Can We Empower Attentive E-reading with a Social Robot? An Introductory Study with a Novel Multimodal Dataset and Deep Learning Approaches

Lee, Yoon; Specht, Marcus

DOI
10.1145/3576050.3576122

Publication date
2023

Document Version
Final published version

Published in
LAK 2023 Conference Proceedings - Towards Trustworthy Learning Analytics - 13th International Conference on Learning Analytics and Knowledge

Citation (APA)
Lee, Y., & Specht, M. (2023). Can We Empower Attentive E-reading with a Social Robot? An Introductory Study with a Novel Multimodal Dataset and Deep Learning Approaches. In *LAK 2023 Conference Proceedings - Towards Trustworthy Learning Analytics - 13th International Conference on Learning Analytics and Knowledge* (pp. 520-530). (ACM International Conference Proceeding Series). ACM. https://doi.org/10.1145/3576050.3576122

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Can We Empower Attentive E-reading with a Social Robot? An Introductory Study with a Novel Multimodal Dataset and Deep Learning Approaches

Yoon Lee
Delft University of Technology, Delft, Netherlands
Leiden-Delft-Erasmus Centre for Education and Learning, Delft, Netherlands
y.lee@tudelft.nl

Marcus Specht
Delft University of Technology, Delft, Netherlands
Leiden-Delft-Erasmus Centre for Education and Learning, Delft, Netherlands
m.m.specht@tudelft.nl

ABSTRACT

Reading on digital devices has become more commonplace, while it often poses challenges to learners' attention. In this study, we hypothesized that allowing learners to reflect on their reading phases with an empathic social robot companion might enhance learners' attention in e-reading. To verify our assumption, we collected a novel dataset (SKEP) in an e-reading setting with social robot support. It contains 25 multimodal features from various sensors and logged data that are direct and indirect cues of attention. Based on the SKEP dataset, we comprehensively compared the difference between HRI-based (treatment) and GUI-based (control) feedback and obtained insights for intervention design. Based on the human annotation of the nearly 40 hours of video data streams from 60 subjects, we developed a machine learning model to capture attention-regulation behaviors in e-reading. We exploited a two-stage framework to recognize learners' observable self-regulatory behaviors and conducted attention analysis. The proposed system showed a promising performance with high prediction results of e-reading with HRI, such as 72.97% accuracy in recognizing attention regulation behaviors, 74.29% accuracy in predicting knowledge gain, 75.00% for perceived interaction experience, and 75.00% for perceived social presence. We believe our work can inspire the future design of HRI-based e-reading and its analysis.

CCS CONCEPTS

• **Applied computing** → **E-learning**; *Interactive learning environments.*

KEYWORDS

Novel dataset, Attention Self-regulation, E-reading, Human-Robot Interaction, Deep Learning

ACM Reference Format:
Yoon Lee and Marcus Specht. 2023. Can We Empower Attentive E-reading with a Social Robot? An Introductory Study with a Novel Multimodal Dataset and Deep Learning Approaches. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3576050.3576122

LAK 2023, March 13–17, 2023, Arlington, TX, USA

ACM ISBN 978-1-4503-9865-7/23/03.
https://doi.org/10.1145/3576050.3576122

1 INTRODUCTION

With the convergence of diverse e-learning platforms and peripheral device usage, e-learning has become a mainstream education form over the last decade. The previous year's pandemic accelerated the need for e-learning due to the rapid transformation into online and hybrid settings. In e-learning, many learners have trouble managing their learning processes with less feedback on learning progress and support from educators. Research on Learning Analytics (LA) has developed a variety of methods and approaches to support self-regulation for learners in online and hybrid environments [7, 22, 26]. At the same time, educators have difficulty checking learners' engagement and progress and thus cannot provide timely learning support.

Reading documents on screen and tablet devices is essential to online and self-regulated learning. In the context of e-reading, attention management and keeping up attentive e-reading has been a difficult challenge for learners [47]. Especially, young readers in the previous years have suffered from attention span reduction by heavy usage of social media and primarily video-based content [47]. On the one hand, low attention of learners in e-reading leads to less effective and efficient learning [42]. On the other hand, it can also form a negative loop resulting in learners losing interest and engaging less in reading activities [52]. In this regard, our research investigates the impact of Human-Robot Interaction (HRI) design with affective and meta-cognitive support as an added intervention for e-reading.

In recent years, HRI has been implemented in diverse education practices and domains (e.g., physics, math, handwriting, reading, vocabulary, and chess [36]). Educational support has been implemented for various learning objectives (e.g., vocational training [53]) and different target groups (e.g., elementary school students [24]), taking different roles in the educational dialogue as educators, co-learners, and companions [48] in and outside the classroom [4].

In our research, we focus on HRI for reading support as we consider reading a core activity in most of today's higher education activities, and more and more reading is done on digital devices, from classical computer screens to tablets and mobile devices. We design our Furhat Robot[1] to function as a feedback agent in e-reading, which forms a social relationship with its empathic feedback and human-like features with appearance, speech, and

[1] https://furhatrobotics.com/

gestures. Educators' feedback with empathy and meta-cognitive prompts have been directly related to learners' cognitive, affective, and behavioral development in learning, leading to positive experiences and effective learning outcomes [34, 54]. Likewise, feedback with empathy and reflection is considered desirable for the educational HRI design to establish social relationships with learners and promote their critical thinking and meta-cognition [29, 38]. In this regard, we have the following research questions that we would like to focus on:

- How can HRI with empathic and meta-cognitive prompts support attention self-regulation in e-reading?
- How can self-regulatory learner behaviors in e-reading be recognized through a machine-learning approach?
- How can we develop an automatic system to predict learning outcomes, perceived learning experience, and perceived social presence of the social robot through the self-regulatory behaviors of learners?

2 BACKGROUND AND RELATED WORKS

2.1 Attention theories and indicators

Human attention has been defined and interpreted diversely at an intersection of education, psychology, neuroscience, and affective computing. [9] found that external attention toward different objects, modalities, and features is closely interlinked with internal attention. For instance, emotional arousal, triggered by external stimuli, can change the level of attention when acquiring information [35], form different internal associations [37], and affect the levels of working memory involved [46]. [55] also revealed that affective signals from sensory stimuli are one source that regulates various levels of awareness, perception, and attention. Such a link between sensory stimulation and attention emphasizes the importance of engaging in intervention for more productive, motivating, and better-perceived learning experiences [42]. [45] defined social attention as behaviors and motivations to engage in learning as a part of social communication, followed by visual attention towards learning materials.

However, in the context of e-reading and the implementation of HRI, the understanding of attention seems to be more specific since it is an educational environment where human agents (i.e., educators and peers) are absent. In this regard, our focus is to investigate the HRI effects on e-reading via diverse measurements. As discussed above and argued in the framework of Attention Network [40], human attention is characterized by not only cognition but also by temperamental differences such as expression and control of emotions and internal thoughts. In this regard, we examine multimodal cues that are direct and indirect clues of attention: attention self-regulation behaviors, knowledge gain, perceived interaction experience, and perceived social presence of the HRI.

2.2 Learning Analytics on HRI

We adopted the Analytics4Action Evaluation Framework (A4AEF) for our HRI analytics, an evidence-based LA intervention evaluation protocol that can be applied to online learning [41]. A4AEF has suggested teaching presence, cognitive presence, emotional presence, and social presence as core components of learning interventions to assist learners in planning, meaning construction, and facilitating engagement with the community of inquiry (e.g., learning technologies, contents, peers, and instructor). It is typically achieved by establishing a social learning space which is especially important in blended and online settings. A4AEF has further emphasized the usefulness of predictive models for instructors and learners based on learner data and analysis. We focus on four variables in our HRI analytics approach related to learners' attention: 1) *attention self-regulation* that are found as self-regulatory behaviors, 2) *knowledge gain* as a cognitive learning outcome, 3) *perceived interaction experience* from learners through their learning practices, and 4) *perceived social presence* of a social robot as a learning companion.

2.2.1 Attention self-regulation. With the convergence of sensor-driven approaches and machine learning techniques, diverse multimodal datasets have helped to gain insights into learners' cognitive and non-cognitive processes [6]. [22] indicated that there had been only a few studies about behavioral and measurable indicators of self-regulation in learning compared to its well-established theoretical and conceptual frameworks. Self-reporting is a traditional measure to collect learners' responses during or after learning activities, which is also often criticized due to the high dependency on learners' perception and awareness [58]. Biological signals from the body, brain, actions, and language have been implemented to measure brain activity, while learner behaviors have been coded and combined with diverse log data [22]. For instance, diverse parameters from the eyes, such as pupil diameter [49], fixations [21], and the number of blinks [23] have been investigated as cues of attention with the implementation of dedicated eye trackers and computer-vision approaches. Learners' emotions and arousal, which are known to be critical elements for attention changes, have been interpreted through facial expression changes, combined with various data points [18]. Gestural cues from the hands and body have been studied for individual, and group level attention [15]. In this work, we implement a framework of [27] for the data collection and behavior labeling, which combines the classification of self-regulatory learner behaviors and associated self-reported distractions in an e-reading setting. Specifically, the *behavioral cues of attention self-regulation* include movements from eyebrows, blinking, mumbling, hands, and body. We found such behavioral cues vital since it is the moment when learners are aware of changes in their own attention, which are also observable, that could directly lead to relevant intervention. Model building for attention regulation behavior recognition could also help to develop real-time loops for future LA and instructional design research.

2.2.2 Knowledge gain. Attention is known to enhance learning outcomes, and the acquisition of new knowledge, promoting the effectiveness of the learning [31]. Especially in e-reading, attention contributes to sifting vital information from others in the information selection process, which is a critical process of managing cognitive load during the learning process [47]. [47] suggested that the fundamental learning goals in e-reading as comprehension, reducing reading times, and increasing meta-cognition. Diverse learning strategies have been developed for e-reading as exploring, finding, analyzing, and evaluating documents to enhance the

knowledge gain [44]. For the knowledge gain assessment in reading, standard practices have been: asking questions about global or local information, text organization tasks, identifying main ideas, matching the sequence of events, and conclusions [47]. Meanwhile, three interrelated measures have been suggested to assess learners' knowledge gain, having their particular functions: diagnostic, formative, and summative assessments. While diagnostic assessment aims to determine the existing knowledge levels of learners, the formative assessment focuses on the current ongoing learning process and knowledge, both in formal and informal forms. The summative assessment evaluates the mastery level of the learning, which provides an overview of final learning achievements [13]. In this regard, we collect *diagnostic, formative, and summative assessment results* through pre-session, in-session, and post-session in diverse formats (e.g., multiple choice, true or false, multiple answers) to measure the knowledge gained in e-reading.

2.2.3 Perceived interaction experience. Attention span is known to be highly associated with the motivations, and emotional arousal of learners [2]. From the instructional design perspective, interaction is a critical component that affects motivation and emotional arousal in e-learning, where learners get better self-efficacy and adjust the cognitive load through sensory stimuli [51]. In this regard, the concept of User Experience (UX) and interaction experience [57] has often been adopted to understand learners' emotions, beliefs, preferences, perceptions, and accomplishments and applied to HRI and social robot evaluation, too [32]. The traditional circumplex model of affects has interpreted affects by dividing them into two dimensions: positive or negative valence and degree or extent of activation [39]. [12] has suggested an emotion measurement by categorizing users' perceptions based on appealingness, legitimacy, motive compliance, and novelty of emotions. The usability aspect of the interface has been scrutinized through the System Usability Scale (SUS) [3], while Attrakdiff measurement [20] has been developed for investigating diverse interface experiences and values that are delivered to users, having Pragmatic, Hedonic-Identity (Hedonic-I), Hedonic-Stimulation (Hedonic-S), and Attractiveness as its sub-dimensions. We implemented the *Attrakdiff measurement* [20] in our study since it has been a measurement developed especially for evaluating the interaction quality and focused on users' affective perceptions, which is our focus of interaction experience analysis.

2.2.4 Perceived social presence. In e-learning, social presence has been understood as a key component for deep and meaningful learning, contributing to learner participation and satisfaction towards learning [43]. Furthermore, it is known to encourage the cognitive actions of learners, and their critical thinking in learning processes [10]. Especially for e-reading with HRI, understanding the perceived social presence seems to be especially critical since the social robot forms an additional layer in the learning environment compared to the GUI-based interface. Traditionally, social robots have been evaluated for their interaction quality [45], perception of the robot appearance [1], rapport building, and relationship dynamics [25]. Immersion, parasocial interaction, parasocial relationships, physiological responses, social reality, and general social richness have been found as crucial factors of media as presence, which has been explicitly applied to compare the robot interaction with animated characters as social presence [33]. The framework of Social Presence [19] has emphasized attentional allocation, perceived message understanding, perceived affective understanding, perceived emotional interdependence, and perceived behavioral interdependence as criteria to evaluate the social presence, which has been adopted for HRI evaluation for the iCat, a companion robot for chess play [28]. We implement the modified *Social Presence measurement* since it is a measurement that has been well-established for diverse domains, including HRI evaluation, with diverse sub-dimensions and its validity.

2.3 Behavior-based attention prediction

To our best knowledge, very little behavior-based attention prediction research has been conducted in e-reading. [30] developed an attention prediction model in e-reading based on multimodal cues, such as eyebrow, lip, head movements, and mouse orientation. [56] used head orientation, eyelid, mouth height, gaze direction, and emotion to predict the six levels of attention labeled by annotators (i,e., sleepiness, drowsiness, fatigue, distraction, attention shift, concentration). [27] focused on self-regulatory learner behaviors (i.e., attention regulation behaviors) to regain attention during the e-reading and used it as a predictor of self-reported distractions from learners. In this work, we collect *attention regulation behavior* [27] to identify learning behavior differences in HRI. As we found that behavior patterns and analysis should vary based on a specific scenario [17], we collected a novel dataset containing the HRI analytics on attention self-regulation, knowledge gain, perceived interaction experience, and perceived social presence.

All in all, our contributions to the fields of Learning Analytics, Affective Computing, and Human-Robot Interaction are as stated as follows:

- We developed preliminary HRI interventions with empathic and meta-cognitive support for attentive e-reading. We analyzed learners' e-reading with HRI from diverse perspectives through direct and indirect attentional cues: attention self-regulation, knowledge gain, perceived interaction experience, and perceived social presence. It enables HRI analytics for both learners and instructors and further assists the design of e-reading support.
- We collected a novel dataset (SKEP) with five measurements and 25 features, spanning a total duration of nearly 40 hours with 4,210,860 frames, which includes data from sophisticated sensors, such as an eye tracker, and data layers with easy reproducibility, with a commercialized webcam and questionnaires. Rich data layers and intensive human annotations are provided as ground truths that enable more comprehensive analysis.
- A data-driven system has been proposed with state-of-the-art deep learning models for recognizing attention regulation behaviors (i.e., low-level recognition) and predicting knowledge gain, perceived interaction experience, and perceived social presence (i.e., high-level understanding). Our webcam-based approach is applicable to diverse reading-based e-learning scenarios, that can further be used to design and assess feedback.

3 A NOVEL DATASET FOR HRI-BASED E-READING ANALYTICS

3.1 Apparatus

We designed two interfaces: 1) a GUI-based system, with a monitor, mouse, and eye tracker implemented, and 2) an HRI-based system, which has a monitor, mouse, eye tracker, and Furhat Robot as physical components. See the footnote to check the specification of the Pupil Core eye tracker[2] and Logitech C505 HD Webcam[3], that were implemented. For both conditions, an informative e-reading material with technicality, "Waste management and critical raw materials," has been provided through a screen-based reader, which we explicitly developed for this study. The content has been chosen, aiming for an equal baseline knowledge for general readers. The text contains 4,750 words, divided into 29 pages, which cover seven subtopics. The text has been implemented with 47pt on a 27-inch monitor, having 2560×1440 resolution. The setting was optimized for the eye tracker implementation, which requires a bigger font size than the usual PDF readers for high-resolution data collection. See Figure 1 for a procedural summary of experimental settings.

3.2 Materials

We implemented four measurements that are direct and indirect attentional cues. Data features and granularity varies based on the data collection methods, collection timing, and post-processing of data.

3.2.1 Attention self-regulation. Learners' self-regulatory behavior has been collected through a video feed and annotated second-by-second by human labelers as post hoc. Labels are observable behavioral cues that indicate learners' attentional shifts. As [27] revealed that movements from the 1) *eyebrow*, 2) *blink*, 3) *mumble*, 4) *hands*, and 5) *body* works as good predictors of learners' self-awareness on attention loss, we annotated 60 video samples by applying six labels, including 6) *neutral* state as opposed to five attention regulation behaviour labels.

3.2.2 Knowledge gain. Knowledge levels have been measured pre-session, in-session, and post-session, to understand learners' baseline knowledge and knowledge gained through the reading session. Questionnaires with the same content have followed diverse formats (e.g., multiple choice, true or false, multiple answers) to prevent learners from getting familiarized with questions and making judgments based on their memory. We followed the formula below to reduce the complication in calculating *knowledge gain*

$$Score_{pre} = \sum_{i=1}^{Npre} S_i^{pre}, \tag{1}$$

$$Score_{post} = \sum_{i=1}^{Nin} S_i^{in} + \sum_{i=1}^{Npost} S_i^{post}, \tag{2}$$

$$KnowledgeGain = Score_{post} - Score_{pre}, \tag{3}$$

where S_i^{pre} is the pre-session score (0 or 1) for question i, while S_i^{in} is the in-session score (0 or 1) for question i and S_i^{post} is the post-session score (0 or 1) for question i. $Npre$, Nin, and $Npost$ that indicate the total number of questions in practices for pre-session, in-session, and post-session, which are 14, 7, and 7, respectively.

3.2.3 Perceived interaction experience. Attrakdiff measurement [20] provides assessments of learners' perceived interaction. The questionnaire has 28 questions with four sub-dimensions and seven scales between word pairs: 1) *Pragmatic* quality refers to users' perceived usability of the system (e.g., technical, complicated, practical, straightforward, predictable, clearly structured, manageable). 2) *Hedonic-I* focuses on characteristics that identify the system (e.g., connective, professional, stylish, premium, integrating, brings me closer, presentable). 3) *Hedonic-S* investigates perceived advancements of the system (e.g., inventive, creative, bold, innovative, captivating, challenging, novel). 4) *Attractiveness* measurement assesses the likeability of the system (e.g., pleasant, attractive, likable, inviting, good, appealing, motivating).

3.2.4 Perceived social presence. Social presence measurement [19] represents learners' evaluation of interfaces as perceived social beings. The questionnaire has 36 questions with six sub-dimensions: 1) *Co-presence* refers to users' perceived mutual awareness between the interface and the user. 2) *Attentional allocation* refers to a users' impression of exchanging attention with the interface. 3) *Perceived message understanding* is users' interpretation of mutual message understanding with the interface. 4) *Perceived affective understanding* is users' perception that both interface and users can interpret each others' affective states. 5) *Perceived emotional interdependence* conveys perceived mutual emotional impacts on each other. 6) *Perceived behavioral interdependence* shows the perceived behavioral changes triggered between the user and the interface.

3.3 Procedure

We recruited bachelor's and master's students on campus who use the English language for their daily education. We kept nearly equal gender ratios and non-significant age differences to prevent cognitive capability differences and following distinctions among participants. GUI condition had 18 males and 12 females with an age range of 19 to 33 (M=25.8, SD=3.35). HRI condition had 19 males and 11 females with an age range of 19 to 37 (M=24.1, SD=4.30). Participants have been invited to an experiment individually for an e-reading task. While a researcher in the GUI condition solely gave instructions about the interface and the procedure, a Furhat Robot helped the researcher's instruction in the HRI setting so that participants could internalize how to make the speech input to the robot. A screen-based pre-test questionnaire with 14 questions was given to measure the baseline knowledge about the topic. There were 10 minutes of time limitations for the pre-test. Once the pre-test was finished, a researcher entered the room, let learners wear an eye tracker, and further calibrated it. A webcam was activated when learners clicked the "start reading" button. Participants proceeded with the reading session by reviewing the text on the screen reader. Throughout the process, seven pop-up questions were given to both conditions at the end of each subtopic, while emphatic & meta-cognitive robot feedback (Figure 2) was given only in the HRI condition, two seconds after the last page of each subtopic was triggered. Once the reading session had finished,

[2]https://pupil-labs.com/
[3]https://www.logitech.com/

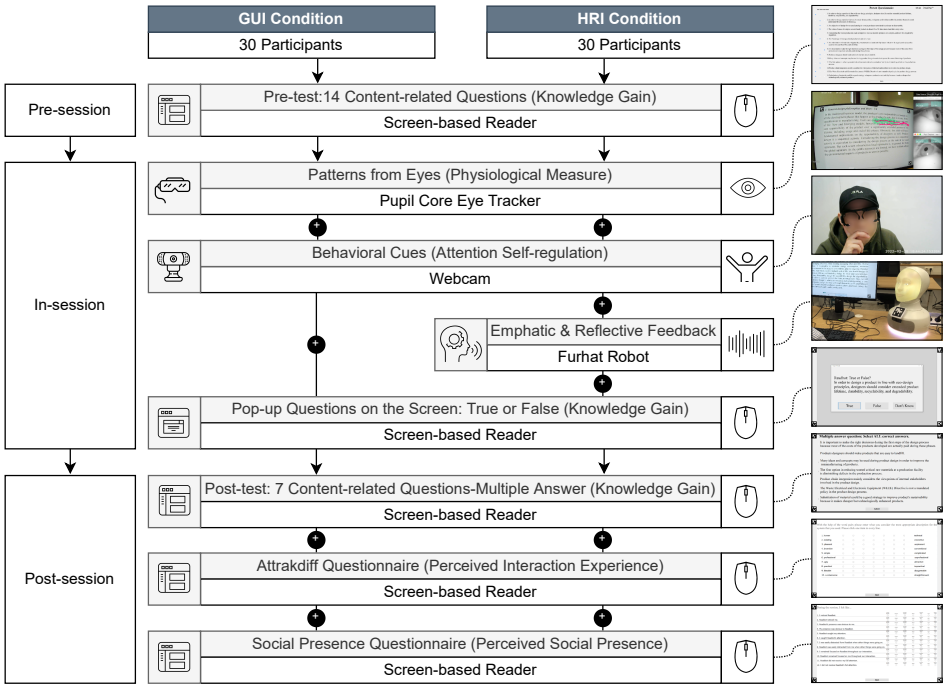

Figure 1: A procedural summary of the GUI and HRI settings.

participants were given a post-test questionnaire with seven statements as multiple-answer questions in both conditions. Likewise, all participants received an Attrakdiff questionnaire with 28 questions and a Social Presence questionnaire with 36 questions as the final post-reading session.

3.4 Dataset construction

As illustrated in Table 1, our SKEP dataset contains multimodal data with diverse objectives, input channels, features, granularity, and data formats in different collection timing, which gives insights into direct and indirect cues of attention. Note that the data from the eye tracker has not been used in this study.

3.5 Data processing and annotation

Sixty video samples from the GUI and HRI conditions with nearly 40 hours (2,339 minutes) have been collected. The raw data has been segmented into every 30 frames (1 second) for the second-to-second labeling from annotators. In total, the video data that has been annotated are 4,210,860 frames. Two labelers (one doctoral student and one master's student) have been instructed about the labeling criteria for the annotation. Six labels have been used, including neutral state, as opposed to five attention regulation behaviors: movements in eyebrow, blink, mumble, hand, and body. In the second round, the labels were summarized and cross-checked to address the inconsistent cases for validation. Note that the behavior labels should be able to provide nearly homogeneous judgments regardless of observers' expertise in attention analysis since labeling only requires factual judgments based on the criteria. See Figure 3 for an overview of the data processing and annotation criteria.

4 STATISTICAL ANALYSIS ON ATTENTIONAL CUES IN E-READING: GUI VS. HRI

In the following, we present descriptive and statistical analysis to show the overall effects of the treatment (GUI, control group and HRI, treatment group) on learners' 1) attention regulation behaviors, 2) knowledge gain, 3) perceived interaction experience, and 4) perceived social presence. Note that the average of all subdimensions has been derived to get the overall Attrakdiff and Social Presence evaluation. Furthermore, a one-way ANOVA (Welch's) analysis has been conducted to find the statistically significant differences between GUI and HRI conditions.

4.1 Attention self-regulation behaviors

We labeled five attention regulation behaviors, which are sound indicators of learners' perceived distractions [27], every second. The neutral behavior indicates the status without any attention regulation behaviors. The dataset showed that the movements on the body (1,048,170) as the most frequent form of attention regulation behavior, while the blink (196,590 frames) and the eyebrow (59,010 frames) have minor cases among labeled attention regulation behaviors. Mumble has recorded 563,640 frames, while hand movements have shown 268,230 frames. As shown in Table 2, more neutral behavior has been observed in the HRI (M=1198.0, SD=273.99) than in the GUI (M=1198.0, SD=273.99), while more eyebrow, mumble, and body movements have taken more places in the HRI with statistical significance. More mumbling and body movements have occurred in HRI since speech-based interaction, and robot-looking has been a part of HRI design. According to our observation, different individuals' unique behavioral patterns, such as expressiveness in behaviors,

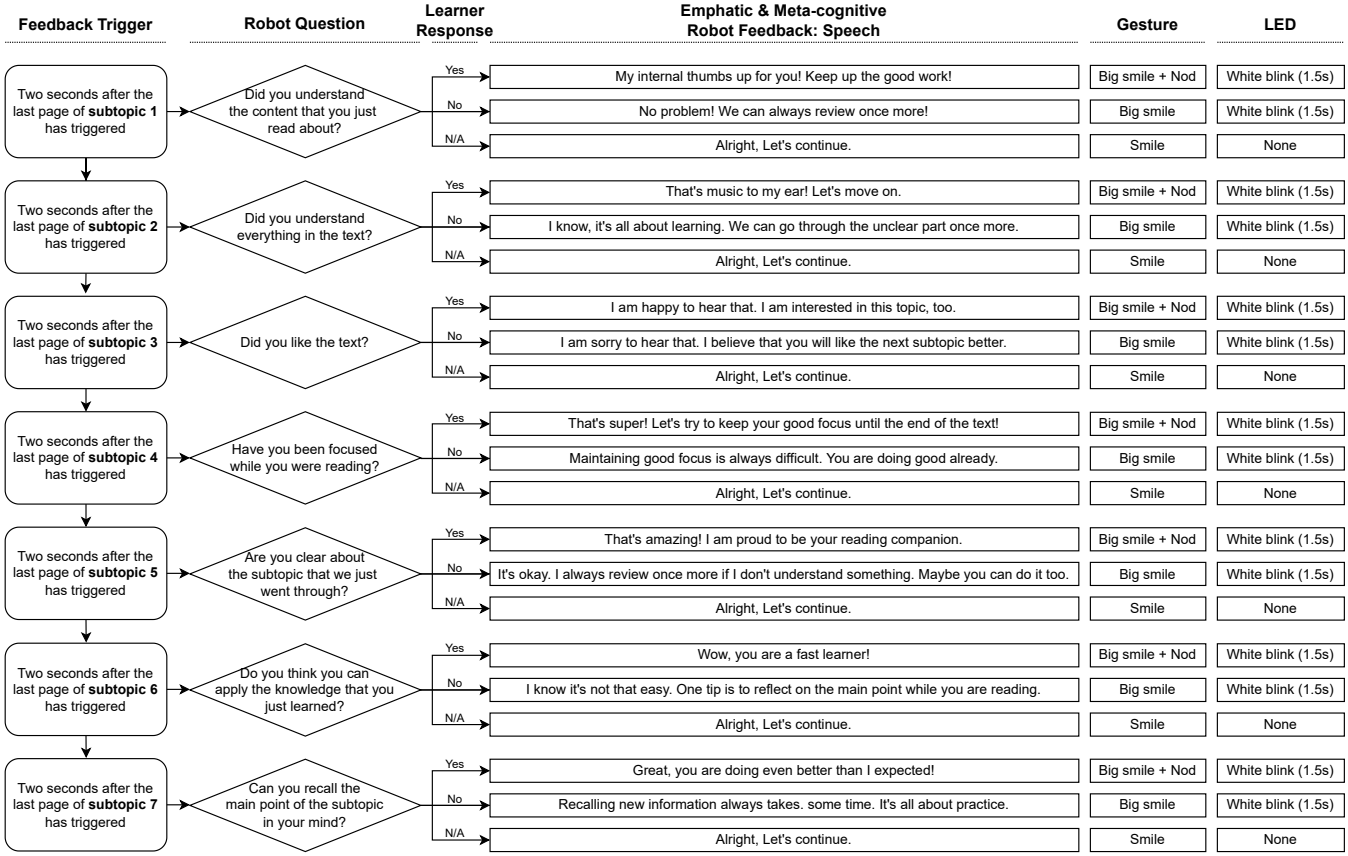

Figure 2: Emphatic & meta-cognitive HRI feedback protocol.

Table 1: A summary of our novel attention self-regulation, knowledge gain, perceived interaction experience, and perceived social presence with HRI in the e-reading (SKEP) dataset.

Objectives	Measurements	Collection Timing	Input Channels	Modalities	Features	Granularity	Data Formats
Attention Self-regulation	Attention Regulation Behaviors	-Throughout the session	-Webcam	-Behaviors -Annotations	-Eyebrow -Blink -Mumble -Hands -Body	-Video -Human Annotation on Every Second (30 fps on 4,210,860 Frames)	-AVI -CSV
Patterns from eyes	Eye Tracking	-Throughout the session	-Eye Tracker	-Eye movements	-Pupil Diameter -Gaze Positions -Gaze on Surface/Markers -Blinks -Fixation -Video (Head Mounted) -Video (Infrared for Eyes)	-Infrared Cameras: 120Hz -Frontal Camera: 30Hz	-AVI -JSON -CSV
Knowledge Gain	Diagnostic, formative, and summative assessments	-Pre-session -In-session -Post-session	-Mouse Click	-Text	-Pre-test -In-session -Post-test	-14 Instances on Each Subtopic	-CSV
Perceived Interaction Experience	Attrakdiff Measurement	-Post-session	-Mouse Click	-Text	-Pragmatic Quality -Hedonic-I Quality -Hedonic-S Quality -Attractiveness	-28 Questions on Overall Interface (7-Scale Likert)	-CSV
Perceived Social Presence	Social Presence Measurement	-Post-session	-Mouse Click	-Text	-Co-presence -Attentional Allocation -Perceived Message Understanding -Perceived Affective Understanding -Perceived Emotional Interdependence -Perceived Behavioral Interdependence	-36 Questions on Overall Interface (7-Scale Likert)	-CSV

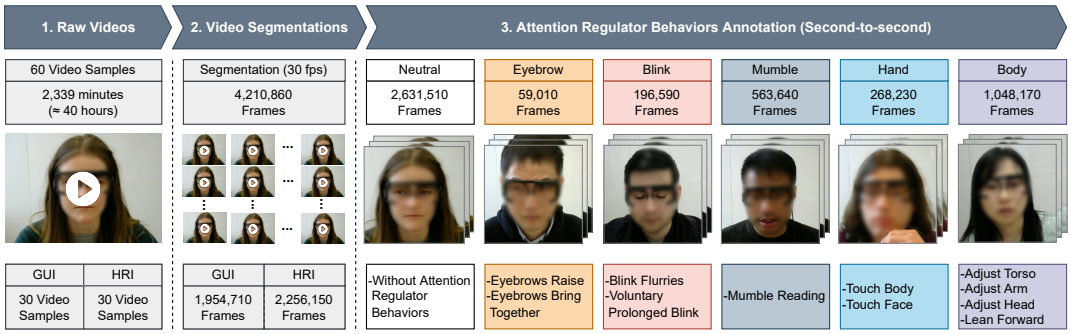

Figure 3: Data processing and annotation criteria.[1]

[1] Images were blurred for identity protection purposes. All images were consented to be used for publication.

Table 2: Attention regulation behaviors from GUI & HRI.

Measurement	GUI	HRI	One-way ANOVA			
	M(SD)		F	df1	df2	p
Neutral	1081.13(317.82)	**1198.0(273.99)**	**118.73**	1	83991	<.001
Eyebrow	3.50(3.57)	**16.10(16.13)**	**11.78**	1	87792	<.001
Blink	**31.03(13.36)**	28.83(11.18)	**13.62**	1	86616	<.001
Mumble	3.97(3.74)	**40.43(34.14)**	**98.96**	1	87040	<.001
Hand	**65.93(93.50)**	21.60(16.46)	1.41	1	84239	0.234
Body	189.43(100.48)	**264.83(115.50)**	**425.43**	1	81155	<.001

Table 3: Knowledge gain from GUI & HRI.

Measurement	GUI	HUM	One-way ANOVA			
	M(SD)		F	df1	df2	p
Pre-test Score	**3.47(2.52)**	2.47(2.18)	2.711	1	56.8	0.105
Post-test Score	9(1.66)	**9.3(1.86)**	0.434	1	57.3	0.513
Knowledge Gain	5.53(2.86)	**6.83(3.04)**	2.908	1	57.8	0.094
Perceived Knowledge Gain	4.1(1.47)	**5(1.14)**	**4.337**	1	55.2	0.042

Table 4: Perceived interaction experience from GUI & HRI.

Measurement	GUI	HUM	One-way ANOVA			
	M(SD)		F	df1	df2	p
Overall Attrikdiff	**0.583(0.633)**	0.537(0.511)	0.09777	1	55.5	0.756
Pragmatic Quality	**1.1(0.721)**	0.676(0.824)	**4.49836**	1	57.0	0.038
Hedonic Quality-I	**0.324(0.833)**	0.314(0.597)	0.00259	1	52.6	0.960
Hedonic Quality-S	0.348(1.12)	**0.652(0.718)**	1.56827	1	49.3	0.216
Attractiveness	**0.562(0.852)**	0.505(0.958)	0.05956	1	57.2	0.808

Table 5: Perceived social presence from GUI & HRI.

Measurement	GUI	HUM	One-way ANOVA			
	M(SD)		F	df1	df2	p
Overall Social Presence	3.59(0.671)	**4.14(0.484)**	**13.07**	1	52.7	<.001
Co-presence	4.32(1.4)	**5.45(0.796)**	**14.81**	1	45.9	<.001
Attentional Allocation	3.59(0.823)	**4.03(0.683)**	**5.18**	1	56.1	0.027
Perceived Message Understanding	4.14(0.51)	**4.48(0.437)**	**7.39**	1	56.7	0.009
Perceived Affective Understanding	3.47(0.907)	**3.73(0.606)**	1.65	1	50.6	0.205
Perceived Emotional Interdependence	2.64(1.05)	**3.33(1.02)**	**6.63**	1	57.9	0.013
Perceived Behavioral Interdependence	3.39(1.4)	**3.81(1.17)**	1.60	1	56.2	0.211

frequent usage of particular behaviors, and significant behaviors as attentional cues, have been derived mainly from individual differences than experimental conditions. In this regard, further model training does not differentiate attention regulation behavior labels by experimental conditions. We combine both conditions as a whole to achieve attention regulation behavior recognition and further predict other attentional cues.

4.2 Knowledge gain

Table 3 summarizes the overall knowledge gained in both conditions, with the pre-test score, post-test score, and perceived knowledge gain. The GUI (M=3.47, SD=2.52) recorded a higher pre-test score than the HRI (M=2.47, SD=2.18). However, a higher post-test score has been documented in the HRI (M=9.3, SD=1.86) than in the GUI (M=9, SD=1.66), representing higher knowledge gain in the HRI. However, the difference between groups did not show statistical significance. The perceived knowledge gain after the reading practice was higher in the HRI (M=5, SD=1.14) setting compared to the GUI (M=4.1, SD=1.47) on a significant level (p=0.042). It indicates that empathic and meta-cognitive HRI feedback has helped learners' self-efficacy. We conducted a further Pearson's correlation analysis between the perceived knowledge gain and the actual knowledge gain to find if learners' perception of their learning achievement correlates to the objective learning outcomes. However, the perceived knowledge gain did not show a correlation with actual knowledge gain (r=.071, p=.589) both in the GUI (r=.052, p=.786) and the HRI (r=-.030, p=.876) settings.

4.3 Perceived interaction experience

Overall Attrakdiff. As shown in Table 4, the overall Attrakdiff measurement on the GUI (M=0.583, SD=0.633) has gained higher scores than the HRI (M=0.537, SD=0.511). However, our ANOVA analysis has shown significance only in Pragmatic Quality measurement between two conditions.

Pragmatic Quality. Table 4 shows that the GUI (M=1.1, SD=0.721) has been evaluated to be more pragmatic than the HRI (M=0.676, SD=0.824). Participants highly appreciated the simplicity, practicality, straightforwardness, predictability, and clear structure of the GUI compared to the HRI. The assessment of the HRI has shown a wide distribution, especially in the "technical-human" measure, representing users' contradicting perceptions. It indicates that the presence of the reflective & empathic robot has often been perceived differently than the original system design intention: we premised the HRI will be consistently perceived as more "human" than the GUI system, but the evaluation results have varied. We assume participants' preconceptions of robots and human-robot interactions impacted their current evaluation, which should be further investigated.

Hedonic-S. The overall hedonic-S measure was highly evaluated in the HRI (M=0.652, SD=0.718) compared to the GUI (M=0.348, SD=1.12). The HRI has been perceived as inventive, creative, innovative, captivating, challenging, and novel than the GUI system. A wide distribution of participant responses was found in the overall GUI for hedonic-S evaluation. It seems to be because some users have perceived our GUI system as a traditional e-reading system, while some perceived the pop-up questions as creative and novel stimuli, which could be developed as a potential intervention with improvements.

Hedonic-I and Attractiveness. In hedonic-I (GUI: M=0.324, SD=0.833; HRI: M=0.314, SD=0.597) and attractiveness (GUI: M=0.562, SD=0.852; HRI: M=0.505, SD=0.958) measurements, the GUI has received slightly higher scores than the HRI without significance. However, the HRI has been evaluated as more premium in the hedonic-I measure while being evaluated as more likable, inviting, and motivating in the Attractiveness measure.

4.4 Perceived social presence

Perceived social presence. The overall Social Presence measurement has gained higher scores in the HRI (M=4.14, SD=0.484) compared to the GUI (M=3.59, SD=0.671) on all sub-dimensions (Table 5). An ANOVA analysis has shown significance in the overall Social Presence, Co-presence, Attentional Allocation, Perceived Message Understanding, and Perceived Emotional Interdependence.

Co-presence. Most participants perceived the HRI as a "presence", while evaluation of the GUI has varied. Co-presence has shown the highest evaluation result among all sub-dimensions in the HRI (M=5.45, SD=0.796) while showing the widest distribution in the GUI (M=4.32, SD=1.4). The same tendency has been observed from the perceived behavioral independence measurement, showing that HRI is more often perceived as a "presence" than the GUI.

Attentional Allocation, Perceived Message Understanding, Perceived Affective Understanding, Perceived Emotional Interdependence, and Perceived Behavioral Interdependence. Unlike the GUI, users expected a certain attentional, intentional, emotional connectivity with the HRI, showing different role expectations towards different interfaces. Such perception toward HRI has likely to affect learners' emotional (M=3.33, SD=1.02) and behavioral (M=3.81, SD=1.17) susceptibility to the HRI, leading to higher interdependence on emotional and behavioral levels. On the other hand, the broad spectrum in the Attentional Allocation (M=3.59, SD=0.823) and Perceived Behavioral Interdependence (M=3.39, SD=1.4) measurements in the GUI indicates that it was unclear for some users whether the GUI reacts based on their behaviors (i.e., intelligent system) or if the feedback was independent to participants. It seems to be because participants premised the HRI as an intelligent system, though robot behavior has been pre-designed regardless of learners' behaviors or speech: it indicates the necessity of developing an intelligent system based on real-time learning analytics.

5 A DATA-DRIVEN SYSTEM DEVELOPMENT WITH DEEP LEARNING APPROACHES FOR ATTENTIVE E-READING ANALYSIS

This section introduces a data-driven system with deep learning approaches for developing an attentive e-reading analysis. Specifically, we exploit a two-stage framework to build the system by leveraging the rich data streams collected from the SKEP dataset. In the first stage of low-level processing, we implemented vision-based behavior recognition of the subjects with computer vision technologies. In the latter stage of high-level analysis, we utilized recognized subjects' behaviors as feature vectors to achieve the attentive e-reading analysis with machine learning models in a holistic way.

5.1 Recognizing attention regulation behaviors with computer vision techniques

In recent years, the deep learning and computer vision fields have made remarkable achievements in various vision tasks [5, 8, 14]. Inspired by powerful AI models, we tried to leverage them to enhance HRI-based attentive e-reading. More precisely, we implemented three of the most standard temporal neural networks: CNN-RNN, CNN-LSTM, and CNN-Transformer to achieve the low-level behavior recognition of subjects during their e-reading. To have standard evaluations for all the reported results on the SKEP dataset, we utilized the cross-subject evaluation protocol, which divides the 60 subjects into a training group of 40 subjects (94,519 samples), and a testing group of 20 subjects (45,843 samples). We used the six classes of annotated attention regulation behaviors as the ground truth to train and evaluate the models' performances. Table 6, presents the performances of baseline networks.

With the result listed in Table 6, our observations are listed as follows: 1) the best methods' accuracy went up to 72.79 %, which is much higher compared to a random guess over six classes (16.67%). It verifies the powerful video recognition ability of deep learning models. 2) RNN-based model has shown the highest performance of 72.97% since larger-scale models like LSTM and Transformer models easily overfit on our SKEP dataset. 3) Capturing shorter temporal dynamics (temporal reasoning) is vital for better performance which proves again that fewer parameters can avoid the overfitting issue (the best two performances are obtained by setting the temporal step as 5). Note that the vast performance drop in 112-size images with an accuracy of 47.72% (compared to 224 size with 72.97%) has been mainly caused by information loss due to the smaller image size. For instance, movements from mumbling, eyebrows, and blinking are extremely subtle. It only takes 2-10 pixels to present those regions at an image size of 112, which provides insufficient image information. However, when it comes to size 224, feature learning can be significantly improved.

5.2 Automatic e-reading-based attention analysis using attention regulation behaviors

In this section, we applied classical machine learning models to predict knowledge gain, perceived interaction experience, and perceived social presence, using attention regulation behaviors obtained from the previous stage as the feature vectors. Similar to the

Table 6: attention regulation behavior recognition using deep neural networks on SKEP dataset. The highest result is marked in bold. The second highest result is marked underline.

Model Type	Temporal Step	Video Input Size	Accuracy
CNN-RNN	5	112	47.72%
		224	**72.97%**
	10	112	63.92%
		224	62.03%
CNN-LSTM	5	112	58.17%
		224	49.86%
	10	112	34.19%
		224	65.61%
CNN-Transformer	5	112	36.91%
		224	<u>72.84%</u>
	10	112	55.88%
		224	65.89%

previous stage, we utilized the cross-subject evaluation protocol. Note that the measurement of attentive analysis (e.g. knowledge gain) is obtained based on the whole e-reading progress. Thus, one subject can have 60 samples in total (40 for training and 20 for testing). We deployed five of the most classical machine learning models to learn the various attention patterns as shown in Table 7, 8, and 9. See the footnote[4] for the implementation details.

5.2.1 Knowledge gain prediction. Knowledge gain prediction is of the highest importance among all measurements since knowledge gain is the most fundamental objective of e-reading activities. We encoded the distribution of attention regulation behaviors that happened within a given attention span as feature vectors with dimensions of $1 \times N$. N is the number of attention regulation behaviors and neutral behavior, as six in practice. Then, we fed the feature vectors into classifiers to predict learners' knowledge gain with the probability distribution. We present two evaluation settings: 1) fine-grained knowledge gain prediction (5-level): excellent-good-average-poor-very poor; and 2) coarse knowledge gain prediction (3-level): good-average-poor. Even through human observation, differentiating fine-grained knowledge gains is difficult or nearly impossible. As shown in Table 7, all the classifiers can achieve encouraging results (above 63.57% accuracy) in the coarse(3-level) knowledge gain prediction, while relatively lower accuracy (around 40%) has been achieved on challenging fine-grained knowledge gain prediction. The SVM classifier has achieved the highest accuracy for both fine-grained and coarse 45.0% and 74.29%, respectively.

5.2.2 Perceived interaction experience prediction. Similar to knowledge gain, we trained the classifiers to predict the perceived interaction experience of subjects. Instead of making it a regression task, we converted the task into a classification task by assigning learners' scores into positive, neutral, and negative, based on the medium value of "4" from the Attrakdiff 7-Likert scale. The prediction with the raw score shows whether learners will have positive, neutral, or negative interaction experiences. However, using the raw score has a limitation in that it leads to nearly-binary prediction (positive or negative) as it is improbable that the evaluation result of a specific sub-dimension takes the exact neutral value. Thus, we further defined the three classes into a normalized distribution [16] with the percentile of participants' scores (below 25%, 25-75%, and above 75%). As described in Table 8, Random Forest provides the best performance for all sub-dimensions of Attrakdiff measurement, scoring the highest performance in the raw score for the Pragmatic Quality prediction with 92.5% of accuracy. The best accuracy lies in the Hedonic-I prediction with 87.5%.

Table 7: Knowledge Gain (KG) prediction using attention regulation behaviors as a predictor.

Method	Accuracy (%)	
	Fine-grained KG (5-level)	Coarse KG (3-level)
Random Guess	20.00	33.33
Random Forest	38.57	69.29
AdaBoost	37.14	63.57
MLP	40.00	70.00
kNN	40.71	70.00
SVM	**45.00**	**74.29**

5.2.3 Perceived social presence prediction. Perceived social presence prediction has followed the protocol of perceived interaction experience prediction: 1) splitting raw distribution to positive, neutral, and negative levels and 2) dividing normalized distribution into the first (25%), second (25-75%), and third quartiles (75%). Table 9 shows that the Random Forest classifier best predicted the overall Social Presence (SP), Co-presence (CP), Attentional Allocation (AA), and Perceived message understanding (PMU) for both raw and normalized distributions. The MLP also has shown high performance in predicting the Perceived Behavioral Interdependence measurement (PBI). From the raw distribution, the highest result has been achieved with 92.5% accuracy in both Co-presence (CP) and Perceived Emotional Interdependence measurement (PEI) predictions. For the classes obtained from normalized distribution, the prediction results went up to 100%, 97.5%, and 95% for predicting Co-presence (CP), Perceived Message Understanding (PMU), and Perceived Emotional Interdependence (PEI), respectively, representing the attention regulation behaviors as effective predictors.

6 CONCLUSION

We comprehensively investigated the effect of social robots in e-reading by collecting the novel SKEP dataset. In the SKEP dataset, we set HRI-based (treatment) and GUI-based (control) conditions and captured rich multimodal features. The SKEP dataset includes more than four-million frames of various sensor data and intensive human annotated ground truths, which function as learners' direct and indirect attentional cues shown during the e-reading. We found that there have been specific role expectations toward different interface types, which leads to more attentional, emotional, and social connectivity with the HRI. We developed a data-driven system using the SKEP dataset with cutting-edge deep-learning approaches. The proposed system showed a promising performance with high

[4]In the above models, we set the following architecture hyper-parameters: CNN: ImageNet-pre-trained [11] InceptionV3 [50] with N = 2048 feature dimensions and average pooling for the last layer. RNN: LSTM: 1-Layer LSTM with N = 256 units. Transformer: Positional Embedding, TransformerEncoder with N=2048 units, GlobalMaxPooling1D, and a fully connected layer to Softmax output. The learning rate is all set as 0.0002 with a decay factor of 0.999 for every five training epochs with a Titan RTX GPU. All other configurations follow the original network architectures unless stated otherwise, such as temporal step and video input size in Table 6. We used Tensorflow/2.8 platform for deploying the deep learning models and scikit-learn Python for machine learning models.

Table 8: Perceived interaction experience prediction using attention regulation behaviors as a predictor.

Method	Accuracy (%)									
	Overall Attrakdiff		Pragmatic Quality		Hedonic Quality-I		Hedonic Quality-S		Attractiveness	
	Raw	Normalized	Raw	Normalized	Raw	Normalized	Raw	Normalized	Raw	Normalized
Random Guess	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33
SVM	62.50	62.50	87.50	52.50	50.00	60.00	52.50	52.50	62.50	62.50
Random Forest	**72.50**	72.50	**92.50**	**72.50**	**82.50**	82.50	**77.50**	72.50	**70.0**	**72.5**
AdaBoost	52.50	67.50	90.00	57.50	57.50	**87.50**	70.00	**70.00**	60.00	67.50
MLP	62.50	**75.00**	87.50	45.00	65.00	47.50	42.50	40.00	**70.00**	57.50
kNN	60.00	62.50	87.50	47.50	57.50	62.50	42.50	42.50	62.50	62.50

Table 9: Perceived social presence measurement prediction using attention regulation behaviors as a predictor. [1] SP: Social Presence, CP: Co-presence, AA: Attentional Allocation, PMU: Perceived Message Understanding, PAU: Perceived Affective Understanding, PEI: Perceived Emotional Interdependence, PBI: Perceived Behavioral Interdependence.

Method	Accuracy (%)													
	Overall SP		CP		AA		PMU		PAU		PEI		PBI	
	Raw	Normalized	Raw	Normalized	Raw	Normalized	Raw	Normalized	Raw	Normalized	Raw	Normalized	Raw	Normalized
Random Guess	50.00	33.33	50.00	33.33	50.00	33.33	50.00	33.33	50.00	33.33	50.00	33.33	50.00	33.33
SVM	62.50	52.50	87.50	97.50	50.00	60.00	52.50	87.50	62.50	47.50	90.00	**95.00**	75.00	75.00
Random Forest	**72.50**	**75.00**	**92.50**	**100.0**	**80.00**	**85.00**	**70.00**	**97.50**	70.0	**60.00**	**92.50**	92.5	80.0	**82.50**
AdaBoost	52.50	65.00	90.00	97.50	57.50	67.50	67.50	87.50	60.00	57.50	85.00	90.00	70.00	70.00
MLP	67.50	67.50	90.00	95.00	70.00	72.50	**70.00**	90.00	**75.00**	37.50	77.50	92.50	**85.00**	**82.50**
kNN	60.00	55.00	87.50	97.50	57.50	57.50	42.50	87.50	62.50	37.50	90.00	95.00	67.50	67.50

attention regulation behavior recognition and high prediction results for knowledge gain, perceived interaction experience, and perceived social presence. It proves the attention regulation behavior as sound observable cues of direct and indirect attention cues in e-reading.

REFERENCES

[1] Daniel Belanche, Luis V Casaló, Jeroen Schepers, and Carlos Flavián. 2021. Examining the effects of robots' physical appearance, warmth, and competence in frontline services: The Humanness-Value-Loyalty model. *Psychology & Marketing* 38, 12 (2021), 2357–2376.

[2] Neil A Bradbury. 2016. Attention span during lectures: 8 seconds, 10 minutes, or more? , 509–513 pages.

[3] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[4] C Bühler and H Knops. 1999. Robots in the classroom-tools for accessible education. *Assistive technology on the threshold of the new millennium* 6 (1999), 448.

[5] Haoyu Chen, Xin Liu, Jingang Shi, and Guoying Zhao. 2020. Temporal Hierarchical Dictionary Guided Decoding for Online Gesture Segmentation and Recognition. *IEEE Transactions on Image Processing* 29 (2020), 9689–9702.

[6] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. 2023. SMG: A Micro-Gesture Dataset Towards Spontaneous Body Gestures for Emotional Stress State Analysis. In *International Journal of Computer Vision*.

[7] Haoyu Chen, E Tan, Y Lee, S Praharaj, M Specht, and G Zhao. 2020. Developing AI into explanatory supporting models: An explanation-visualized deep learning prototype. In *The International Conference of Learning Science (ICLS)*.

[8] Haoyu Chen, Hao Tang, Henglin Shi, Wei Peng, Nicu Sebe, and Guoying Zhao. 2021. Intrinsic-Extrinsic Preserved GANs for Unsupervised 3D Pose Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8630–8639.

[9] Marvin M Chun, Julie D Golomb, Nicholas B Turk-Browne, et al. 2011. A taxonomy of external and internal attention. *Annual review of psychology* 62, 1 (2011), 73–101.

[10] Susan Copley Cobb. 2009. Social presence and online learning: A current view from a research perspective. *Journal of Interactive Online Learning* 8, 3 (2009).

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[12] Pieter Desmet. 2002. Designing emotions.

[13] Nargiza Gaybullaevna Dilova. 2021. FORMATIVE ASSESSMENT OF STUDENTS'KNOWLEDGE–AS A MEANS OF IMPROVING THE QUALITY OF EDUCATION. *Scientific reports of Bukhara State University* 5, 3 (2021), 144–155.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[15] Virginia J Flood, Francois G Amar, Ricardo Nemirovsky, Benedikt W Harrer, Mitchell RM Bruce, and Michael C Wittmann. 2015. Paying attention to gesture when students talk chemistry: Interactional resources for responsive teaching. *Journal of Chemical Education* 92, 1 (2015), 11–22.

[16] Richard D Goffin, R Blake Jelley, Deborah M Powell, and Norman G Johnston. 2009. Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management* 48, 2 (2009), 251–268.

[17] Katharine H Greenaway, Elise K Kalokerinos, and Lisa A Williams. 2018. Context is everything (in emotion research). *Social and Personality Psychology Compass* 12, 6 (2018), e12393.

[18] Jason M Harley, François Bouchet, M Sazzad Hussain, Roger Azevedo, and Rafael Calvo. 2015. A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior* 48 (2015), 615–625.

[19] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. In *Seventh annual international workshop: Presence*, Vol. 2004. Universidad Politecnica de Valencia Valencia, Spain.

[20] Marc Hassenzahl, Annika Wiklund-Engblom, Anette Bengs, Susanne Hägglund, and Sarah Diefenbach. 2015. Experience-oriented and product-oriented evaluation: psychological need fulfillment, positive affect, and product perception. *International journal of human-computer interaction* 31, 8 (2015), 530–544.

[21] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R Brockmole, and Sidney K D'Mello. 2019. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction* 29, 4 (2019), 821–867.

[22] Sanna Järvelä, Jonna Malmberg, Eetu Haataja, Marta Sobocinski, and Paul A Kirschner. 2019. What multimodal data can tell us about the students' regulation of their learning process. *Learning and Instruction* 72, 7 (2019), 4.

[23] Bryant J Jongkees and Lorenza S Colzato. 2016. Spontaneous eye blink rate as predictor of dopamine-related cognitive function A review. *Neuroscience & Biobehavioral Reviews* 71 (2016), 58–82.

[24] Takayuki Kanda, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. 2007. A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on robotics* 23, 5 (2007), 962–971.

[25] Jacqueline M Kory-Westlund and Cynthia Breazeal. 2019. Exploring the effects of a social robot's speech entrainment and backstory on young children's emotion, rapport, relationship, and learning. *Frontiers in Robotics and AI* 6 (2019), 54.

[26] Yoon Lee. 2020. FLOWer: Feedback Loop for Group Work Supporter. *The International Learning Analytics and Knowledge Conference (LAK demo session)* (2020).

[27] Yoon Lee, Haoyu Chen, Guoying Zhao, and Marcus Specht. 2022. WEDAR: Webcam-based Attention Analysis via Attention Regulator Behavior Recognition with a Novel E-reading Dataset. In *24th ACM International Conference on Multimodal Interaction (ICMI)*. ACM, 319–328.

[28] Iolanda Leite, Carlos Martinho, Andre Pereira, and Ana Paiva. 2009. As time goes by: Long-term evaluation of social presence in robotic companions. In *RO-MAN 2009-the 18th IEEE international symposium on robot and human interactive communication*. IEEE, 669–674.

[29] Iolanda Leite, André Pereira, Samuel Mascarenhas, Carlos Martinho, Rui Prada, and Ana Paiva. 2013. The influence of empathy in human–robot relations. *International journal of human-computer studies* 71, 3 (2013), 250–260.

[30] Jiajia Li, Grace Ngai, Hong Va Leong, and Stephen CF Chan. 2016. Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics. *ACM SIGAPP Applied Computing Review* 16, 3 (2016), 37–49.

[31] Ming-Hung Lin, Huang-g Chen, et al. 2017. A study of the effects of digital learning on learning motivation and learning outcome. *Eurasia Journal of Mathematics, Science and Technology Education* 13, 7 (2017), 3553–3564.

[32] Jessica Lindblom and Rebecca Andreasson. 2016. Current challenges for UX evaluation of human-robot interaction. In *Advances in ergonomics of manufacturing: Managing the enterprise of the future*. Springer, 267–277.

[33] Matthew Lombard, Theresa B Ditton, Daliza Crane, Bill Davis, Gisela Gil-Egui, Karl Horvath, Jessica Rossman, and S Park. 2000. Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument. In *Third international workshop on presence, delft, the netherlands*, Vol. 240. 2–4.

[34] Gretchen McAllister and Jacqueline Jordan Irvine. 2002. The role of empathy in teaching culturally diverse students: A qualitative study of teachers' beliefs. *Journal of teacher education* 53, 5 (2002), 433–443.

[35] Steven B Most, Marvin M Chun, David M Widders, and David H Zald. 2005. Attentional rubbernecking: Cognitive control and personality in emotion-induced blindness. *Psychonomic bulletin & review* 12, 4 (2005), 654–661.

[36] Jauwairia Nasir, Utku Norman, Wafa Johal, Jennifer K Olsen, Sina Shahmoradi, and Pierre Dillenbourg. 2019. Robot analytics: What do human-robot interaction traces tell us about learning?. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–7.

[37] Elizabeth A Phelps, Sam Ling, and Marisa Carrasco. 2006. Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological science* 17, 4 (2006), 292–299.

[38] Dimitrios Pnevmatikos, Panagiota Christodoulou, and Nikolaos Fachantidis. 2018. Promoting Critical Thinking Dispositions in Children and Adolescents Through Human-Robot Interaction with Socially Assistive Robots. In *International Conference on Technology and Innovation in Learning, Teaching and Education*. Springer, 153–165.

[39] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734.

[40] Michael I Posner, Mary K Rothbart, et al. 2007. Research on attention networks as a model for the integration of psychological science. *Annual review of psychology* 58 (2007), 1.

[41] Bart Rienties, Avinash Boroowa, Simon Cross, Chris Kubiak, Kevin Mayles, and Sam Murphy. 2016. Analytics4Action Evaluation Framework: A Review of Evidence-Based Learning Analytics Interventions at the Open University UK. *Journal of Interactive Media in Education* 2016, 1 (2016).

[42] Ian Roffe. 2002. E-learning: engagement, enhancement and execution. *Quality assurance in education* 10, 1 (2002), 40–50.

[43] Liam Rourke, Terry Anderson, D Randy Garrison, and Walter Archer. 1999. Assessing social presence in asynchronous text-based computer conferencing. *The Journal of Distance Education/Revue de l'ducation Distance* 14, 2 (1999), 50–71.

[44] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 2021. Note the highlight: incorporating active reading tools in a search as learning environment. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 229–238.

[45] Brenda Salley and John Colombo. 2016. Conceptualizing social attention in developmental research. *Social Development* 25, 4 (2016), 687–703.

[46] Tali Sharot and Elizabeth A Phelps. 2004. How arousal modulates memory: Disentangling the effects of attention and retention. *Cognitive, Affective, & Behavioral Neuroscience* 4, 3 (2004), 294–306.

[47] Michael F Shaughnessy. 2020. An Interview with Miriam Schcolnik: Reading, E-Reading and Writing and Their Assessment. (2020).

[48] Thomas B Sheridan. 2016. Human–robot interaction: status and challenges. *Human factors* 58, 4 (2016), 525–532.

[49] Jonathan Smallwood, Kevin S Brown, Christine Tipper, Barry Giesbrecht, Michael S Franklin, Michael D Mrazek, Jean M Carlson, and Jonathan W Schooler. 2011. Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PloS one* 6, 3 (2011), e18298.

[50] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[51] Mary Thorpe and Steve Godwin. 2006. Interaction and e-learning: The student experience. *Studies in continuing education* 28, 3 (2006), 203–221.

[52] Vicki Trowler. 2010. Student engagement literature review. *The higher education academy* 11, 1 (2010), 1–15.

[53] Konstantinos Tsiakas, Maher Abujelala, Alexandros Lioulemes, and Fillia Makedon. 2016. An intelligent interactive learning and adaptation framework for robot-based vocational training. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–6.

[54] Marieke Van der Schaaf, Liesbeth Baartman, Frans Prins, Anne Oosterbaan, and Harmen Schaap. 2013. Feedback dialogues that stimulate students' reflective thinking. *Scandinavian Journal of Educational Research* 57, 3 (2013), 227–245.

[55] Patrik Vuilleumier and Yang-Ming Huang. 2009. Emotional attention: Uncovering the mechanisms of affective biases in perception. *Current Directions in Psychological Science* 18, 3 (2009), 148–152.

[56] Liying Wang. 2018. Attention decrease detection based on video analysis in e-learning. In *Transactions on Edutainment XIV*. Springer, 166–179.

[57] James E Young, JaYoung Sung, Amy Voida, Ehud Sharlin, Takeo Igarashi, Henrik I Christensen, and Rebecca E Grinter. 2011. Evaluating human-robot interaction. *International Journal of Social Robotics* 3, 1 (2011), 53–67.

[58] Barry J Zimmerman. 2008. Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American educational research journal* 45, 1 (2008), 166–183.