasks decreased in difficulty from left to right).

In this evaluation, the effectiveness of cognitive models is measured using metrics such as accuracy (Acc), precision (Pr), and recall (Rec). These metrics hold particular importance when addressing problems involving binary classification, where the objective is to categorize instances into either of two classes, namely true answer or false answer:

$$Accuracy = \frac{Predictions}{Correct\ Predictions} \tag{3}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{4}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{5}$$

In our research endeavor, we aim to leverage state-of-the-art algorithms to achieve a more profound comprehension of the data patterns involved in cognitive processes. Nevertheless, it is crucial to emphasize that this study does not revolve around optimizing and refining parameters to improve prediction outcomes or to make comparisons with other algorithms currently in use.

B. PERFORMANCE ANALYSIS

During the analysis, we utilize the mean (M) and standard deviation (SD) of the response time and accuracy to assess and compare the effectiveness of cognitive CAPTCHA models. These metrics offer a concise and comprehensive perspective on performance across various cognitive models. In Figure 10, we can observe the average solving time (M) and standard deviation (SD) of the five cognitive models. These models include story completion (M=16.7s, SD=4.51s), different feature identification (M=19.08s, SD=3.73s), similar feature identification (M=13.3s, SD=4.71s), object composition (M=10.03s, SD=4.49s), and object association (M=9.31s, SD=3.97s). Furthermore, Figure 10 also provides the accuracy rate for each of the cognitive models. The accuracy rates are as follows: story completion (44.82%), different feature identification (55.17%), similar feature identification (72.43%), object composition (86.2%), and object association (93.1%).

According to the ranking in Table 2, the cognitive models of story completion and different feature identification are considered the most challenging. These models have the

TABLE 5. ML hyperparameters.

#	Hyperparameter	Value	Description
1	Initial learning rate	0.001	The most crucial hyperparameter in a neural network is the learning rate. Its value plays a pivotal role in determining various aspects during the network's training process.
2	Optimizer	Adam	The optimizer implements the Adam algorithm.
3	Batch size	128	The batch size refers to the quantity of data samples that are processed before making updates to the model.
4	Maximum epochs	1,000	Epoch is a term that encompasses the entire process of transmitting training data through a machine-learning model. It acts as a hyperparameter, exerting control over the number of iterations employed to train the machine-learning model. model. It is a hyperparameter that controls the number of iterations used to train the machine-learning model.

TABLE 6. Task satisfaction (Tasks decreased in difficulty from left to right).

#	Story completion		Different feature identification		Similar feature identification		Object composition		Object association	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Q1	2.23	0.93	3.14	1.15	3.37	0.96	3.67	1.06	3.54	0.76
Q2	3.17	1.01	3.4	0.86	3.6	0.91	3.51	0.89	3.72	0.95
Q3	3.52	0.95	3.1	0.79	3.79	0.94	3.78	1.23	3.91	1.12

longest average completion times, with story completion taking 16.7 seconds and different feature identification taking 19.08 seconds. However, they also have the lowest accuracy rates, with story completion at 44.82% and different feature identification at 55.17%. On the other hand, the cognitive model of object association is deemed the simplest. It has the shortest completion time and the highest accuracy rate. It is evident that the testing accuracy rate is not solely determined by the completion time. Instead, it depends on an individual's cognitive potential, skills, and knowledge.

The satisfaction results of questionnaires are summarized in Table 6 in terms of:

- (Q1) Visual comfort - is it pleasing to the eye?
- (Q2) Ease of use - is it easy to identify and resolve?
- (Q3) Applicability - is it suitable for the eye-tracking purposes?

To evaluate and compare users' satisfaction with cognitive CAPTCHA models, we employ the average (M) and standard deviation (SD) of the satisfaction. The responses from the questionnaire provide insights into the user's sentiments towards each testing item. In doing so, it captures the user's personal perceptions and emotions. In Table 6, we can observe that elaborate cognitive models tend to make users feel uneasy during the examination, consequently leading to lower outcomes in comparison to simpler ones.

C. EYE FEATURE ANALYSIS

1) BASIC FEATURES

Table 7 displays the actual statistical findings regarding various eye attributes during cognitive tests. The results indicate that higher fixation frequencies observed for the most challenging cognitive models suggest that testers struggle to comprehend the task or encounter difficulties in distinguishing relevant from irrelevant information. Moreover, longer fixation durations generally indicate more profound cognitive processing and increased effort. In the case of complex cognitive models, higher saccade rates indicate the utilization of advanced search strategies, whereas simpler models exhibit higher saccade amplitudes due to lower cognitive effort.

The velocity of saccades directly corresponds to the speed at which information is processed while transitioning between elements within a test. Harder cognitive models demand testers to concentrate intensely to enhance their information processing speed, resulting in higher saccade velocities compared to easier models. Pupil dilation serves as a reflection of the relevance of search results and provides insights into decision-making. Larger pupil sizes are associated with improved detection performance. Testers dealing with more complex cognitive models tend to have larger pupil diameters, enabling better information detection and enhancing their cognition and visualization skills. Blinks are linked to information processing during test exposure, leading to subsequent actions. Harder cognitive models require faster information processing, leading to shorter blink durations.

2) HEAT MAPS

Figures 11–15 depict grayscale heat maps that highlight varying levels of attention (with bolder shades indicating higher attention) for five cognitive models. These models have been sorted from the most challenging to the easiest based on the difficulty scores presented in Table 2. The heat maps are generated by aggregating the eye-gazing data of all participants. When examining each individual model, it is evident that participants' attention is evenly distributed across different Areas of Interest (AOIs). None of the investigated cognitive models show a specific location that consistently attracts more attention than others. Consequently, the pattern observed in the heat maps remains consistent regardless of the difficulty level of the cognitive models.

3) SCAN PATHS

The grayscale eye-tracking scan paths, illustrated in Figures 16-20, portray the collective eye movements of all participants. These pathways are ranked based on the level of difficulty, as indicated in Table 2. It is crucial not only to comprehend the distribution of attention through heat maps but also to understand the trajectory of eye movements from one location to another. Scan paths are patterns formed by fixations and saccades, representing the

TABLE 7. Eye performance with sample size = 29 (Tasks decreased in difficulty from left to right).

Feature	Story completion		Different feature identification		Similar feature identification		Object composition		Object association	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fixation frequency (count/s)	2.81	0.47	2.5	0.51	2.23	0.43	2.35	0.53	2.24	0.54
Fixation duration (ms)	264.13	13.5	254.43	16.06	249.3	13.3	243.6	15.97	242.3	11.42
Saccade frequency (count/s)	3.41	2.34	3.25	2.55	2.71	1.73	1.89	0.85	1.76	0.92
Saccade amplitude (degree)	13.12	0.93	13.46	1.45	14.06	1.29	15.67	0.89	15.94	1.15
Saccade velocity (degree/s)	261.71	103.74	267.78	116.42	146.32	68.53	132.54	75.26	157.68	92.35
Saccade duration (ms)	19.16	3.51	20.23	4.59	15.33	3.27	15.76	2.69	14.37	2.51
Blink frequency (count/s)	0.19	0.17	0.17	0.21	0.07	0.05	0.05	0.03	0.09	0.04
Blink duration (ms)	209.23	32.97	202.29	40.48	225.54	48.35	220.31	47.53	223.54	51.26
Pupil diameter (mm)	3.82	0.71	3.71	0.58	3.49	0.64	3.56	0.54	3.61	0.73

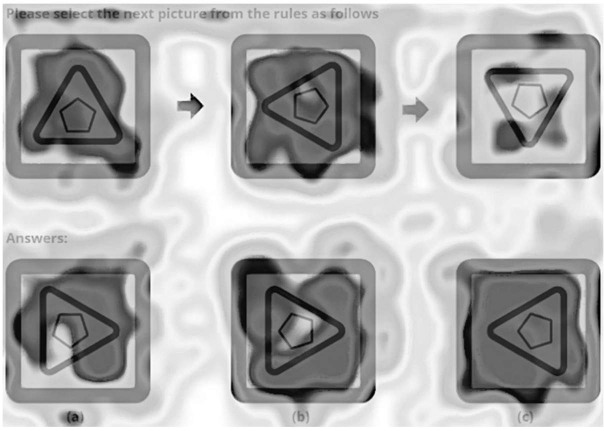

FIGURE 11. Heat map of story completion.

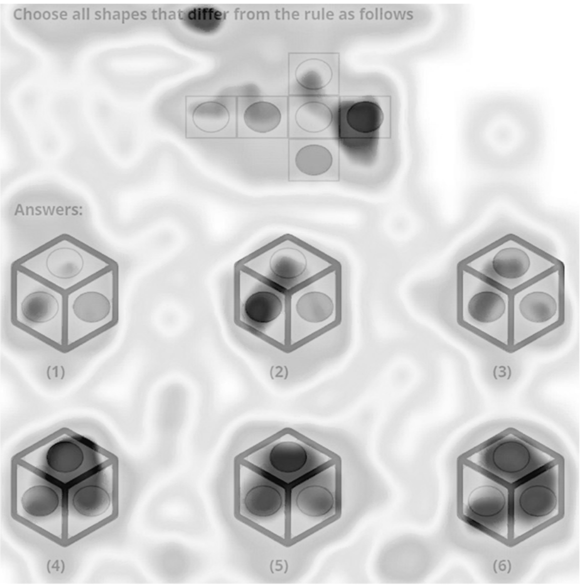

FIGURE 12. Heat map of different feature identification.

path of eye movements over time. The diameter of the fixation circles in the scan path provides insights into the duration of task fixations. A larger circle diameter indicates increased attentiveness from the tester. The participants' eye orientations are randomly distributed across all feasible locations in all cognitive models. Complex cognitive models exhibit a greater number of fixation circles compared to simpler ones. Additionally, we observe that

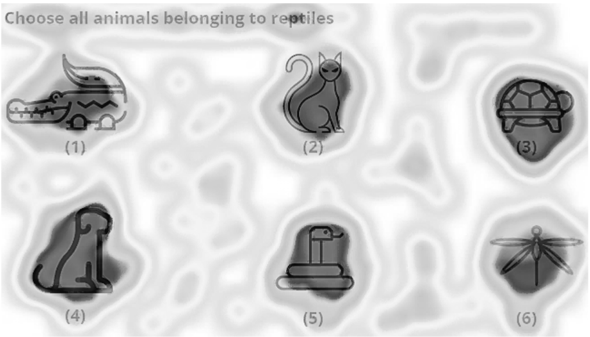

FIGURE 13. Heat map of similar feature identification.

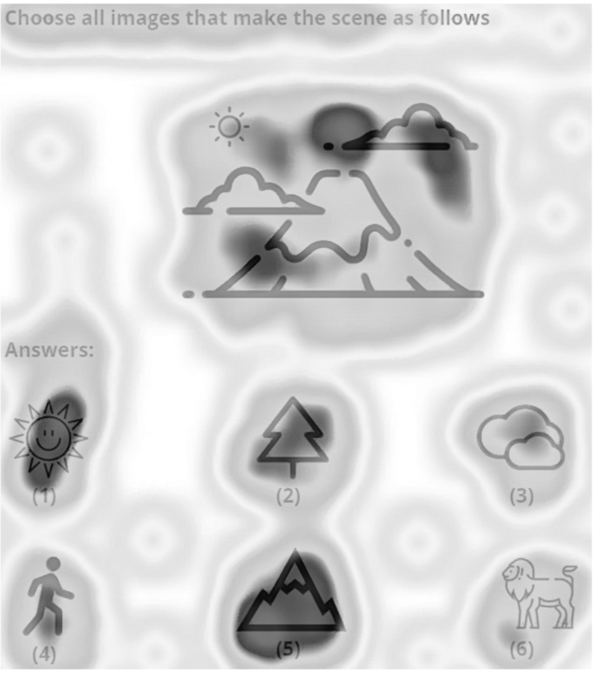

FIGURE 14. Heat map of object composition.

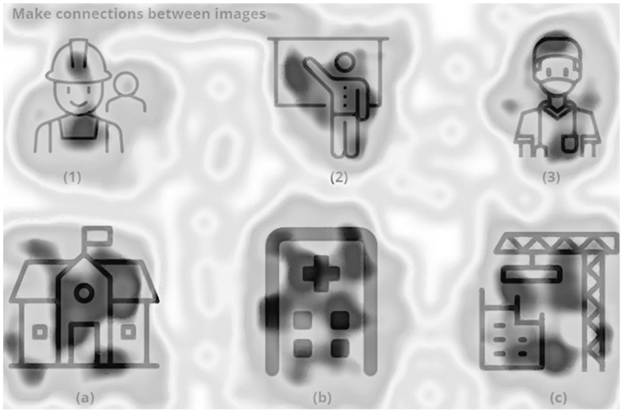

FIGURE 15. Heat map of object association.

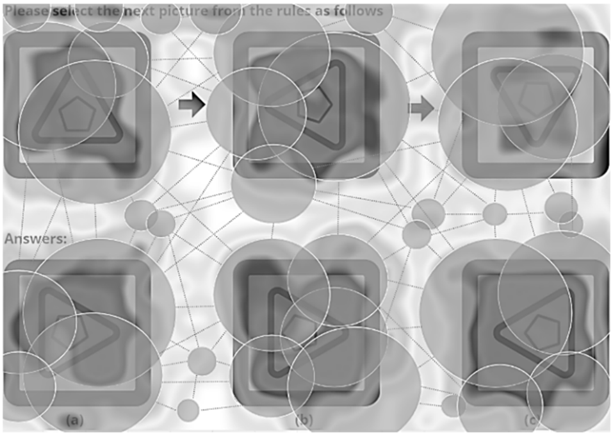

FIGURE 16. Scan path of story completion.

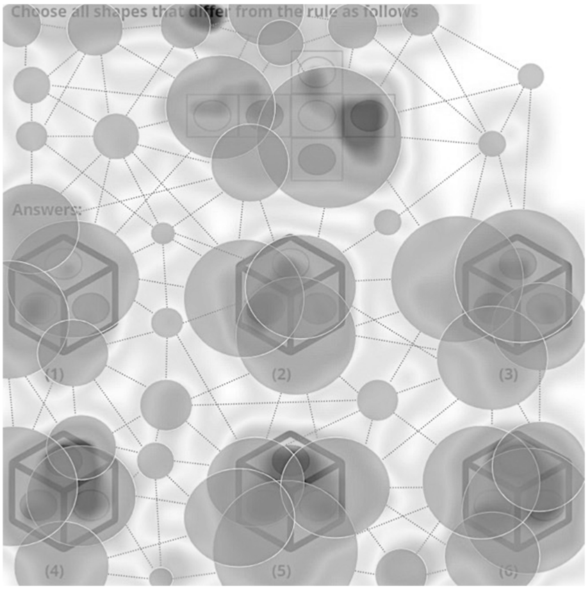

FIGURE 17. Scan path of different feature identification.

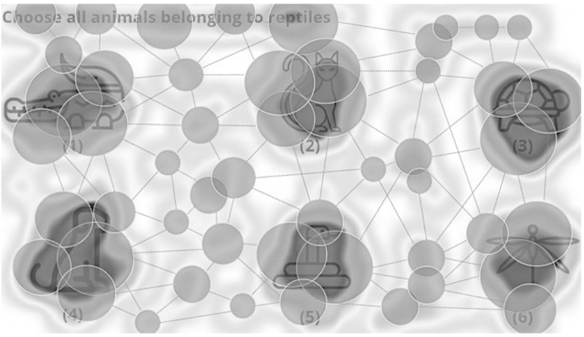

FIGURE 18. Scan path of similar feature identification.

more sophisticated cognitive models have longer fixation durations compared to simpler models. Consequently, there are instances where saccadic trails move back and forth within a particular region, leading to revisits. The scan route also demonstrates that complex cognitive models result in a higher frequency of revisits compared to simpler models. Therefore, users need to allocate more attention to higher-level cognitive models and maintain their gaze on essential areas of interest (AOIs) to gather crucial information that supports their decision-making process.

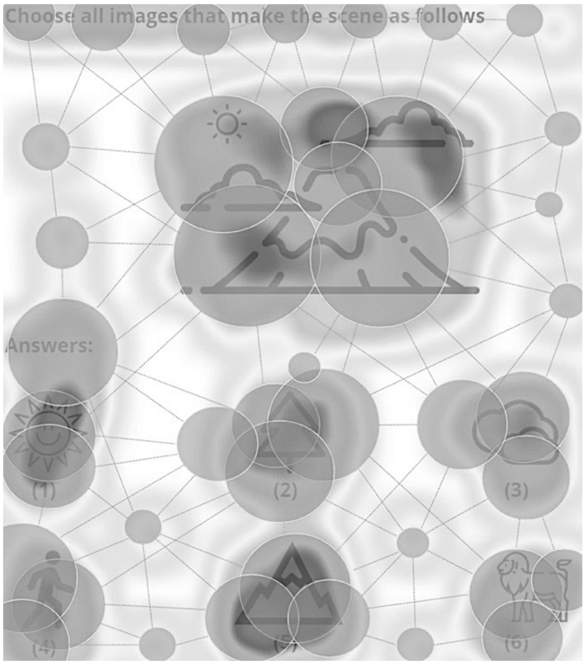

FIGURE 19. Scan path of object composition.

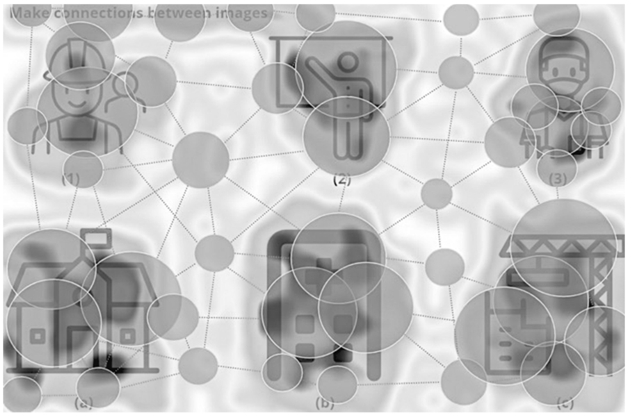

FIGURE 20. Scan path of object association.

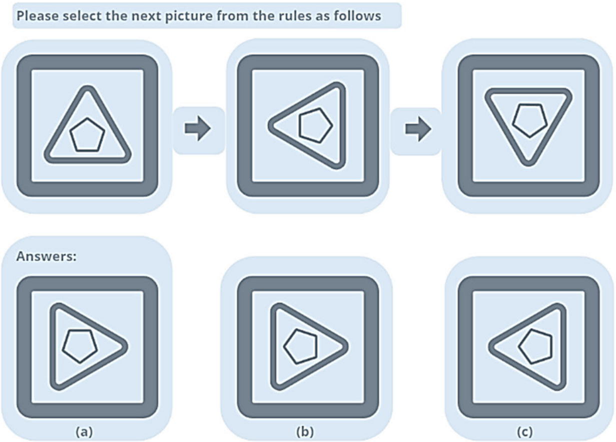

FIGURE 21. Key AOIs of story completion.

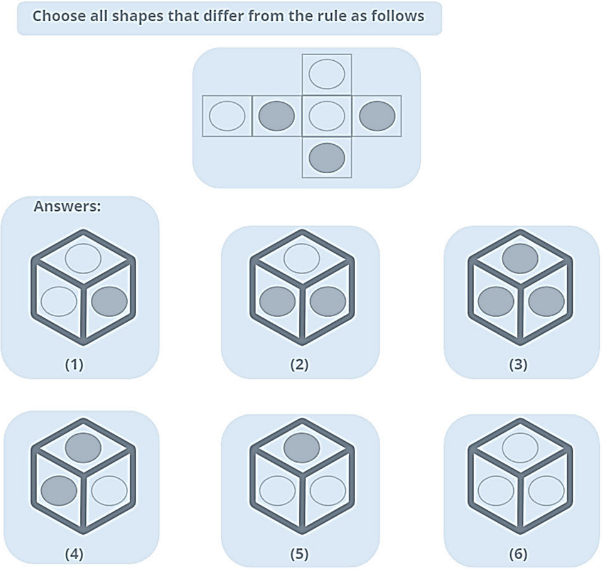

FIGURE 22. Key AOIs of different feature identification.

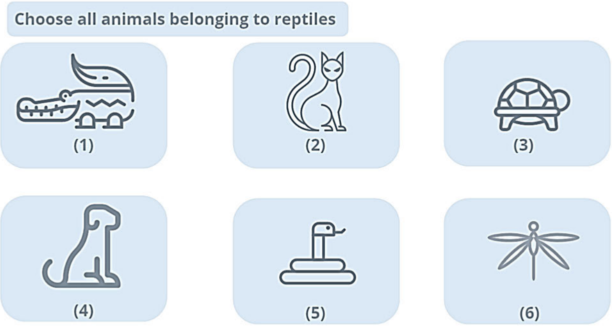

FIGURE 23. Key AOIs of similar feature identification.

D. HYPOTHESIS STATISTICAL ANALYSIS

In this study, we aim to validate our hypothesis regarding the level of attention exhibited by individuals. This validation is presented in Table 4. To investigate these assumptions, we analyze the experimental data using statistical techniques such as ANOVA (Analysis of Variance) and T-test. Drawing on the insights gained from Section IV-C, we can identify the specific areas of interest (AOIs) that users tend to focus on when seeking crucial information to overcome challenges. These findings are visually represented in Figures 21-25. Given the limited sample size (size < 50), it is necessary to verify the normality of the eye-tracking measures employed. In this hypothesis testing, we employ the metric of fixation lengths observed in key AOIs to demonstrate that individuals with higher levels of attention perform better in the test.

1) NORMALITY TEST

We have gathered data on the duration of fixations on specific Areas of Interest (AOIs), as illustrated in Figure 21-25. From this data, we computed the average (mean) and variability (standard deviation) values. Upon examining the significance

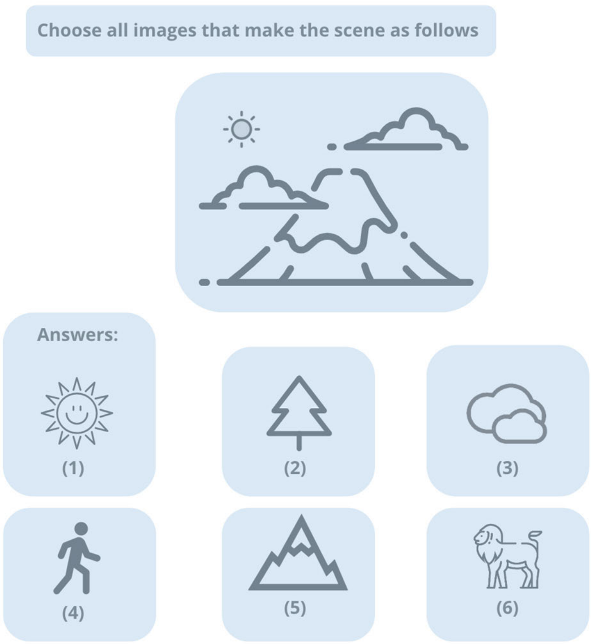

FIGURE 24. Key AOIs of object composition.

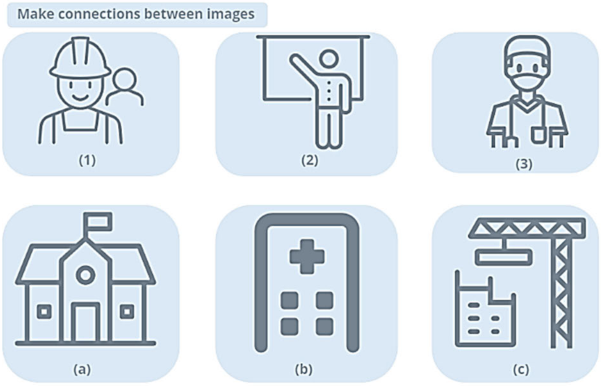

FIGURE 25. Key AOIs of object association.

TABLE 8. Shapiro-Wilk Test result of fixation durations on AOIs (size = 29, $\alpha = 0.05$, unit = ms).

Cognitive model	Mean	Standard Deviation (SD)	Statistic (W)	Degree of Freedom (DF)	Significance (P)
Story completion	264.13	11.15	0.932	29	0.106
Different feature identification	254.43	13.57	0.934	29	0.107
Similar feature identification	249.3	13.61	0.96	29	0.441
Object composition	243.6	14.35	0.957	29	0.385
Object association	242.3	11.74	0.951	29	0.283

TABLE 9. One-tailed t-test's t-values comparison (size = 29, $\alpha = 0.05$).

#	M1	M2	M3	M4	M5
M1		2.04	4.6	6.9	8.35
M2			2.35	4.62	5.54
M3				2.32	3.01
M4					**0.43**
M5					

(**M1**: Story completion, **M2**: Different feature identification, **M3**: Similar feature identification, **M4**: Object composition, **M5**: Object association)

values presented in Table 8, we observed that they exceed 0.05. Based on this evidence, we can infer that the fixation durations on key AOIs follow a normal distribution pattern.

2) HYPOTHESIS TEST

The independent t-test serves the purpose of comparing the average values among different groups under examination. To determine whether the mean of one population is greater or smaller than the other, we employ one-tailed t-tests. The calculation of the t-value is carried out in the following manner:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}} \tag{6}$$

where $\bar{x}_1$ and $\bar{x}_2$ are testing groups' fixation duration means, SD_1 and SD_2 are testing groups' standard deviations of fixation durations, and n_1 and n_2 are testing groups' sample sizes.

a: H1 HYPOTHESIS ANALYSIS

This hypothesis explores whether more intricate CAPTCHAs elicit longer fixation durations on key AOIs (Areas of Interest) in comparison to less complex ones. In Table 9, the t-values obtained from cognitive models, calculated using Formula 6, are presented for comparison. The majority of t-values surpass the critical value (t-value = 1.676,

TABLE 10. Two score groups' t-value.

SMD	Higher score group				Lower score group				t-value
	N	SM	FM	FSD	N	SM	FM	FSD	
2	12	3.75	257.71	12.46	17	1.33	247.29	13.58	2.13

(**SMD**: Score Median, **N**: size, **SM**: Score Mean, **FM**: Fixation Duration Mean, **FSD**: Fixation Duration Standard Deviation)

DF = 50, $\alpha = 0.05$), except for the t-value associated with the model of object composition with object association. Consequently, the hypothesis holds true for highly complex cognitive models that demand users to allocate more attention to solve them. On the other hand, it can be argued that simple cognitive models do not effectively differentiate between individuals' attention and effort in problem-solving, as there is not much discrepancy in fixation durations among individuals.

b: H2 HYPOTHESIS ANALYSIS

This hypothesis aims to examine whether individuals who have longer periods of fixation show similar or higher levels of accuracy in response compared to individuals with shorter periods of fixation on specific areas of interest (AOIs). As illustrated in Table 10, the participants were divided into two groups based on their scores, with the median score (score median = 2) serving as the dividing point. The lower score group consisted of 12 participants, while the higher score group had 17 individuals. By utilizing Formula 6, the calculated t-value exceeded the critical value (t-value = 1.703, DF = 27, $\alpha = 0.05$). Consequently, we can confirm this hypothesis.

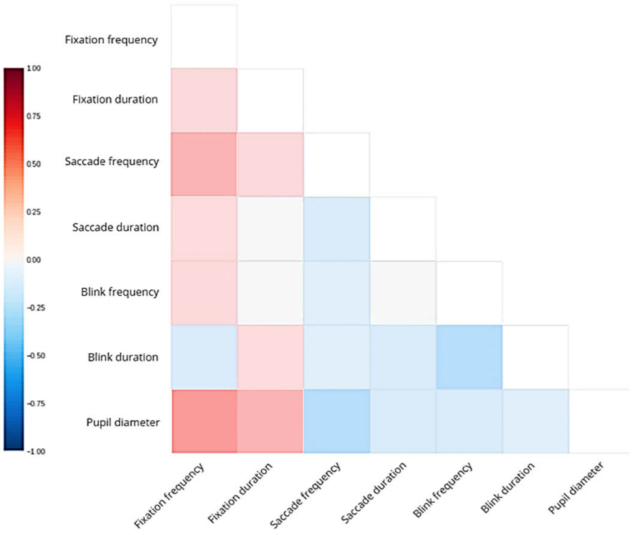

FIGURE 26. Feature correlation.

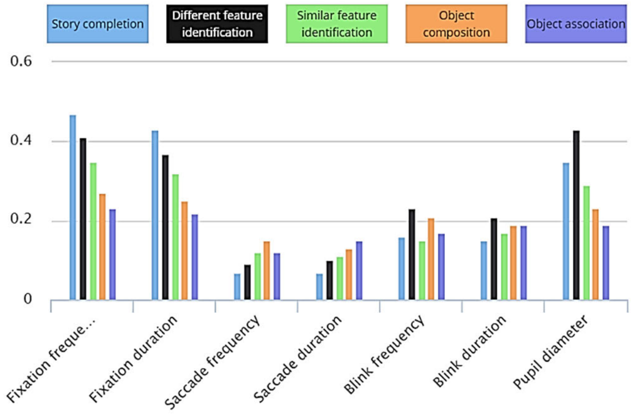

FIGURE 27. Feature importance.

E. MACHINE LEARNING ANALYSIS

1) FEATURE ENGINEERING

Correlation helps us uncover patterns in data by using the relationship between different features. Figure 26 illustrates the correlation among the features in the eye-tracking data. The fixation feature shows positive associations with the pupil, saccade, and blink features. When it comes to predicting performance, not only the fixation feature but also these highly correlated features with the variable fixation seem to be suitable options as exploratory variables in simple linear regression models.

In accordance with Figure 27, we assessed the significance of characteristics in forecasting performance using the Random Forest Classifier, mentioned in Section IV-A3.b. Alongside fixation features, pupil features also contribute significantly to performance prediction.

2) PERFORMANCE PREDICTION

Following the details provided in Section IV-A3.b, during the classification phase, we conducted training and testing on three classifiers utilizing the features chosen in Section IV-A2. These classifiers include Logistic Regression (LR), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM). These classifiers have proven to be exceptionally effective in handling continuous and normally distributed data, as well as data that demonstrates an approximately linear relationship between its features and the desired outcome. The effectiveness of these classifiers can be attributed to their simplicity, computational efficiency, and their ability to produce satisfactory results in real-world scenarios. Table 11 reveals that, interestingly, the more complex cognitive models seem to have inferior performance compared to the simpler ones. This suggests that attention is not the sole factor influencing performance; the abilities and knowledge of each individual also play a significant role.

3) TIME SERIES ANALYSIS

This methodology examines a sequence of data points collected over a period of time and predicts desired values solely based on a known history of target values. It is a

TABLE 11. Accuracy performance (Acc: Accuracy, Pr: Precision, Rec: Recall).

ML	Story completion			Different feature identification			Similar feature identification			Object composition			Object association		
	Acc	Pr	Rec	Acc	Pr	Rec	Acc	Pr	Rec	Acc	Pr	Rec	Acc	Pr	Rec
GNB	0.52	0.52	0.54	0.61	0.59	0.59	0.81	0.79	0.79	0.81	0.81	0.81	0.83	0.84	0.84
SVM	0.63	0.62	0.64	0.65	0.66	0.65	0.86	0.85	0.87	0.84	0.83	0.83	0.85	0.86	0.85
LR	0.51	0.52	0.51	0.59	0.58	0.58	0.79	0.79	0.79	0.81	0.79	0.79	0.81	0.82	0.81

TABLE 12. Full-length performance (Acc: Accuracy, Pr: Precision, Rec: Recall).

ML	Story completion			Different feature identification			Similar feature identification			Object composition			Object association		
	Acc	Pr	Rec	Acc	Pr	Rec	Acc	Pr	Rec	Acc	Pr	Rec	Acc	Pr	Rec
MLSTM-FCN	0.79	0.78	0.79	0.81	0.82	0.81	0.82	0.79	0.81	0.84	0.81	0.83	0.83	0.82	0.83
ResNet	0.65	0.64	0.63	0.68	0.66	0.67	0.67	0.65	0.66	0.69	0.67	0.68	0.71	0.69	0.72

specific type of regression referred to as an auto-regressive model in the literature. In time series analysis, data points are consistently captured at regular intervals over a fixed duration of time, rather than sporadically or randomly. For this study, as mentioned in Section IV-A3.b, we employed two advanced deep learning classifiers, MLSTM-FCN and ResNet.

The fusion of MLSTM and FCN within the MLSTM-FCN framework presents a model design that leverages the strengths of LSTM's ability to model sequences and FCN's effectiveness in retaining spatial information. This architecture proves particularly advantageous in tasks that require the consideration of both temporal and spatial dependencies. Examples of such tasks include certain forms of time series analysis, video analysis, and other sequential data tasks. ResNet, originally designed for computer vision tasks, has demonstrated its efficacy in time series classification as well. ResNet architectures have gained widespread acceptance in various computer vision projects, encompassing image classification, object detection, and segmentation. The implementation of skip connections and the concept of residual learning have not only inspired subsequent architectures but also established them as a vital component in the advancement of deep neural networks for a diverse range of applications.

These algorithms have been trained using 10-second time sequences to anticipate whether a user will correctly answer a particular question. The efficiency of the models is evaluated in terms of milliseconds of inference time per sample (ms).

a: FULL-LENGTH ANALYSIS

The comprehensive outcome of the performance yields a standard against which the interval performances can be assessed. The outcomes, as illustrated in Table 12, align with the discoveries mentioned in Section IV-E2, indicating that more intricate cognitive models tend to underperform simpler ones. Additionally, it is worth noting that ResNet exhibits a short inference time of 4 milliseconds, whereas MLSTM-FCN showcases the lengthiest inference time of 105 milliseconds per sample.

b: TIME INTERVAL ANALYSIS

The complete sequence is divided into various lengths, ranging from 1 to 10 seconds, to assess how well the classifiers perform on shorter time intervals. As depicted in Table 13, MLSTM-FCN exhibits a maximum accuracy drop of approximately 11%, while ResNet experiences a maximum accuracy decrease of around 23.5% compared to its performance on the full-length sequence. When considering the time taken for each sample in the full-length sequence, MLSTM-FCN's inference time increases linearly by approximately 175 ms for every additional second of sequence length. On the other hand, ResNet's inference time

TABLE 13. Interval performance (Acc: Accuracy, Pr: Precision, Rec: Recall).

Cognitive model	ML	Baseline			Time sequence 1			Time sequence 2			Time sequence 3		
		Acc	Pr	Rec	Acc	Pr	Rec	Acc	Pr	Rec	Acc	Pr	Rec
Story completion	MLSTM-FCN	0.79	0.78	0.79	0.71	0.72	0.71	0.73	0.72	0.73	0.75	0.74	0.74
	ResNet	0.65	0.64	0.63	0.51	0.52	0.51	0.57	0.58	0.57	0.63	0.62	0.62
Different feature identification	MLSTM-FCN	0.81	0.82	0.81	0.73	0.72	0.72	0.77	0.76	0.76	0.79	0.81	0.81
	ResNet	0.68	0.66	0.67	0.52	0.51	0.52	0.58	0.59	0.59	0.67	0.68	0.67
Similar feature identification	MLSTM-FCN	0.82	0.79	0.81	0.73	0.74	0.75	0.79	0.79	0.79	0.83	0.82	0.82
	ResNet	0.67	0.65	0.66	0.54	0.55	0.53	0.59	0.61	0.61	0.65	0.66	0.67
Object composition	MLSTM-FCN	0.84	0.81	0.83	0.77	0.78	0.77	0.82	0.81	0.83	0.81	0.82	0.82
	ResNet	0.69	0.67	0.68	0.61	0.62	0.61	0.67	0.66	0.67	0.69	0.68	0.69
Object association	MLSTM-FCN	0.83	0.82	0.83	0.78	0.77	0.77	0.82	0.81	0.83	0.83	0.83	0.83
	ResNet	0.71	0.69	0.72	0.64	0.63	0.62	0.67	0.68	0.69	0.72	0.71	0.72

remains constant across different time intervals. Additionally, the results indicate that cognitive models demonstrate superior performance in later time sequences. Moreover, for simple cognitive models, they exhibit better performance in earlier time sequences compared to more complex models.

V. CONCLUSION

In this research study, we performed eye-tracking experiments on participants to analyze how their visual engagement with cognitive CAPTCHAs changes as the difficulty level varies. Our objective was to evaluate users' visual activity by examining their eye movement performance and process metrics. To accomplish this, we utilized statistical techniques such as ANOVA (Analysis of Variance) and T-test to validate our proposed attention theories in assessing the relationship between individuals' attention and performance. Additionally, we employed Machine Learning (ML) to train the collected data and explore the correlation between these factors in predicting cognitive function.

Based on the findings of our experiment, we observed that the most challenging cognitive models took the longest time, on average, to complete but had the lowest accuracy rates. Conversely, the simplest cognitive model had the shortest completion time but the highest accuracy rate. Complex cognitive models exhibited higher saccade rates, saccade velocities, larger pupil diameters, and shorter blinks of shorter duration, which enhanced search strategies and processing speed. On the other hand, simpler models had higher amplitudes, indicating less cognitive effort.

The attention distribution provided by the heatmap did not display any preference for a specific site in any of the analyzed cognitive models. This consistent finding was observed across both difficult and simple cognitive models. Furthermore, participants exhibited random eye-gazing orientations along the scan route, covering all viable areas in all cognitive models. The scan path also revealed that more complex cognitive models resulted in larger fixation circles and more revisits compared to simpler models.

To investigate the impact of CAPTCHA complexity on attention on key Areas of Interest (AOIs), we conducted hypothesis testing. The hypothesis was only supported for more complex cognitive models, which required greater user attention to resolve. Additionally, it was confirmed that individuals with longer fixation durations on key AOIs achieved similar or higher accuracy rates compared to those with shorter fixation durations.

In predicting performance, fixation characteristics showed positive associations with pupil, saccade, and blink features. Pupil features also played a significant role in performance prediction, alongside fixation characteristics. Besides, through time series analysis, we found that cognitive models performed better in later time sequences. However, in the early time sequences, easier cognitive models outperformed more complex models.

The study highlights certain limitations that should be considered when interpreting the findings. The short duration of the tests may not capture significant changes in gaze patterns over time, and the average eye characteristics may be influenced by outliers, such as familiarity with the test method or discomfort felt during observation. Additionally, the size of the stimuli might impact the results, as smaller images could potentially enhance cognitive function and affect visual search. The small size of the participant pool also poses a challenge, influencing training and overall prediction results.

The findings of this investigation underscore the potential of utilizing eye-tracking methodology to evaluate attention and performance in solving cognitive CAPTCHA models. By employing advanced techniques such as hypothesis testing and machine learning, we can gain a deeper understanding of cognitive processes through the analysis of eye metrics and visual behavior. Subsequent research endeavors can further enhance these techniques by incorporating advanced hypotheses and algorithms that are applicable in real-world scenarios. This study particularly highlights the opportunity to enhance any cognitive CAPTCHA model by comprehending the inner workings of cognitive processes, including distinguishing between less effective and more effective Areas of Interest (AOIs). The removal of less effective AOIs can save users' time, while improving the more effective AOIs can enhance users' proficiency in solving CAPTCHAs. This advancement could potentially pave the way for the development of a decision-making system that evaluates the effectiveness and efficiency of specific cognitive CAPTCHA models, providing guidance for CAPTCHA design.

ABBREVIATIONS

#	Abbr	Description
1	CAPTCHA	Automated Public Turing Test toTell Computers and Humans Apart
2	HCI	Human Computer Interaction
3	ML	Machine Learning
4	VLAT	Visualization Literacy Assessment Test
5	CVR	Content Validity Ratio
6	ANOVA	Analysis of Variance
7	POI	Point of Interest
8	SVM	Support Vector Machine
9	NASA-TLX	NASA Task Load Index
10	CTT	Classical Test Theory
11	DNA	Non-Disclosure Agreement
12	FIR	Finite Impulse Response
13	I-VT	Velocity Threshold Identification
14	SD	Standard Deviation
15	GNB	Gaussian Naive Bayes
16	LR	Logistic Regression
17	MLSTM-FCN	Multivariate Long Short-Term Memory Full Convolutional Networks
18	LSTM	Long Short-Term Memory
19	AOI	Area of Interest
20	ResNet	Residual Neural Network
21	RFC	Random Forest Classifier

DISCLOSE STATEMENT

No potential conflict of interest was reported by the author(s).

REFERENCES

[1] N. Dinh and L. Ogiela, "Human-artificial intelligence approaches for secure analysis in CAPTCHA codes," *EURASIP J. Inf. Secur.*, vol. 2022, no. 8, 2022, doi: 10.1186/s13635-022-00134-9.

[2] N. D. Trong, T. H. Huong, and V. T. Hoang, "New cognitive deep-learning CAPTCHA," *Sensors*, vol. 23, no. 4, p. 2338, Feb. 2023.

[3] S. Lee, S.-H. Kim, and B. C. Kwon, "VLAT: Development of a visualization literacy assessment test," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 551–560, Jan. 2017.

[4] J. Pascual-Leone, "A mathematical model for the transition rule in piaget's developmental stages," *Acta Psychologica*, vol. 32, pp. 301–345, Jan. 1970.

[5] D. C. Richardson, R. Dale, and M. J. Spivey, "Eye movements in language and cognition," *Methods Cogn. Linguist.*, vol. 18, pp. 323–344, Jun. 2007.

[6] M. Borys, S. Barakate, K. Hachmoud, M. Plechawska-Wòjcik, P. Krukow, and M. Kamiński, "Classification of user performance in the ruff figural fluency test based on eye-tracking features," in *Proc. ITM Web Conf.*, vol. 15, 2017, p. 02002.

[7] M. S. Vendetti, A. Starr, E. L. Johnson, K. Modavi, and S. A. Bunge, "Eye movements reveal optimal strategies for analogical reasoning," *Frontiers Psychol.*, vol. 8, p. 932, Jun. 2017.

[8] E. Davis and L. Morgenstern, "Introduction: Progress in formal commonsense reasoning," *Artif. Intell.*, vol. 153, nos. 1–2, pp. 1–12, Mar. 2004.

[9] C. H. Lawshe, "A quantitative approach to content validity," *Personnel Psychol.*, vol. 28, no. 4, pp. 563–575, Dec. 1975.

[10] D. E. Kieras and D. E. Meyer, "An overview of the EPIC architecture for cognition and performance with application to human-computer interaction," *Hum.–Comput. Interact.*, vol. 12, no. 4, pp. 391–438, Dec. 1997.

[11] *Tobii Eye Tracker*. Accessed: May 21, 2023. [Online]. Available: https://connect.tobii.com

[12] *EyeTribe*. Accessed: 21 May 2023. [Online]. Available: https://theeyetribe.com

[13] *EyeLink*. Accessed: May 21, 2023. [Online]. Available: https://www.sr-research.com

[14] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behav. Res. Methods, Instrum., Comput.*, vol. 34, no. 4, pp. 455–470, Nov. 2002.

[15] L. Rozenblit, M. Spivey, and J. Wojslawowicz, "Mechanical reasoning about gear-and-belt diagrams: Do eye movements predict performance? Diagrammatic representation and reasoning," *Diagrammatic Represent. Reasoning*. London, U.K.: Springer, 2002, pp. 223–240.

[16] S.-C. Chen, H.-C. She, M.-H. Chuang, J.-Y. Wu, J.-L. Tsai, and T.-P. Jung, "Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities," *Comput. Educ.*, vol. 74, pp. 61–72, May 2014.

[17] V. Bachurina, S. Sushchinskaya, M. Sharaev, E. Burnaev, and M. Arsalidou, "A machine learning investigation of factors that contribute to predicting cognitive performance: Difficulty level, reaction time and eye-movements," *Decis. Support Syst.*, vol. 155, Apr. 2022, Art. no. 113713.

[18] S. Lee, D. Hooshyar, H. Ji, K. Nam, and H. Lim, "Mining biometric data to predict programmer expertise and task difficulty," *Cluster Comput.*, vol. 21, no. 1, pp. 1097–1107, Mar. 2018.

[19] H. S. Al-Khalifa, "An empirical pilot study of CAPTCHA complexity using eye tracking," in *Proc. 16th Int. Conf. Inf. Integr. Web-Based Appl. Services*, New York, NY, USA, Dec. 2014, pp. 175–179, doi: 10.1145/2684200.2684330.

[20] M. R. Ogiela and L. Ogiela, "Eye tracking solutions in cognitive CAPTCHA authentication," in *Proc. 12th Int. Conf. Comput. Autom. Eng.*, New York, NY, USA, Feb. 2020, pp. 173–176, doi: 10.1145/3384613.3384655.

[21] E. Ktistakis, V. Skaramagkas, D. Manousos, N. S. Tachos, E. Tripoliti, D. I. Fotiadis, and M. Tsiknakis, "OLET: A dataset for cognitive workload estimation based on eye-tracking. Computer methods and programs in biomedicine," *Comput. Methods Programs Biomed.*, vol. 224, Sep. 2022, Art. no. 106989.

[22] R. P. McDonald, *Test Theory: A Unified Treatment*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1999.

[23] R. L. Thorndike and E. Hagen, *Measurement and Evaluation in Psychology and Education*. London, U.K.: Pearson, 2010.

[24] *Educational Assessment*. Accessed: May 21, 2023. [Online]. Available: http://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis

[25] S. Mejia-Romero, J. E. Lugo, D. Bernardin, and J. Faubert, "An effective filtering process for the noise suppression in eye movement signals," in *Proc. Int. Conf. Data Sci. Appl.*, vol. 148, 2021, pp. 33–46.

[26] A. T. Duchowski, *Eye Movement Analysis. In: Eye Tracking Methodology: Theory and Practice*. Cham, Switzerland: Springer, 2003.

[27] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, New York, NY, USA, 2000, pp. 71–78.

[28] K. Rayner, "Eye movements and attention in reading, scene perception, and visual search," *Quart. J. Exp. Psychol.*, vol. 62, no. 8, pp. 1457–1506, Aug. 2009.

[29] B. Steichen, C. Conati, and G. Carenini, "Inferring visualization task properties, user performance, and user cognitive abilities from eye gaze data," *ACM Trans. Interact. Intell. Syst.*, vol. 4, no. 2, pp. 1–29, Jul. 2014.

[30] C. Strauch, L. Greiter, and A. Huckauf, "Pupil dilation but not microsaccade rate robustly reveals decision formation," *Sci. Rep.*, vol. 8, no. 1, p. 13165, Sep. 2018.

[31] F. T. P. Oliveira, A. Aula, and D. M. Russell, "Discriminating the relevance of web search results with measures of pupil size," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2009, pp. 2209–2212.

[32] L. L. Di Stasi, R. Renner, P. Staehr, J. R. Helmert, B. M. Velichkovsky, J. J. Cañas, A. Catena, and S. Pannasch, "Saccadic peak velocity sensitivity to variations in mental workload," *Aviation, Space, Environ. Med.*, vol. 81, no. 4, pp. 413–417, Apr. 2010.

[33] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for time series classification," *Neural Netw.*, vol. 116, pp. 237–245, Aug. 2019.

[34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015.

[35] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1578–1585.

[36] J. Lu and K.-Y. Tong, "Robust single accelerometer-based activity recognition using modified recurrence plot," *IEEE Sensors J.*, vol. 19, no. 15, pp. 6317–6324, Aug. 2019.

NGHIA DINH received the M.Sc. degree in software engineering from Bordeaux University, France. He is currently pursuing the Ph.D. degree with the VSB—Technical University of Ostrava, Czech Republic. He is a Software Architecture Enthusiast and a Computer Scientist. He has contributed to the success of many open sources and technology companies.

LIDIA DOMINIKA OGIELA received the Habilitation degree in computer science from the Faculty of Electrical Engineering and Computer Science, VŠB – Technical University of Ostrava, Czech Republic, in 2016, and the Doctor degree in computer science and telecommunication from Hosei University, Tokyo, Japan, in 2018, for her thesis and research on human-centered computing for future-generation computer systems. In 2022 she received Professor Title in Computer Science. She is a Computer Scientist, Mathematician, and Economist. She is an author of more than 240 scientific international publications on information systems, cognitive analysis techniques, and computational intelligence methods. She is a member of prestigious international scientific societies as well as: Fellow Lifetime Member of SPIE, Member of SIAM, ACM, OSA, CSS, and IPSJ Information Processing Society of Japan. In 2005, she was awarded the title of Doctor of Computer Science and Engineering at the Faculty of Electrical, Automatic Control, Computer Science and Electronic Engineering of the AGH University of Science and Technology, for her thesis and research on cognitive informatics and its application in intelligent information systems.

KIET TRAN-TRUNG received the master's degree from Ho Chi Minh City Pedagogical University, Vietnam. He is currently a Lecturer with Ho Chi Minh City Open University. His research interests include machine learning and computer vision.

TUAN LE-VIET received the master's degree from Ho Chi Minh City University of Science, Vietnam, and the Ph.D. degree from Sejong University, South Korea. He is currently a Lecturer with Ho Chi Minh City Open University. His research interests include deep learning and computer vision.

VINH TRUONG HOANG (Graduate Student Member, IEEE) received the master's degree from the University of Montpellier, in 2009, and the Ph.D. degree in computer science from the University of the Littoral Opal Coast, France. He is currently an Assistant Professor and the Head of the Image Processing and Computer Graphics Department, Ho Chi Minh City Open University, Vietnam. His research interests include image analysis and feature selection.