Interactive Intelligence

Multimodal AI for Real-Time Interaction Loop towards Attentive E-Reading

Lee, Y.

Citation (APA)
Lee, Y. (2024). *Interactive Intelligence: Multimodal AI for Real-Time Interaction Loop towards Attentive E-Reading*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:764408e4-72c1-4cf9-8bff-1ce20b8944b2

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

Interactive Intelligence

Multimodal AI for Real-Time Interaction
Loop towards Attentive E-Reading

Yoon Lee

Interactive Intelligence:
Multimodal AI for Real-Time Interaction Loop towards Attentive E-Reading

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Tuesday, 27 February 2024 at 15:00 o'clock

by

Yoon LEE

Master of Science in Integrated Product Design,
Delft University of Technology, The Netherlands,
born in Seoul, Republic of Korea.

This dissertation has been approved by the promoters.

Composition of the doctoral committee:

Rector Magnificus,	Chairperson
Prof. dr. M. Specht	Delft University of Technology, Promptor
Dr. G. Migut	Delft University of Technology, Copromotor

Independent members:

Prof. dr. M.A Neerincx	Delft University of Technology
Prof. dr. H. Drachsler	DIFT, Germany
Prof. dr. M. Cukurova	University College London, United Kingdom
Prof. dr. I. Molenaar	Radboud University
Prof. dr. H. Jarodzka	Open University
Prof.dr.ir. G.J.P.M. Houben	Delft University of Technology, reserve member

Our greatest glory is not in never failing, but in rising up every time we fail.

–Ralph Waldo Emerson

Contents

1

General Introduction

E-learning has shifted the traditional learning paradigms in higher education, offering more flexible, ubiquitous, and personalized learning experiences. The previous years' COVID-19 pandemic required a re-calibration of education to accommodate virtual learning environments from the traditional classroom-based education [1]. Widespread learning platforms and digital devices have accelerated the adoption of e-learning [2], and now, it plays a central role in formal and informal education.

E-reading, a term used in this dissertation to describe digital reading on computers, has its unique position in higher education because extensive self-directed reading [3], information processing [4], knowledge comprehension [5], critical thinking [6], and knowledge reproduction and application [7] through reading, are required as a part of regular studies. Thus, it directly affects learners' self-efficacy [8, 9], learning effectiveness [1], and success [10]. However, despite its significance in daily higher education, e-reading support has yet to be implemented in previous studies, which is the focus of this dissertation.

Learning supports for e-learning have emerged as the focus of educational researchers and practitioners in the last decade, considering its specific environmental contexts. Unlike traditional on-site education, students can only interact with the interfaces without the physical presence of instructors and peers [11]. In this context, Self-Regulated Learning (SRL), which refers to learners' voluntary efforts to understand and control their education based on proactive goal setting, self-monitoring, self-instruction, and self-reinforcements, has been emphasized, which can further be benefited via implementing additional learning supports.

On the other hand, there are emerging opportunities in intervening e-learning, thanks to the prolifications of various sensor-based technologies and computers as co-existing ecosystems in e-learning. Though educators bring invaluable expertise, empathy, and contextual understanding to the educational experience, [12, 13], they are constrained by their capacity to interpret and respond to the myriad learner needs in real-time and at scale. However, insights regarding learning and learners can be captured through multiple data streams, processed based on multimodal reasoning, and turned into insightful feedback, leveraged by deep neural networks, complementing the existing limitations of human educators.

This work adopts a holistic approach, intertwining learning analytics and feedback provision as an iterative loop, utilizing machine learning as a means of multimodal reasoning and feedback via the computer and peripheral interfaces, such as Graphical User Interface (GUI) and speech-based Robot interface. The focus of this dissertation is two-fold: 1) successfully capturing learners' distractions in e-reading using multimodal indicators of human attention and machine learning technologies, and 2) assisting learners' behavioral, cognitive, and affective states based on the alignment of feedback interfaces, contents, traits, and timing. The feedback aims at learners' fewer distractions, and longer attention spans in e-reading as intervention objectives, exploring the possibility of interactive AI for e-reading.

1.1 Challenges of capturing learners' real-time attention and distractions in e-reading and approaches adopted in this dissertation

The main challenge in capturing learners' attention in e-reading has been creating and using systems that gather various types of data and develop machine learning models effectively and efficiently. Thus, this dissertation focuses on 1) defining learners' attention and finding indicators critical for data reasoning and 2) training the models for accurately capturing learners' attention and deploying them in parallel with the feedback for attentive e-reading.

The first challenge of designing indicators for multimodal reasoning comes from the fact that learners' attention has been understood based on various frameworks with varied definitions. At the intersection of cognitive science, education, and affective computing, multiple frameworks have strived to interpret learners' attention as mind-wandering [14], switches of inner thoughts [15], working memory [16], level of interest [17], and goal-directed thoughts [18]. Such segmental frameworks suggest varied definitions, concept coverage, and attention measures. This thesis defines learners' attention as consciousness toward an ongoing task without attention redirection and strives to find measures that could apply non-intrusively in e-reading.

As a means to effectively capture learners' internal states (e.g., cognitive and affective status), various observable indicators [19, 20], such as engagement [21, 22], affects [23, 24], and emotion [25], have been studied. Multimodal indicators, such as diverse parameters from eyes (e.g., pupil diameter, blinks, and saccades [14]), facial expression (e.g., valence and arousal [25]), and pose and gestures [21, 23, 24] have been commonly used as measures. Those data streams with the pre-implemented sensor arrays derived from its physical architecture were commonly combined with analytical methods and machine learning to predict learners' states and used to find critical components for attentive e-reading.

Therefore, this dissertation specifically focused on webcam-based computer vision methods based on learners' behavioral cues during e-reading to design the non-intrusive measure applicable to real-life e-reading scenarios. Other layers of behavior-based log data have been examined through case studies to find the best-performing model for predicting learners' attention, assisted through behavioral, cognitive, and affective supports in e-reading.

To address the practical deployment of the model for real-time interventions, various models based on image, video, and skeleton-based methods were experimented with through various case studies. The strengths and weaknesses of each method were com-

pared in perspectives of computational requirements, model performances, and real-time applicability in line with feedback. At the same time, the hybrid approach based on model fusion in practice has endeavored to perform attention recognition accurately in e-reading.

1.2 Challenges of designing and implementing real-time feedback for e-reading intervention and considerations in this dissertation

Real-time feedback design and implementation require considerations of various feedback components, such as feedback modality, interface, content, traits, and timing [26]. The first challenge of real-time feedback design and implementation for e-reading intervention comes from 1) evaluation of such feedback components is often context-specific, which means that one type of feedback works nicely in one scenario, while the same type of feedback does not add value to another scenario. However, feedback implementation for e-reading intervention has yet to be conducted, which this dissertation aimed at with case studies with empirical implementations and shared datasets as outputs.

Specifically, this dissertation strived to understand how feedback design can affect learning outcomes, experience, and perceptions based on various measures, such as pre-post knowledge test, AttrakDiff measure, and social presence measure. First, the effects of various interface types have been compared by implementing the GUI-based interface and speech-based robot interface with their meta-cognitive and empathic prompts. Also, the effect of the explainability of the feedback has been studied to understand the feedback component that affects the learning process and outcomes.

Another challenge focused on in this dissertation has been 2) the same feedback does not work for everyone, and learners have different learning needs and styles. In this regard, balancing generalization and personalization has often been an essential challenge in designing and implementing feedback. This dissertation strived for adaptive feedback design strategies to accommodate general feedback needs while adapting for specific learning needs to address the given challenge.

Based on seven works derived from three case studies, in line with multimodal data input stream, learning analytics, and machine learning, this work aimed to lay a foundational architecture for understanding how the AI-based interface can assist the e-reading experience of learners and lead them to better attention management.

1.3 Research questions

The general research questions for this dissertation have been drawn as follows:

- Main research question: How can a multimodal feedback loop, informed by automatic attention recognition, enhance e-reading experiences for higher education learners?

Sub-research questions have been articulated as below:

- RQ1. What are the state-of-the-art advancements and challenges in multimodal data aggregation, feedback design, and implementations in the context of Technology-Enhanced Learning (TEL)?

- RQ2. What theoretical and technical approaches can be taken to recognize learners' attention regulation in e-reading for higher education?

- RQ3. How can AI-based real-time feedback in e-reading assist attention management for higher education learners and further affect their learning outcomes, perceptions, and interactions?

In tackling RQ1 through **Part I**, a scoping review has been conducted to study multimodal feedback loops in Technology-Enhanced Learning (TEL) scenarios designed for various users. Based on the research, this work tried to align the multimodal data aggregation and feedback design applied across multiple learning domains.

Part I.

Chapter 2: Unveiling the Multimodal Feedback Loops: A Comprehensive Literature Review in Technology Enhanced Learning (TEL)

In tackling RQ2 through **Part II**, three studies have been conducted to design the indicators effectively to predict learners' distractions. Various behavior-based machine learning approaches were investigated, comparing the strengths and weaknesses of image, video, and skeleton-based recognition model developments, fusions, and deployments. Additionally, through an explainable AI approach, the work endeavored to find critical behavioral characteristics that contribute to recognizing learners' distractions and the usage of lower-order and higher-order thinking skills in the learning process.

Part II.

Chapter 3. Designing Indicators and Predicting Learners' Self-regulation Based on Behaviors: A Video-based Deep Learning Approach

Chapter 4. Investigating Behavioral Indicators for Predicting Learners' Higher and Lower-Level Thinking Skills: An Explainable AI Approach

Chapter 5. Data-Driven Persona Development and Automatic Recognition for Real-Time Applications: An Unsupervised Machine Learning Approach

---

Three studies have been conducted to tackle the RQ3 through **Part III**. They are mainly concerned with feedback strategies, considering the conversational agent as a feedback interface, and finding ways to implement feedback in an adaptive manner, considering the generalization and personalization aspects of feedback provision. Through the last paper, this dissertation designed and implemented the real-time feedback loop with the hybrid model and adaptive feedback. This work evaluated the system from the perspective of learners' behavioral, cognitive, and perceptive changes by adopting AI-based feedback in e-reading.

Part III.

Chapter 6. Feedback Design Strategies: The Impact of Conversational Agents and Empathetic & Metacognitive Feedback

Chapter 7. Designing Feedback Timing: Deep Learning-Based Attention Regulation Recognition and Real-Time Feedback Loop

Chapter 8. Real-time AI-based Feedback Loop Implementation and Its Impacts on Learners' Attention Span, Learning Outcomes, and Perceived Learning Experiences

The thesis concludes with a general discussion summarising the findings of all studies introduced in this dissertation with insights and design recommendations for the real-time feedback loop for e-reading. This work reflects on implications for practice and outlines directions for future research.

2

Unveiling the Multimodal Feedback Loops: A Comprehensive Literature Review in Technology Enhanced Learning (TEL)

Technology-enhanced learning systems, specifically multimodal learning technologies, use sensors to collect data from multiple modalities to provide personalized learning support beyond traditional learning settings. However, many studies surrounding such multimodal learning systems mostly focus on technical aspects concerning data collection and exploitation and therefore overlook theoretical and instructional design aspects such as feedback design in multimodal settings. This paper explores multimodal learning systems as a critical part of technology-enhanced learning used for capturing and analyzing the learning process to exploit the collected multimodal data to generate feedback in multimodal settings. By investigating various studies, we aim to reveal the roles of multimodality in technology-enhanced learning across various learning domains. Our scoping review outlines the conceptual landscape of multimodal learning systems, identifies potential gaps, and provides new perspectives on adaptive multimodal system design: intertwining learning data for meaningful insights into learning, designing effective feedback, and implementing them in diverse learning domains.

With the increasing application of Technology-Enhanced Learning (TEL), the educational roles of teachers and students are constantly changing [28]. The seismic shift was observed during the pandemic in the last few years, which forced the educational focus from traditional classroom learning to online and hybrid environments [28]. Owing to the proliferation of digital platforms and devices designed for educational purposes [29], TEL technologies have resulted in the availability of copious amounts of data on both the learner and their learning process. As a direct consequence, TEL technologies are being further enhanced with sophisticated Artificial Intelligence (AI), particularly Machine Learning (ML) techniques and Learning Analytics (LA).

Such technological advancements have reinforced the role of TEL, not only as a LA tool but also as a form of feedback agent in learning. For instance, the advent of ChatGPT[1] appears to bring transformative development in the field, as it has the potential to change the foundations of learning and education [30]. Although such interactions are currently limited only to text modality, information acquisition will become even more accessible via multiple sensory modalities, with the convergence of diverse speech-based conversational agents [31] and sensor technologies, in the form of multimodal interactions (e.g., Generative AI combined with VR agents) [30]. In this context, the importance of multimodality is not only confined to TEL as data input from the digital world [32], but also as outputs in both the physical and the virtual world, which can trigger cognitive, behavioral, and emotional changes in learners.

Multimodal learning systems, a subgroup of TEL, frequently employ multiple sensors and AI techniques to gather contextual learning data from diverse modalities to provide a comprehensive understanding of learning processes. This understanding can assist us, as practitioners and researchers, to reflect on the efficacy of the design of multimodal learning systems: how to digitize learning and learner information as data [29], how to process and intertwine multimodal data to best contextualize learning [29, 33, 34], and how to design and implement feedback and LA, also called multimodal learning analytics (MMLA) in learning systems to address students necessities [29, 35].

The field of MMLA combines different types of data from multiple modalities and sources to gain contextual insights into the learning process. Di Mitri et al. [32], in their conceptual framework called "Multimodal Learning Analytics Model (MLeAM)", portrayed multimodality in learning systems as a series of steps involving sensor capturing, annotation, predictions, and feedback implementation, in a loop. Although their conceptual framework has precisely aligned the multimodal data stream in input space, the framework has yet to be extended to the dimensions of feedback design and its implications for learning domains. In order to get insights into the design of feedback and MMLA, a critical component of TEL, we examine through a review how previous studies utilize multimodality in their learning systems from data collection to feedback implementation, which has yet to be collectively understood in previous research. Therefore, we investigate multimodal learning systems in three primary stages: 1) data collection and integration, 2) design decisions for the design of multimodal feedback, and 3) implications for system implementation in diverse learning domains. The following three research questions will be tackled by reviewing and analyzing studies in the field.

---

[1] https://openai.com/

- RQ1. How is multimodal data collected and processed to get insights about learning in MMLA?

- RQ2. How is learner feedback designed in the context of multimodal learning systems?

- RQ3. What are the considerations for implementing multimodal learning systems in various learning domains?

2.1 Methodology

The literature search was conducted with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach. The review itself was later adapted to a *scoping review* due to the erratic landscape of the multimodal learning systems we observed based on our preliminary searches, as our focus was on investigating the emerging topic, multimodality, as a critical component of TEL. Therefore, we adopted the five-stage approach of a scoping review of Arksey and O'Malley [36]: 1) identifying the research questions, 2) identifying relevant literature based on inclusion and exclusion criteria, 3) selecting studies, 4) analyzing and synthesizing the data, 5) and summarizing and reporting the results.

Through various search engines, such as Scopus and Web of Science, 1,794 search results were found based on keyword search (i.e. (multimodal OR multisensory) AND feedback AND (learning OR education)). Only results that included a description of learning systems designed for human users were selected, resulting in 274 papers. The results included papers from various subject areas, such as computer science, engineering, social sciences, psychology, and art and humanities. Six researchers further coded and filtered the remaining 274 papers in the eligibility check process with inclusion and exclusion criteria, such as having multimodal components as both the input and output of the system implementation. Using Cohen's Kappa coefficient by comparing observed and random probabilities, the inter-rater reliability among six coders' scores has been evaluated (Cohen's Kappa: good, $0.9 > 0.81 \geq 0.8$). The primary author solely proceeded with the rest of the overall review, and 27 papers were chosen for the final review. To compensate limitation of the PRISMA methodology caused by its strict application of inclusion and exclusion criteria, we applied the snowball method to extend the discussion with other relevant research in the field for a further scoping review, which resulted in 30 papers ranging from 2010 to 2023.

2.2 Results

The search conducted with the methodology described above resulted in 30 papers published between 2010 and 2023. Most of the systems in the resulting studies were based on a sensor-based approach and built for on-site learning than online learning. Similarly, they were often geared towards individual learning scenarios instead of collaborative learning. The majority of the selected studies' primary intervention was in the form of real-time feedback rather than post-hoc feedback, and most targeted learners more than teachers. A significant proportion of studies were conducted in K-12 education and higher education. Table 2.1 provides an overview of all the selected studies and their learning domains, data inputs, and feedback modalities.

2.3 Discussion

2.3.1 RQ1. How is multimodal data collected and processed to get insights about learning in MMLA?

Multimodal data collection in MMLA is performed using a host of sensors that correspond to the five primary modalities used by humans (i.e., visual, auditory, tactile, taste, and smell [37]), with various information layers, such as data types, frequencies, and resolutions. Our literature search yielded no studies that addressed the modalities of taste and smell, indicating a dearth of technology capable of capturing them. Of the 30 papers, six studies (20.0%) collected visual and auditory data, two studies (6.7%) collected visual and tactile data, two studies (6.7%) collected auditory and tactile data, and two studies (6.7%) collected all three of them. Tactile data has been most frequently used as the major data stream in eleven studies (36.7%), while visual and auditory data have been used in five (16.7%) and two studies (6.7%), respectively.

Sensor-based data collection

Visual and auditory sensors are frequently used to collect audio and video data. Visual data is collected using different types of cameras (e.g., webcam [38], infrared camera [31], motion capture camera [39]), which consist of various information layers such as RGB [22], shapes, sizes, and textures. Visual data is further processed, often with AI and ML techniques, for various purposes such as image recognition [40], facial expression analysis [41], gaze and posture analysis [42, 43], and trajectory tracking of the body [44, 45] and objects.

Auditory data is captured through the microphone, having volume and frequency as essential features. The human voice is commonly captured as an auditory modality that is used for corpus analysis [46], speech analysis, voice trait analysis [42, 47], and musical trait analysis [39, 48].

Tactile sensors, such as Inertial Motion Units (IMUs), are used to capture learners' physical movement and orientation detection [47, 49] while force trajectory is tracked via force sensors [44]. Additionally, *environmental sensors* collect information about the physical learning environments, such as temperature, humidity, noise level, and air quality [50]. More sophisticated *physiological sensors* technologies (e.g., eye tracker [31, 51], electroencephalogram (EEG)) are also applied for more accurate and deeper insights into the physiological state of the learner. Such sensors collect physiological information such as heart rate and skin conductance, which are further interpreted as clues of learners' stress levels, arousal, and emotional states [52]. For example, one dominant tendency in MMLA is the inclusion of physiological sensors to evaluate learners' affective states (e.g., cognitive load from pupil dilation and blinks based on eye tracking data [38, 51]). However, such applications are often criticized for their obstructiveness. To compensate for such limitations, remote detection technologies that are developed and implemented in affective computing can be utilized: assessing learners' bio-data based on vision-based detection (e.g., heart rate [53]) and behavior recognition algorithms [38], without having to have intrusive biosensors implemented, which allows more stealth monitoring of learner activities.

Table 2.1: List of papers included in the scoping review.

Lit.	Learning Domains	Data Input	Feedback Modalities
[46]	**Language Learning (Foreign)**: Personalized foreign language training	**Auditory:** Vocal traits and sentence formulation from microphone	**Visual / Auditory :** Dashboard (analytic dashboards), Text (paraphrasing), Voice (model voice, dialogue simulation)
[54]	**Medical Education :** Gamified palpation training	**Tactile :** Pressure sensitivity and timing from wearable glove (ParsGlove)	**Visual / Auditory :** Graphics (gamified task), Sound Effects (rewarding purpose for successful task completion)
[55]	**Medical Education :** Palpation analytics for training	**Tactile :** Orientation, position, and pressure of hand from wearable glove (ParsGlove)	**Visual :** Dashboard (real-time monitoring panel, feedback panel, post-hoc examogram with tables, graphs, and scores)
[56]	**Conceptual Learning :** Teaching properties of graphs and diagrams for visually impaired students	**Tactile :** Stylus orientation from Phantom Omni	**Auditory / Tactile :** Voice (task guidance), Physical Movements (Phantom Omni)
[44]	**Language Learning (First) :** Teaching handwriting for children	**Tactile :** Characters written on touch screen	**Visual / Tactile :** Graphics (cartoon face, color trajectory for guidance), Text (message alerts), Vibrations (error), Physical Movements (haptic rendering of correct trajectory)
[22]	**Conceptual Learning :** Gamified interface for better knowledge gain in collaborative learning	**Visual / Auditory :** Learner poses from a camera, Group corpus from microphone	**Visual :** Graphics (virtual farmland getting mature with better participation in the class)
[41]	**Clear Communication :** Practicing public speaking	**Visual / Auditory :** Behavior (gaze, facial expressions) from webcam and speech, voice activity from a microphone (Multisense)	**Visual :** Graphics (interactive virtual audiences), Dashboard (after-action report with analysis, containing text and graphs, and video recording)
[42]	**Clear Communication :** Presentation training	**Visual / Auditory :** Speech rate and filled pauses from six microphones, Gaze detection from six cameras (HAAR Cascade), posture skeleton (Kinect)	**Visual :** Graphics (symbolic icons-thumbs up and down with color coding, video capture of presentation), Text (evaluation summary with statistics)
[57]	**Conceptual Learning :** Immersive organic chemistry education with gamified VR	**Tactile :** Finger tracking via haptic glove	**Visual / Tactile :** Graphics (3D-gamified molecule graphic with color changes via VR headset), Vibrations (for approval or disapproval for task performance)
[58]	**Language Learning (First) :** Teaching reading to children with dyslexia	**Tactile :** Letters and spatial arrangements recognized by pogo pins	**Visual / Auditory :** 3D Physical Prototype alphabets with colored LED, Graphics (colored text graphics), Voice (associated sound for letters)
[49]	**Sports Education :** Dancing supports from expert performances	**Auditory / Tactile :** Movement features (timing, intensity) from a 3D accelerometer, electromyography sensor, and vocalization from a microphone	**Auditory :** Voice (Pre-recorded experts' vocalization mapped and synthesized with learner's movement)
[59]	**Conceptual Learning :** Haptic rendering for elementary science education	**Visual :** Images that student took using mobile phone cameras	**Tactile :** Vibrations (haptic rendering which represents the texture of photos taken)

2

Lit.	Learning Domains	Data Input	Feedback Modalities
[45]	**Sports Education**: Motion analysis for golf swing	**Visual**: Visual marker recognition from motion capture cameras	**Visual / Auditory**: Graphics (comparing color-coded trajectory traces with reference), Sound Effects (putting sound with different pitch, indicating performance levels)
[60]	**Medical Education**: Injection simulator for veterinary education	**Visual / Tactile**: Visual marker from AR tracking camera and orientation from gyro sensor on syringe simulator	**Visual / Auditory / Tactile**: Graphic (AR/VR graphic with simulator and vein with evaluative color coding), Sound Effects (bell chimes and dog barks as evaluation), Physical Movements (force from syringe simulator)
[38]	**Conceptual Learning**: Companion robot for e-reading in higher education	**Visual**: Webcam-based video recording on learners' behaviors	**Visual / Auditory**: Robot Gestures (facial expressions), Robot LED (approval, disapproval), Robot Speech (dialogue)
[51]	**Language Learning (First)**: Calligraphy trainer based on expert's handwriting	**Tactile**: Touch from tablet and muscle activities, gesture embedded accelerometer and gyroscope on Myo sensor, and eye tracker glasses from SMI	**Visual / Tactile**: Graphics (characters with colored trajectory guidance), Dashboard (summative evaluation with graphs, video recording), Physical Movements (pressure feedback with saturation)
[48]	**Musical Education**: Musical ensemble between human and machine	**Visual / Auditory**: Motion of cueing data from the camera, evaluating tempo from the microphone	**Visual / Auditory**: Graphics (projection of shadow-like pianist, Music (violin, viola, cello, double bass play in coordination)
[61]	**Sports Education**: Indoor rowing training with VR	**Visual / Tactile**: Hand movements from cameras and ergometer on handles	**Visual / Tactile**: Graphics (VR simulation for hand movement trajectories and gauge bar with color coding), Physical Prototype (feedback for hand position based on haptic markers)
[43]	**Clear Communication**: Automatic feedback for oral presentation skills	**Visual / Auditory**: Head gaze, posture detection from the camera, usage of filled pauses and volume from the microphone, pitches from text, and image size/density recognition from presentation slides	**Visual / Auditory**: Dashboard (post-hoc report with statistical scores and advice for weaknesses), Graphics (video, image recording), Audio (audio recording)
[62]	**Medical Education**: VR Epidural Administration	**Tactile**: Interaction with 3D-syringe with Novint Falcon, Leap Motion, and Touch 3D	**Visual / Auditory**: Graphics (virtual patient in a mixed-reality environment with Oculus Rift DK2, Dashboard (score, time completion, and number of attempts performed), Sound Effects (needle injection)
[40]	**Conceptual Learning**: Mathematics education for children with visual impairments	**Visual**: Image recognition based on markers (TopCodes)	**Auditory / Tactile**: Sound Effects (drum and piano), Voice (number reading, verbal rewards), Tangible user interface with braille (iCETA)
[63]	**Language Learning (First)**: Teaching handwriting for blind children	**Auditory / Tactile**: Audio pan and pitch represents from microphone and stylus gestures from Phantom Omni	**Auditory / Tactile**: Sound Effects (Pitch differences for drawing in the appropriate trajectory), Physical Movements (force feedback to minimize the distance with trajectory and playback feedback from Phantom Omni)
[64]	**Sports Education**: Virtual training for rowing skills	**Tactile**: Velocity of rowing tracking from accelerators	**Visual/Auditory/Tactile**: Graphic (immersive virtual reality for immersion and optimal trajectory guidance), Sound Effects (errors), Vibrations (errors)
[65]	**Clear Communication**: Presentation trainer for public speaking	**Visual / Auditory**: Body posture, hand gesture analysis based on a visual marker recognition from Kinect, voice volume, buffer, and pauses from the Kinect microphone	**Visual / Tactile**: Dashboard (mirrored video of the users), Graphics (symbolic icons for approval/disapproval), Text (evaluation and guidance), Vibrations (correction alert. through wristband)

Lit.	Learning Domains	Data Input	Feedback Modalities
[66]	**Language Learning (Foreign)** : Mobile tutoring system for English pronunciation	**Auditory** : Phoneme recognition from the microphone on mobile device	**Visual** : Graphics (flash card, emoji-based mouth shape animation), Dashboard (list of mispronounced words), Text (word-to-text with color-coded highlights)
[67]	**Medical Education** : Ultrasonography training	**Visual** : Position recognition of simulator from Kinect	**Visual** : Graphics (3D organ models and simulations, suggestive trajectory in guided mode)
[68]	**Clear Communication** : Presentation trainer	**Visual/Auditory/Tactile** : Body posture and movements from Kinect and voice volume, voice modulation, and phasing of speaking from microphone	**Visual** : Dashboard (posture analysis with skeleton, color-coded symbolic icons, suggestive text)
[39]	**Musical Education** : Violin play training	**Visual / Auditory / Tactile** : Video with visual markers from motion capture camera and Kinect, ambient audio from the microphone, and physiological (EMG) data from Myo sensors	**Visual / Auditory** : Dashboard (audio, video, and motion-capture recording with data visualization as post-hoc feedback)
[69]	**Medical Education** : Gross anatomy education for undergraduate medical students	**Tactile** : Stylus orientation from Phantom Omni haptic stylus	**Visual / Tactile** : Graphics (3D human anatomical structures), Physical Movements (rotation, touch feedback from Phantom Omni haptic stylus)
[70]	**Conceptual Learning** : Geometrical concept teaching for children with virtual 3D space	**Tactile** : Stylus orientation from Phantom Omni haptic stylus	**Visual / Auditory / Tactile** : Graphics (3D graphics with gamification), Sound Effects (immersive object moving sound), Physical Movements from Phantom Omni haptic stylus

With the current advancement of deep learning technologies, a subset of ML, and increased computational capabilities, high-resolution sensor data, with an unstructured data form, has become the resource of deep neural network developments. Deep neural networks can be used to make a sophisticated prediction about the learners' performance in LA, such as their attention prediction during e-reading [38], and provide personalized support. Before the emergence of deep learning technologies, only structured data with statistical explainability, such as log data from learning management systems, had been the target resources of traditional ML and LA [71]. However, dynamic data with uninterpretable patterns, such as image, video, sound, and text [72], are now widely used for the various model developments for the classification, prediction, and detection tasks [31]. It is expected that such emerging models developed based on unstructured or semi-structured information with non-numeric organizations, such as data from eye trackers and EEG data, will expand the horizons of MMLA.

Log-based data collection

Log data, collected through learners' interaction with learning management systems via mouse clicks, keyboard inputs, and touch, in quantified forms (e.g., number, frequency) [73], have been the traditional sources of data in LA. Such data collection is mainly done in online platforms, such as MOOCs [74], with bigger sample scales and broader demographics than the sensor-based data collection. Although the collection of log data is often more accessible due to the absence of complex hardware infrastructure and sensors and can be interpreted with relative ease [71], the insights from log data have been subject to criticism for its superficial interpretations [75]. Also, its smaller computation requirements make it suitable for extensive data collection at larger scales. With the emergence of generative AI, the log data, such as the discourse between learners and the system, will become much more valuable due to its potential for personalized chat-based learning assistants, which will become more common with current advancements in Transformer-based Natural Language Processing (NLP) models (e.g., GPT-4) [30].

Questionnaire data collection

The questionnaire is one traditional data collection method for learning analytics: evaluating learning on objective (e.g., knowledge gain) and subjective levels (e.g., learning experience). One common approach has been a pre-post questionnaire to measure the objective learning outcomes (e.g., knowledge gain) through task performances. With increasing emphasis on the User Experience (UX) in computer-assisted systems, more measures are developed and implemented for gauging the UX of the system (e.g., System Usability Scale (SUS) [76], Attrakdiff questionnaire [77]). Most existing measures have been developed especially for Human-Computer Interaction (HCI), which focus primarily on computer-based artifacts [78]. However, with the expansion of the physical and virtual ecosystems of TEL, there are emerging necessities for more standardized measures for evaluating the UX of various peripheral devices (e.g., AR/VR, conversational agents) and agents (e.g., virtual robots) [11]. Deciding the timing of questionnaire-based data collection (e.g., real-time, post-hoc) is one challenge, where researchers should balance the timely aspect and obstructiveness of the data collection.

Observation-based data collection

Observation-based data collection is typical where the targeted learners are not fully capable of expressing their own perspectives (e.g., children with intellectual disability) or experts' opinion takes an essential role in evaluation (e.g., evaluating collaborative learning [22]). In such cases, observers' evaluation of observable indicators becomes the means of gauging learners' learning progress and performances [29]. The evaluation objectives are often learners' internal states, such as affects, attention, and perceived experiences [24, 31, 38], that influence learning experiences and potential learning outcomes. Since the evaluation is dependent on third-person observation, having clear annotation standards and frameworks is essential for the validity of the data. However, in some cases, practitioners often design and execute the measures themselves without having solid standards or frameworks [78]. Another challenge comes from individual differences: behaviors occur differently due to cultural backgrounds and individual differences [24], such as behavioral or emotional expressiveness. Alternative methods of combining human annotations with other layers of ground truths are suggested to compensate for such limitations: implementation of biosensor data (e.g., eye tracker [31] electroencephalography (EEG) [24]) and collecting self-reported ground truths [38] from learners.

2.3.2 RQ2. How is learner feedback designed in the context of multimodal learning systems?

Multimodal Feedback		
Feedback Modality	Visual	Auditory	Tactile
Feedback Characteristics	Spacial / Temporal	Temporal	Temporal
Feedback Timing	Post-hoc / Real-time		Real-time
Feedback Functions	Semantic / Intuitive		Intuitive
Feedback Types	-Graphics -Dashboards -Text	-Sound Effects -Music -Voice	-Physical Movements -Vibrations

Figure 2.1: Multimodal feedback involves decision-making regarding the feedback modalities, characteristics, timing, functions, and specific types of feedback.

Multimodal learning feedback in the form of in situ real-time feedback has often been provided via physical components, such as touch-based devices, wearables, haptic devices, physical prototypes, and speakers. In our literature search, in situ real-time feedback was more predominant than post-hoc feedback, while some took the hybrid approach. Such feedback often employed intuitive pictograms, color-coding, sound effects, vibration, and force feedback to provide immediate responses as learning interventions. In the meantime, post-hoc feedback has often been provided as dashboards, narrative text, and personalized voice feedback. Learner dashboards continue to be a prominent tool for supporting self-regulated learning in LA and MMLA. Dashboards in LA provide an easy-to-understand visual representation of complex learning data in real-time, which allows educators and students to make informed decisions. Generally, the feedback design in MMLA includes the following design elements: feedback modalities, characteristics, timing, types, and functions (see Figure 2.1). In the following sections, we investigate the feedback elements found in the previous studies concerning their modalities.

Visual

Graphics   are the primary visual element with intuitive delivery. The realistic graphical features have often been combined with engaging virtual environments [57, 64], gamification elements [22, 54, 57] for specific learning objectives and contexts with enhanced immersion [70]. The symbolic features of the graphic have been used to communicate complex constructs. Symbolic pictograms, icons, and emojis are used [42, 65] to help reinforce or correct learners' behaviors. Visual effects, such as 2D/3D effects and color coding [42, 44, 45, 58, 66], are used for highlighting specific information in the visual message delivery. Additionally, motion graphics/animations are used to convey dynamic information, such as a reference for the model trajectory and movements [45, 64].

Dashboards   commonly use post-hoc visual language with extensive and collective information. For example, statistical analysis of learning progress and performances has been shown through data visualization via graphs [46, 47, 55], tables [46, 55], and gauge bars [61]. Multimodal learning systems track more sophisticated data from sensors capable of monitoring latent constructs in learners. Video [41, 51] and audio recordings [39] are used as feedback for summative evaluation via dashboards so learners can reflect on their learning.

Text   is used for its descriptive nature, capable of delivering narratives and details. It is a distinctive feature compared to other visual languages since the text relies on its semantic nature and the meaning layer, while visual languages mainly depend on intuitive understanding. Thanks to its clarity in message delivery, text feedback has often been used in dashboards for written descriptions [42, 65] and message alerts [44]. To differentiate the information hierarchy, some visual traits, such as font size [58], highlighting [66], and colors [42, 66], have partially been applied to texts.

Auditory

Sound effects   refer to types of auditory stimuli that are artificially made. Sound effects are used for positive feedback in dashboards for showing approval and rewards (e.g., bell chime [60]), while alerting sound effects are used for intuitively signaling learners for behavior corrections (e.g., dog barks [60]). Sound effects are also used for in situ real-time feedback [51] in multimodal learning systems for better immersion in certain educational scenarios (e.g., golf putting sound with different pitches [45]).

Voice   feedback has been commonly implemented for its semantic and phonetic features. Since lexical meaning can be delivered through voice messages, vocal instructions are given for the concept delivery [56], guidance [43], and dialogue simulations [46]. The acoustic features have been mainly emphasized for assisting pronunciations of second language learners and young learners [46]. Various tonal differences were applied to the vocal feedback to highlight specific information or certain sound units.

Music   has been implemented for musical education [39, 48] and context-giving for the immersions. Musical traits of learners' instrumental play, such as tempo, pitch, and timbre,

have been corrected by providing specific parts of the musical recording as guidance. Music can also create a certain ambiance with immersive visual aids.

Tactile
Physical movements have often been used for providing feedback on the psychomotor aspects of complex skill learning. For instance, model movements have been demonstrated for sports training (e.g., rowing [61, 64]) and delivering abstract concepts [40, 56, 57, 70]. Fine motor movements were given for handwriting education with the trajectory (e.g., handwriting [44, 51, 63]). The syringe prototype provided the force feedback [60], intertwining with physical probes for more effective veterinary training. Movement feedback has often been helpful for learners with visual impairments for compensating their limitations in visual knowledge acquisition [56].

Vibrations constitute the majority of tactile feedback found in literature, often referred to as vibrotactile feedback, in the form of small vibrations and frictions. Vibrations are simplistic and are not able to encode complex information. Vibrations have been implemented for concept delivery (e.g., texture rendering [59]), guidance (e.g., haptic trajectory [44]), and as corrective feedback (e.g., vibration buzzers [57, 65]). All vibration feedback, and other tactile feedback, have mostly been adopted as real-time feedback due to their temporal context-specific nature and, therefore, not used in the dashboards.

2.3.3 RQ3. What are the considerations for implementing multi-modal learning systems in various learning domains?

To answer RQ3, we analyzed the multimodal learning systems found in our literature according to the three learning domains from revised Bloom's taxonomy [79]: cognitive, psychomotor, and affective domains. The *cognitive domain* involves the development of our mental skills and acquiring knowledge. The *Psychomotor domain* relates to discreet physical functions, reflex actions, and interpretive movements of the human body, while the *affective domain* involves our feelings, emotions, and attitudes. Furthermore, we also cluster the learning systems according to their specific application domain and learning goals and present some of the largest clusters. It should be noted that it is not our intention to present learning systems as exclusive to one domain, and we only seek to categorize the learning systems according to their primary learning objective.

Cognitive domain
Conceptual Learning: Multimodal feedback loops for conceptual learning primarily focus on facilitating knowledge delivery and comprehension. Providing haptic feedback, in addition to visual-oriented course content, to demonstrate various physical phenomena has shown an enhanced understanding of the phenomena in learners [57, 59]. The inclusion of additional modalities in instructions of conceptual learning can assist learners with visual impairments (e.g., haptic feedback from a stylus on Phantom Omni[2] [56]).

Language Learning: Multimodality is also beneficial in various aspects of language learning, which has traditionally been considered a predominately cognitive domain. Im-

---

[2]http://www.immersion.fr/

proving pronunciation and intonation in learning a foreign language has been a commonly targeted learning objective through various methods, such as visual aids with ideal mouth movements [66] and audio feedback with standard pronunciation [46]. Children and learners with disabilities have been the main end user of learning systems for first-language learning. For children, teaching how to read [58], write with characters [44], and improve handwriting skills [44] has been the main focus. For writing tasks, force feedback has been commonly given through stylus [63, 69, 70] and colored trajectory feedback [44] to indicate learners' errors intuitively.

Medical Education: Multimodal learning systems for medical education have been implemented to compensate for textbook-oriented education, aiming at more practice-based learning. Yeom et al. [69] suggested a 3D visual and tactile education system offering vivid visuals and tactile structures of human organs for gross anatomy class. Similarly, Palpation education tools [54], ultrasonography simulator [67], injection simulators for human patients [62], and animals [60] have been designed to promote authentic real-life practices of such complex skills with mock-ups. Those mock-ups have embedded collective sensors and software for learning analytics and feedback so that learners can receive real-time feedback during their learning practices [54, 60]. Medical education systems also tend to involve physical props, mainly for tactile data collection and embedded performance assessment algorithms to provide real-time instructions and feedback.

Affective domain
Clear Communication Skills: Systems have been developed to improve clear communication skills during learners' presentations. With real-time evaluation, learners were asked to reflect on their performances and improve their skills over practice [42, 43, 65, 68]. Combining visual and auditory data collected from a webcam, microphone, and Kinect [42, 65, 68], learners' posture, gaze, facial expression, and voice traits for clear communication have been evaluated. Systems gave the correction in real-time, by short written descriptions [42, 65], real-time posture analysis [65, 68], and performance analysis on the dashboards as post-hoc feedback [42, 43, 65].

Psychomotor domain
Sports Education: In sports education or training, learning goals are predominantly psychomotor. As such, during sporting activities, real-time physical features have been evaluated: posture and strike patterns [45] for the golf swing, body orientation and posture [61] for rowing, and body movement [49] for dancing. These learning systems aim to correct learner errors in real time and offer actionable plans to improve learners' motor skills. As motor-skills development demands conscious repetitive practices [80] where learners rely on apprenticeship-based education, systems can computationally model experts or mentors [51] and use ML to provide real-time feedback.

Musical Education: Systems in musical education were implemented to support human-machine ensemble [48] and violin play [39]. Based on visual and auditory indicators collected from a camera and a microphone, the phasing of violin play was analyzed on the dashboard [39]. To support the human-machine play [48], the shadow visual of the

pianist and pre-recorded music piece has been played along with the learner's play during repetitive practices. Through visual aid, learners were taught to understand the current issues, correct errors, and internalize better techniques in an analytic and reflective manner.

2.3.4 Challenges and Opportunities

Generalization vs. Personalization of multimodal learning systems.   While most systems aim at the best generalizability in the application, more and more learning systems target feedback provisions with personalization since one system should be general enough to cover targeted user groups while it should effectively reflect individuals' critical learning necessities. In this regard, future studies for refining the generalizability and personalization of critical learning necessities, timing, frequencies, and effects in multimodal learning systems design would greatly benefit the community.

Overarching MMLA frameworks for higher-level learning objectives.   Multimodal learning systems are often modeled as domain-specific and context-based since most systems aim to improve concrete learning activities with clear system goals. However, in many cases, such goals are set based on fragmentary frameworks, lacking overarching models for higher-level learning objectives that can be universally applied to general domains or even domain-specific instructional design. Having such an overarching framework could work as a common ground where practitioners and researchers can exchange and share their knowledge and grow as a community while defining learning features is often the biggest challenge in MMLA with advancements in a data-driven approach.

Closing the feedback loop in MMLA.   Our findings suggest that despite the advances in AI and ML algorithms, multimodal learning systems often fail to close the feedback loop. Though the systems we examined in our study included feedback in the system loop, most MMLA systems in the field need to take the current analytics into the context of the next round of feedback provision. In this sense, closing the feedback loop based on various modalities and evaluating the effect of the feedback loop for further optimization seems to be an essential challenge in the field.

2.4 Conclusion

In this scoping review, we investigated multimodal learning systems, an integral extension of modern TEL systems. We investigate systems in three stages as an extension to the MLeAM framework: 1) multimodal data collection and processing, 2) multimodal feedback design decisions, and 3) multimodal system implementation for various learning domains. The result indicates the necessity of a more holistic understanding of the whole process in order to design effective systems and multimodal instruction patterns. We also identified critical challenges in multimodal learning systems, such as defining learning indicators, balancing the generalization and personalization of analytics and interventions, and closing the feedback loop in multimodal learning systems. Our paper provides an overview of the role multimodality plays in defining the potential of the next generations of TELs and outlines important considerations for data collection, feedback design, and MMLA design for adaptive TEL system implementations.

With more evidence-based, data-driven approaches taken in LA, the quality of data is getting increasingly important, especially in the context of MMLA. Although data is becoming more accessible through sensors on commercialized devices (e.g., laptops, webcam [38]) and increasing public datasets, engineering competencies are becoming more critical in MMLA [34] as how data is collected and processed impacts the quality of data and therefore, the predictions it makes. Based on our analysis of RQ1, MMLA builds upon, rather than replacing, traditional LA but using data from multiple modalities. By doing so, MMLA is able to make more robust predictions about learners' performance across multiple domains, as evidenced by RQ3, and can also provide more personalized feedback. However, even with the increased roles of data engineering and advancements in AI, researchers' insights, experiences, and domain knowledge are still critical [34]. For example, the black-box nature of ML models makes decisions and predictions in MMLA often not explainable and requires human interpretations. Explainable AI (e.g., tree-based models) [72] can supplement the current MMLA in partially addressing this issue by providing better interpretability of such analysis. This also holds true for the LA dashboards, as evidenced by RQ2, with the majority of multimodal learning systems still relying on the affordances of traditional LA dashboards, which support necessities for the stronger MMLA and feedback design based on multimodalities.

3

3

Designing Indicators and Predicting Learners' Self-regulation Based on Behaviors: A Video-based Deep Learning Approach

Human attention is a critical yet challenging cognitive process to measure due to its diverse definitions and non-standardized evaluation. In this work, we focus on the attention self-regulation of learners, which commonly occurs as an effort to regain focus, contrary to attention loss. We focus on easy-to-observe behavioral signs in the real-world setting to grasp learners' attention in e-reading. We collected a novel dataset of 30 learners, which provides clues of learners' attentional states through various metrics, such as learner behaviors, distraction self-reports, and questionnaires for knowledge gain. To achieve automatic attention regulator behavior recognition, we annotated 931,440 frames into six behavior categories every second in the short clip form, using attention self-regulation from the literature study as our labels. The preliminary Pearson correlation coefficient analysis indicates certain correlations between distraction self-reports and unimodal attention regulator behaviors. Baseline model training has been conducted to recognize the attention regulator behaviors by implementing classical neural networks to our WEDAR dataset, with the highest prediction result of 75.18% and 68.15% in subject-dependent and subject-independent settings, respectively. Furthermore, we present the baseline of using attention regulator behaviors to recognize the attentional states, showing a promising performance of 89.41% (leave-five-subject-out). Our work inspires the detection & feedback loop design for attentive e-reading, connecting multimodal interaction, learning analytics, and affective computing.

K eeping a high level of attention is considered a prerequisite for successful learning, being associated with more effective (e.g., comprehension), efficient (e.g., efforts put per time), and appealing (e.g., duration of engagement) learning experiences and outcomes [81–83]. In this regard, in the fields of learning sciences, multimodal interaction, and affective computing, there have been attempts to measure learners' real-time attention with mind-wandering [14], switches of inner thoughts [15], working memory [16], level of interest [17], and goal-directed thoughts [18]. In this work, we define attention as consciousness towards an ongoing task without an attention redirection. With various sensors and model implementations, attention management through real-time feedback loop design has been endeavored [11, 84]. Significantly, the current transition to hybrid and online learning environments during the pandemic has accelerated the need for attention detection and management in diverse e-learning scenarios.

In e-learning, learners' attention management is different from what they have had in the traditional classroom [85], with limited human educators' involvement and the lack of timely intervention accordingly [86]. Therefore, attention management in e-learning has been highly dependent on learners' self-regulation compared to on-site learning [85]. During e-learning practices, learners experience several iterations of attention fluctuations [87, 88]. In the process, learners recognize their own distractions and try to re-engage in their tasks [89] as a voluntary attentional control [90]. In this work, our focus is on finding learners' self-regulatory behaviors based on learners' own awareness, which leads to self-regulatory efforts to sustain a good level of attention. We define such behaviors as "attention regulator behaviors": Learners' earliest self-awareness of attention loss and following observable behavioral changes as self-regulation. We find those moments important since those are the moments that learners are willing to and are still able to re-engage in their learning tasks.

In previous studies, diverse multimodal cues have been investigated as observable predictors of subjects' diverse internal states (e.g., cognitive and affective status), such as attention [19, 20], engagement [21, 22], affects [23, 24], and emotion [25]. However, they were often criticized for being difficult to measure or interpret. Iris extension, gaze direction, the position of hands and legs, the style of sitting, walking, standing or lying, body posture, and movement are known to be relevant behaviors for a person's internal states [25, 91, 92]. Diverse parameters of eyes, such as pupil diameter, blinks, and saccades [14], have often been directly used to assess learners' attentional states with dedicated eye trackers. Learners' valence and arousal were often understood primarily through facial expressions [25], with expansion with sensors, such as a photoplethysmograph (PPG), Galvanic Skin Response (GSR), Electroencephalography (EEG), and Electrocardiography (ECG) [93]. Poses and gestures have been interpreted as means to assess engagement [21], affective and cognitive states [23, 24].

However, the current framework has shown that the interpretation of internal states should be understood within the context [94] on macro (e.g., cultural) and micro levels (e.g., situational, personal features) [95]. It indicates that specific cues can be significant indicators of attention in one learning activity, while the same cue does not necessarily represent the same in the other type of learning activity. In this sense, we choose to collect a novel dataset in an e-reading scenario with cognitive and behavioral parameters, which we hypothesize interlinking with attentional changes in e-reading. We chose e-reading since

it is the most common and fundamental form of e-learning practice in higher education, which can support other learning activities. This work focuses on which attention regulator behaviors occur following the perceived distraction via the statistical analysis and model implementation. We hope this interdisciplinary study can nurture an understanding of attentive e-reading. Our contributions to the field are listed below.

- To our best knowledge, it is the first attempt to introduce attention regulator behaviors in e-learning for attention analysis and prediction. Compared to conventional subjective measurements of attention, such as self-report, the attention regulator behaviors are easier and intuitive to capture and more objective to evaluate.

- We collected a novel dataset, WEDAR, from 30 subjects with various metrics, including learner status, affects, behaviors, and learning outcomes. Self-report of distraction is also provided as ground truths to verify the effectiveness of the attention regulator behaviors as a predictor.

- Diverse machine learning models are implemented as a baseline to recognize learners' attention regulator behaviors and attentional states. Those baselines can further be applied to diverse e-reading system designs.

- The framework provides a webcam-based attention analysis. It does not require dedicated hardware implementation for obtaining the attention recognition features and can thus be applied to diverse real-world settings.

3.1 Related work
3.1.1 Attention "regulator" behaviors
Diverse learning theories have been constructed to understand learners' internal states through various tangible predictors. Our work is based on the framework of [96], which focuses on how diverse *stimuli* (e.g., external condition, verbal representation, awareness, intentionality, external feedback, delivered information) can be *interpreted* (e.g., arbitrary, iconic, intrinsic) and connected to *functional* nonverbal behaviors (e.g., emblems, illustrators, regulators, affect displays, adaptors).

According to the behavior categorization of [96], "regulator" behaviors occur as a self-regulatory action with the purpose of successful task performance (e.g., head nods, eye contact, slightly forwarded body, small postural shifts, and eyebrow raises in human-to-human interaction). Those are subconscious and habitual actions triggered by behavior agents' "awareness" of their internal and external states (e.g., attention loss). We hypothesize that such self-regulatory behaviors (i.e., attention regulator behaviors) also occur in e-reading. In this work, we try to define the types and frequencies of attention regulator behaviors in e-reading. The framework of [96] also indicates the expandability of their categorical framework, which supports our attempt.

3.1.2 Multimodal attention recognition in real-world e-reading settings
Previous research has highlighted the importance of contextual interpretation of multimodal indicators [95]. Instead of finding global features for attention in diverse learning scenarios,

we explicitly investigate theoretical and empirical behavioral cues of attention regulation in e-reading. We investigate a data collection method that is non-intrusive and closer to real-world settings, which allows a more widespread application of our framework in diverse e-reading scenarios. In the following, we introduce previous research aimed explicitly at real-world implementation based on webcam and mouse-click.

[19] aimed for the subject-independent model development in e-reading based on eyebrow, lip, head movements, and mouse orientation. Specific behaviors (e.g., leaning forward) have been combined and labeled as more generic categories (e.g., body) to avoid overlapping features in different classes. [20] focused on head orientation, eyelid and mouth height, gaze direction, and emotion (i.e., confusion and happiness) during e-reading. Six hand-labeled attention levels (i,e., sleepiness, drowsiness, fatigue, distraction, attention shift, concentration) has been used as ground truths. However, we assume that each attention class is not exclusive enough to the other, so there is a high chance that the machine can not classify different attention levels with higher performance. According to our best knowledge, very little empirical work has been done for attentive e-reading, which premises real-world settings.

3.1.3 Multimodal attention regulator behaviors

This section explicitly explores multimodal learning behaviors that function as attention regulators in e-reading. Instead of investigating features that can be found with dedicated sensors and devices, we focus on features recognizable to observers.

Eyebrow. The movements of eyebrow has been associated with the activation of cognition [23, 97], arousal [98] and emotions [97, 98], having most of the framework applied to social communication with rare empirical studies [99]. Though eyebrow movement is often observed in e-reading, only several empirical works indicated that eyebrow movements correlate to attentional changes [19]. As far as we know, theoretical behavior frameworks dedicated to e-reading have not been established yet. [100] understood eyebrow movements as ritualized behavior of attention signals, while [97, 98] interpreted it as sign of "wanting to know more", which is connected to the cognitive arousal. The framework of [97] defined eyebrow movements as a representation of surprise, question, and fear. In this work, we focus on the arousal function of eyebrow movements that are shown with combinations of inner and upper brow raise and lowering movements [98]. With a few solid evidence, we hypothesize eyebrow movements as voluntary self-regulatory behavior to re-engage in the task, aiming at a better attention level.

Blink. Correlations between various eye movements and cognitive actions have been revealed in diverse task performance scenarios, such as reading, scene perception, and visual search [101]. Based on the environmental task demands, humans are known to adapt their blink patterns spontaneously, voluntarily, and reflexively [102]. In e-reading, a reduced blinking rate by 4% has been observed with higher perceived fatigue, compared to paper-based reading, having dry eyes and eye discomfort as major causes [103]. As a result of fatigue, changes in blink patterns in frequency and duration are observed [104]. Blink flurries, which are defined as three or more blinks within a 3-second window, occur [103] and are interpreted as a spontaneous effort to sustain the attention and increase wakefulness [105]. Voluntary prolonged blinks are observed as a behavior to reduce the fatigue levels in eyes [106], showing different ranges in interblink interval variability,

degree of completeness, duration of the closure, and the force involved, compared to spontaneous blinking [107].

Mumble. Verbalization during reading is one learning strategy known to help readers' cognitive processing, reading development, and comprehension [108]. Verbalization is also known as read out loud, oral reading, and mumble reading, allowing readers to focus and monitor their real-time comprehension, as opposed to silent reading [109]. We use the term "mumble reading" in this work since our target behavior does not indicate active usage of the verbalization technique. However, it is more inclined to semi-spontaneous mumble behavior as a self-regulatory action to achieve better attention. Mumble reading is more commonly applied to teach young learners. However, it is also known to assist adult learners with decoding difficult passages. By mumbling the text, learners internalize the meaning and information of the sentence as coherent sets [109] with auditory stimulation. Diverse eye movement patterns are known to be correlated with mumble reading behavior [110]: Mumble also works as a stimulus to blink [103], showing internal consistency as an attention regulator behavior.

Hand. Self-touch is known to be an action that re-engages people's attention by soothing themselves during stressful moments, causing self-enjoyment [111]. Aside from the stress-release effect, self-touch during the task performance is known to bring better self-regulation, too [111]. Inhibiting effect from such tactile stimulation helps learners ignore distractions and refocus on the task [112]. Especially when working on a task that demands working memory with the presence of distractors, more spontaneous self-touches on the face tend to appear with the increasing necessity of refocusing [113]. Self-touch should also be interpreted within the context since certain self-touching behaviors lead to relaxation, while others work as arousal (e.g., self-squeezing, rubbing, scratching, stroking) [111]. Therefore, we define calming self-touching behaviors on the body and face as one category of attention regulator behaviors.

Body. While the face delivers more information about types of emotions, the body is known to convey affects and intensity [96] of emotions via diverse amplitude, speed, and fluidity of movements [114]. In previous studies, body postures have shown a direct correlation with attentive [115] and affective states [24]. The direction of the body is known to imply the affective states, such as boredom, confusion, delight, flow, and frustration [24] while the leaning forward pose works as a sign of active cognitive state [23]. Head direction indicates the subject of attention [116]. [117] understood postural shift as an action to move on to the next phase during the task performances.

Distraction self-reports. Distraction self-reports are commonly used as the ground truth to reflect people's internal states [118]. The model of [88] introduce two types of distraction: 1) Task-related distraction and 2) task-unrelated distraction. [88] explains that task-related thoughts are correlated with the objective performances, while task-unrelated mind wandering functions as an impairment to the ongoing task performances. In this regard, we collect two types of distraction self-reports in e-reading.

Based on the execution, distraction reports are two types. The first method is to collect distraction reports real-time at the choice of participants during the task performance, putting more importance of more timely aspect of distraction reports. The second method uses a specific time or event to trigger the question regarding the current distraction levels [118]. The first method is criticized as participants might not be aware of their attention

loss or forget about reporting. The second method is faulted for bothering the primary task performance.

We implemented the first method since our objective of the distraction self-reports collection is to find behaviors at the moment of learners' perceived distraction, which is used as the ground truth of the model training in our work. To minimize the possible intrusiveness in the self-report process, we carefully designed a simple and intuitive self-report interface, introduced in the following "Distraction self-reports" section. In this way, we obtained the ground truth of the attention levels of every subject through their frame-level distraction self-reports.

3.2 WEDAR-dataset
3.2.1 Participants
30 learners (gender: 15 males, 15 females; age: M=27.89, SD=3.39) in higher education, who use the English language for their daily education, have been invited for an e-reading task. Participants voluntarily joined the experiment via an advertisement on campus.

3.2.2 Materials
The text "how to make the most of your day at Disneyland Resort Paris" has been implemented on a screen-based e-reader, which we developed in a pdf-reader format. An informative but entertaining text was adopted to capture learners' attentional shifts during knowledge acquisition. The text has 2685 words, distributed over ten pages, with one subtopic on each page (e.g., how to book tickets online the same day). The e-reader has been implemented on a 13-inch laptop monitor with resolutions of 960×720, having the text with 11 pt. A built-in webcam on Mac Pro and a mouse have been used for the data collection, aiming for real-world implementation only with essential computational devices. A height-adjustable laptop stand has been used to compensate for participants' different eye levels.

3.2.3 Measurements
We collected various cues that reflect learners' moment-to-moment and page-to-page cognitive states to understand the learners' attention in e-reading. Fig 3.1 shows an overview of measurements used in the WEDAR dataset collection.

Distraction self-reports. Learners were asked to report their distractions on two levels during the reading: 1) In-text distraction (e.g., still reading the text with low attentiveness) or 2) out-of-text distraction (e.g., thinking of something else while not reading the text anymore). We implemented two noticeably-designed buttons (33×22) on the right-hand side of the screen interface to minimize the possible distraction coming from the reporting task.

Blur stimuli. We implemented blur stimuli on the text in the random range of 20 seconds after the trigger of a new page. It ensures that the blur stimuli occur at least once on each page. This is based on the finding that average learners read 230-250 words per minute [119]. Participants were asked to click the de-blur button on the text area of the screen to proceed with the reading. The button has been implemented to the whole text area, with 400×480 resolutions, so participants can minimize the effort to find and click the button. Reaction time for de-blur has been measured, too, to grasp the arousal of learners during
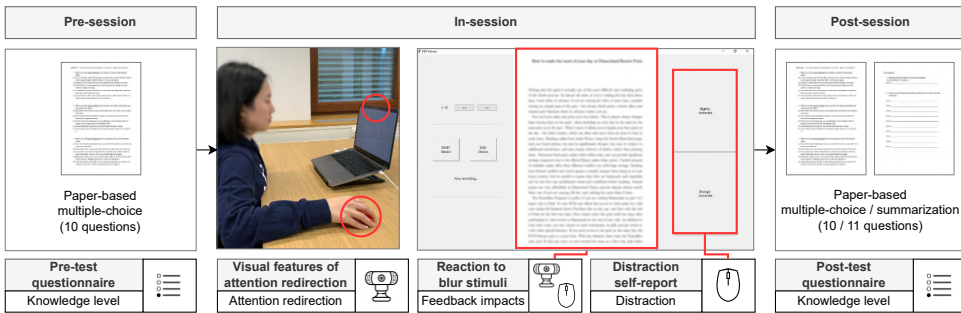
Figure 3.1: The experiment settings show an overview of our WEDAR dataset collection.

the reading.

Pre-test and post-test. We asked participants to answer pre-test and post-test question-naires related to the reading material. Participants were given ten multiple-choice questions before the session, while the same set of questions was given after the reading session (i.e., formative questions) with added subtopic summarization questions (i.e., summative questions). It can provide insights into the quantitative and qualitative knowledge gained through the session and different learning outcomes based on individual differences.

3.2.4 Procedure

30 learners in higher education have been invited for a screen-based e-reading task (M=16.2, SD=5.2 minutes). A pre-test questionnaire with ten multiple-choice questions was given before the reading to check their prior knowledge level about the topic. There was no specific time limit to finish the questionnaire. Afterward, instructions on secondary tasks were given: 1) Deactivating the blur stimuli on the screen by clicking the text area and 2) reporting distractions (i.e., in-text distraction, out-of-text distraction). Learners were left alone in a room to perform a screen-based reading task. Once participants finished the reading, they were given a post-test questionnaire with the same question set as the pre-test. However, in the post-test questionnaire, there were added questions for summarizing ten subtopics by filling in the sentences starting with " How to…".

3.2.5 Dataset: WEDAR

The final outcome of the WEDAR dataset is presented in Table 3.1, including the objectives of data collection, modalities, features, evaluation, and interpretation. In this work, indicators in **bold** are used for the attention regulator recognition and attention prediction. Note that the WEDAR is built not only for attention regulator behavior recognition but, more importantly, for exploring the learners' attentional states during the e-reading events. Thus, we collected various metrics as cues of the learners' attentional states, such as reactions to stimuli, distraction self-reports, and knowledge gain. All those metrics were obtained by learners' self-reports. The annotation is frame-level (one value for one reading case) for the metrics of reactions to stimuli and distraction self-reports. The annotation is instance-level for the metric of knowledge gain, which has been measured before and after the reading.

Table 3.1: Our WEDAR dataset contains diverse dimensions of attention: Objectives, modalities, features, evaluation, and interpretation of attention indicators.

Objectives	Modalities	Features	Evaluation	Interpretation
Learner behaviors	Video (avi.)	-Affective states of learners -Behavioral states of learners	Objective/ Subjective	Short-term attention
Reactions to stimuli	Timestamp (txt.)	-Blur triggered -Blur deactivated -Reaction time	Objective	Short-term attention
Formative & summative assessment	Text (txt.)	-Pre-test (multiple-choice) -Post-test (multiple-choice, summarization) -Knowledge gain	Objective	Long-term/holistic attention
Distraction self-reports	Timestamp (txt.)	**-Distraction in the context of reading -Distraction outside the context of reading**	Subjective	Short-term attention

3.3 Data analysis and results

This section presents preliminary experimental results conducted on the WEDAR dataset. We first report a relevant statistical analysis of the WEDAR dataset using Pearson's correlation coefficient. Several classical models are implemented as the benchmark for recognizing different attention regulator behaviors. Lastly, high-level attention analysis is conducted using the attention regulator behaviors and attention span.

3.3.1 Annotation and baseline analysis

Annotation of attention regulator behaviors. The video dataset of 931,440 frames has been annotated with the attention regulator behaviors using an annotation tool that plays the long sequence clip by clip, which contained 30 frames. Two annotators (doctoral students) have done two stages of labeling. In the first stage, the annotators were trained on the labeling criteria and annotated the attention regulator behaviors separately based on their judgments. In the second round, the labels were summarized and cross-checked to address the inconsistent cases. We used six categories that we found to be relevant to attention regulation based on the literature study: Behaviors shown from eyebrow (26,535 frames), blink (17,761 frames), mumble (22,214 frames), hand (101,700 frames), and body (155,880 frames), contrary to the neutral (607,350 frames) state (Figure 3.2). Since the importance of our work is not merely on the recognition of behaviors itself but on connection with hidden behavioral functions (e.g., attention regulator) [91], we combined multiple specific behaviors (e.g., squint) into a general category (e.g., eyebrow). It also helps avoid redundancy among features [19] which could negatively affect the model's performance. The labeled data has been used as two input formats: images segmented by each frame and videos segmented every second (30 frames).

3.3.2 Preliminary analysis: Pearson's correlation

We conducted a preliminary second-to-second analysis using Pearson's correlation among the overall, in-text, out-of-text self-reported distractions and attention regulator behaviors. We aimed at comprehensive insights into how each behavior category can be correlated to perceived distractions. As can be seen from Table 3.2, the total number of self-reported distractions and in-text distractions showed a significant correlation with eyebrow and body behavior categories. Out-of-text distraction has correlated with the most behavior categories: Eyebrow, blink, hand, and body. Though mumbling did not directly correlate
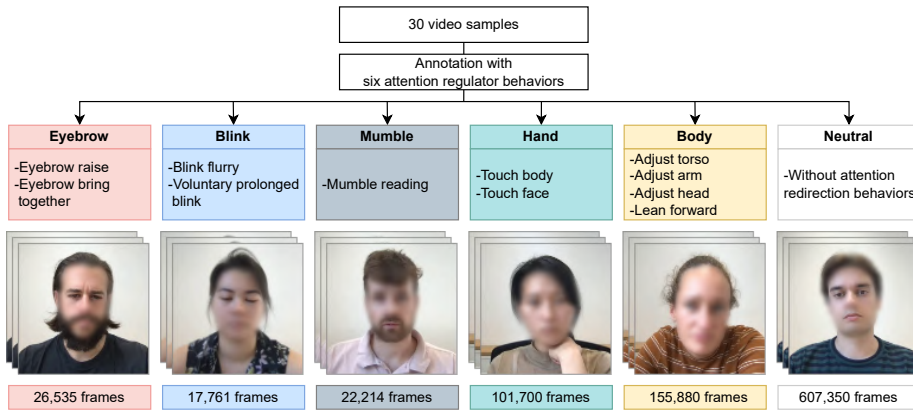
Figure 3.2: Annotation of attention regulator behaviors: Eyebrow, blink, mumble, hand, body, and neutral.[1]
[1] Images were blurred for identity protection purposes. All images were consented to be used for publication.

with any types of distractions, it has been correlated with other behavior categories, such as eyebrow, hand, and body. Various behavior categories have shown correlations among each other. The unimodal correlation analysis based on Pearson's correlation coefficient has presented: 1) The internal consistency among the attention regulator behaviors and 2) the potential of attention prediction model training based on multimodal behavioral cues related to attention self-regulation. Note that Pearson's correlation coefficient is a preliminary examination that only shows the linear correlation of two variables, revealing their potential association in the temporal domain. However, when it comes to attention regulator behavior-based distraction recognition, the performance might vary greatly because the relationship between attention regulator behaviors and distraction level is complex and non-linear, which cannot simply be described by Pearson's factor.

3.3.3 Low-level attention regulator behavior recognition

We propose the benchmark of classical models with two types of frameworks (i.e., frame-level and video-level recognition) on the WEDAR dataset to first recognize attention regulator behaviors.

Here, we followed the classical 70%-30% protocol from other large-scale action/activity datasets, such as ActivityNet [120] and Kinetics-400 [121]. Given 30 video samples with frame-level annotations (931,340 frames), we aimed to recognize the six attention regulator behavior categories accurately. Besides, we conducted an evaluation with both subject-dependent and subject-independent protocols. In subject-dependent protocol, we randomly shuffled all the samples and split the training and testing set with a ratio of 70% and 30%. In subject-independent protocol, we split the subjects with a ratio of 70% and 30%. Thus, all the samples from 21 subjects were used for training, and samples from the remaining nine subjects were used for testing. Note that we used the same protocol and evaluation settings for all the evaluation methods to make a fair comparison. Table 3.3 shows the overall accuracy of the testing set.

Table 3.2: Preliminary Pearson's correlation[2] analysis between distraction self-reports and attention regulator behaviors.

[2]Note. * p < .05, ** p < .01, *** p < .001, [3]Total distraction=In-text distraction+Out-of-text distraction

		Total distraction[3]	In-text distraction	Out-of-text distraction	Eyebrow	Blink	Mumble	Hand	Body
Total distraction	Pearson's r (p-value)	-							
In-text distraction	Pearson's r (p-value)	**0.938 ***** (**<.001**)	-						
Out-of-text distraction	Pearson's r (p-value)	**0.342*** (**<.001**)	-0.004 (0.469)	-					
Eyebrow	Pearson's r (p-value)	**0.030*** (**<.001**)	**0.021*** (**<.001**)	**0.028*** (**<.001**)	-				
Blink	Pearson's r (p-value)	0.019 (0.001)	0.011 (0.055)	**0.025*** (**<.001**)	**0.025*** (**<.001**)	-			
Mumble	Pearson's r (p-value)	0.004 (0.518)	0.006 (0.274)	-0.006 (0.270)	**0.041*** (**<.001**)	-0.005 (0.440)	-		
Hand	Pearson's r (p-value)	0.006 (0.325)	0.002 (0.783)	**0.012*** (**0.036**)	**0.053*** (**<.001**)	**0.028*** (**<.001**)	**0.045*** (**<.001**)	-	
Body	Pearson's r (p-value)	**0.045*** (**<.001**)	**0.035*** (**<.001**)	**0.035*** (**<.001**)	**0.095*** (**<.001**)	**0.044*** (**<.001**)	**0.037*** (**<.001**)	**0.375*** (**<.001**)	-

Table 3.3: Attention regulator behavior recognition performance on the test set of the WEDAR.

Method	Framework	Accuracy (%)	
		Subject-dependent	Subject-independent
ResNet-18 + fine-tuning		39.76	25.90
ResNet-50 + fine-tuning		30.92	23.84
ResNet-101 + fine-tuning	Frame-level	31.26	16.39
ResNet-18 + kNN		69.98	18.43
ResNet-50 + kNN		69.95	18.23
ResNet-101 + kNN		69.73	15.76
CNN-RNN-imbalanced	Video-level	**75.18**	**68.15**
CNN-RNN-balanced		**75.70**	**68.43**

Frame-level attention regulator behavior recognition. In this section, we conduct the attention regulator behaviors recognition using frame-by-frame image inputs. We implemented ResNet architecture as the backbone with its three variants (ResNet-18, ResNet-50, ResNet-101) [122], which are pre-trained on ImageNet [123], and fine-tuningd by fixing the layers $1000d\ fc$ and above. ResNet architecture became one of the most popular architectures in various computer vision tasks. Its shortcut connections architecture yields compelling results. First, each frame has been converted to 224×224 grid RGBs as image inputs. The higher-level features have been extracted with the layer going deeper, combining primitive features from images on earlier layers. To avoid the imbalanced data issue brought by a large number of neutral behaviors, we evenly sampled each category based on the class with the minimum category number (17,761). The number of training features from ResNet-18, ResNet-50, and ResNet-101 were 1000×74778, and testing features were 1000×31788, respectively. All the models have been trained with 32 batch sizes. In the process, fast Stochastic Gradient Descent (SGD) [124] with standard momentum parameters were applied.

Furthermore, we implemented a simple multiclass kNN (k-Nearest Neighbour) classifier stacked to the output features from the layers $1000d\ fc$ of ResNet-18, ResNet-50, and ResNet-101 to achieve the attention regulator behavior recognition. Our rationale lies in

the observation that the target dataset (WEDAR) is relatively small and different from the source dataset (ImageNet). The images in the WEDAR are also with high homogeneity. Thus, the fine-tuning of the WEDAR dataset will highly likely make it overfit. Therefore, we implemented the multiclass kNN classifier to verify it. Various k variables have been applied to ResNets for the comparative performance analysis.

Video-level attention regulator behavior recognition. Since attention regulator behaviors are the aggregation of instant actions over the temporal domain, frame-level recognition tends to lose rich, dynamic information. In that sense, we adopted a video-level framework compatible with the video inputs, having a "temporal" feature in its learning process. Comparative analyses have been conducted between frame-based and video-based models to achieve a better recognition result of attention regulator behaviors. Specifically, we implemented a hybrid architecture that consists of convolutions (for spatial information) and recurrent layers (for temporal information). We used a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) consisting of GRU layers [125], popularly known as a CNN-RNN [126, 127]. We chose the InceptionV3 architecture [128] as the CNN backbone, which has been pre-trained on ImageNet [123], benefiting from its lightweight structure, which is suitable for the temporal modeling. The output features of each frame have been fed into GRU with three layers (with GRU units as 16) and stacked by a fully connected layer as output. Besides, we noticed that imbalanced-data issues brought by a large number of neutral behaviors might affect the model's performance. Thus, we present two types of data sampling strategies: 1) Evenly sampling each category based on the class with the minimal number (balanced) and 2) using all the samples from each category (imbalanced).

Experimental results of the attention regulator behavior recognition. A comparative performance analysis has been conducted among models aimed at recognizing attention regulator behaviors (Table 3.3). Note that all models followed the same evaluation protocol mentioned above for fair comparisons. 1) Video-level models (CNN-RNN) have shown better performances than frame-level models (75.70% vs. 69.73% in subject-dependent settings and 68.43% vs. 25.90% in subject-independent settings) by large margins, with the more temporal information involved. It means capturing temporal dynamics (temporal reasoning) is important for behavior recognition. 2) The performances of the models vary significantly based on the evaluating protocol. ResNet-kNN architecture performed better than ResNet-finetuning architecture on the subject-dependent protocol. ResNet-kNN (-18, -50, -101) has achieved 69.98%, 69.95%, and 69.73% accuracy, respectively. However, when it comes to subject-independent protocol, ResNet-kNN models have a significant performance drop of more than 50% accuracy, while the performances of Resnet-finetuning models are relatively steady. This result indicates that the high performance of the ResNet-kNN model is benefited from the subject-dependent setting via overfitting to our WEDAR dataset. 3) The comparison between different ResNet variants has shown the best result in ResNet-18 with slight performance differences compared to other models with higher learnable parameters. Because WEDAR is a relatively small dataset, learning could have been converged early with smaller learnable layers. Our result emphasizes the importance of the compatibility of the sizes of the model and datasets [129, 130].

Table 3.4: Attention regulator behavior-based attention recognition results from various classifiers. The attention span is the instance duration before and after the distraction self-reports. We show the average and standard deviations over six leave-five-subject-out runs.

Methods	Attentional state recognition (%)			
	Attention span (2s)	Attention span (4s)	Attention span (8s)	Attention span (16s)
Random guess	0.50	0.50	0.50	0.50
kNN [131]	51.69 ± 5.62	61.86 ± 11.10	88.91 ± 7.98	80.02 ± 15.67
SVM [132]	$\mathbf{58.09 \pm 4.95}$	68.83 ± 7.67	89.31 ± 6.92	86.98 ± 7.43
AdaBoost [133]	57.84 ± 5.48	69.14 ± 7.51	88.12 ± 6.92	85.642 ± 6.83
MLP [134]	57.84 ± 5.48	$\mathbf{69.55 \pm 7.83}$	$\mathbf{89.41 \pm 6.91}$	$\mathbf{87.57 \pm 7.46}$

3.3.4 High-level attention analysis with attention regulator behaviors

This section introduces our attention analysis based on attention regulator behaviors. The task is recognizing the attentional states (i.e., attention or distraction) based on the attention regulator behaviors within a given small video instance.

Evaluation protocols. For the attentional state recognition, as the task is highly subject-dependent, we chose to use a leave-subjects-out protocol to verify the generalizability of the method. We obtained the ground truths of attention and distraction instances from participants' distraction self-reports. We took 8-second duration as an average attention span of human beings based on a literature study [135, 136]. Therefore, we set the last 8 seconds to the moment of distraction self-report as "distraction" while following 8 seconds from the moment of distraction self-report as "attention" state. We also took 16-second, 4-second duration and 2-second duration as comparisons. 383 distraction self-reports have been observed in the dataset, resulting in two sets of 383 × instances of "attention" and "distraction" states. We split the 30 subjects into six folds; each fold contains five subjects. To conduct the leave-subjects-out evaluation, we used all the attention instances from 25 subjects for the training and all the instances from the remaining five subjects for testing at each fold evaluation. Each instance belonged to a specific state (attention or distraction). We reported the average and standard deviations of the recognition accuracy in percentage. Note that we only focused on the recognition task of "recalled" and "reported" distractions. Thus, although "false-negative" errors of the self-reports (e.g., participants forgot to or ignored reporting the distraction) exist, they will not be included in this analysis.

Attentional state recognition. We provide six machine learning-based methods for attentional state recognition, using attention regulator behaviors as cues. We first encoded the distribution of attention regulator behaviors that happened within a given attention span as feature vectors with dimensions of $1 \times N$. N is the number of attention regulator behaviors, as six in practice. Since we used 30 fps for the annotation, which is redundant to count the attention regulator behaviors, we downsample the frame rate from 30 to 8. The resulting feature vectors were fed into the classifiers to predict the final binary attentional states (i.e., attention or distraction). We experimented with different classical machine learning classifiers combined with feature embedding: Bayesian network [92], Multi-layer Perceptron with Relu non-linearity (MLP) [134], k-nearest neighbors (kNN) [131], and Adaptive Boosting (AdaBoost) [133]. As can be seen from Table 3.4, the MLP classifier has achieved the best performance (69.55%, 89.41%, and 87.57%) in the attention span settings of 4s, 8s, and 16s while the SVM classifier has shown the best performance over the attention
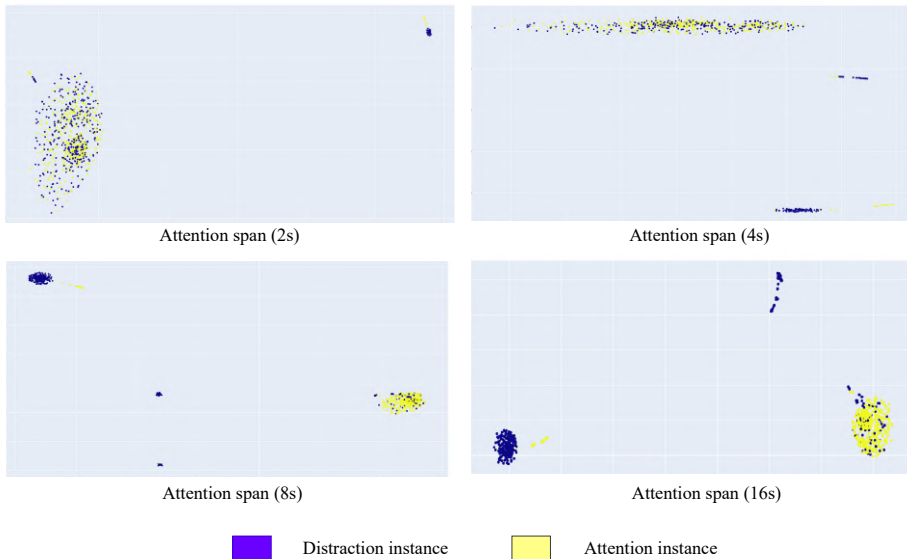
Figure 3.3: The t-SNE visualization of the features for attentional states. The feature embeddings are obtained based on the attention regulator behaviors happened during the given attention span. Each dot stands for attentional states.

span of 2s. We can also observe that a shorter attention span of an instance has brought a significant performance drop (87.57% to 57.84% from 16s to 2s) in the recognition. We assume it is because the shorter attention span implementation does not provide enough information on the attention regulator behaviors to build up the probabilistic distribution model for further inferences.

Visualization of the features for attentional state recognition. In this section, we visualized the feature embeddings constructed from the attention regulator behaviors, using the t-SNE technique [137]. As shown in Figure 3.3, features from a short attention span are not discriminative enough, while features from a longer attention span show much larger margins.

3.4 Discussion and limitations
3.4.1 Discussion
Distraction self-reports vs. attention regulator behaviors. Self-reported distractions during the e-reading practices can be regarded as ground truths of attentional states of participants to some extent, as they are the direct reflection of internal activities provided by participants. However, there are three major limitations of the distraction self-reports. 1) Self-reports are based on a dual-task condition. The participants might be distracted when keeping the reporting task in their minds, which could affect their attention level. 2) self-reported metrics are not always reliable as participants often forget to record their distractions. Thus, false-negative errors are evident in some case, which could be a severe issue when evaluating the performances of online detection algorithms. 3) Lastly, the

self-report is subjective, making it difficult for machines to learn the patterns. In contrast, attention regulators-based attentional state analysis has advantages as follows. 1) Those patterns are concrete and rather easy to observe in the images so that machines can easily learn. 2) Significant correlations found between distraction reports and attention regulator behaviors indicate that observable attention regulator behaviors as a good predictor of attention.

Further implementation in e-learning. Since our work aimed for a real-life application based on a webcam, we believe that the work can be extended to other reading-based e-learning scenarios with an investigation of attention regulator behaviors in the specific learning activity. By combining various feedback types with diverse instructional designs, platforms, and modalities from different feedback agents, more timely feedback provision can be achieved for learners and instructors.

Defining attention span. We defined the attention span by taking the duration before and after the distraction reports (e.g., 2s, 4s, 8s, and 16s). We found that the definition of the attention span can affect the performance of attention recognition by a large amount, as a longer period will contain more behavioral patterns for the recognition. Existing methods [19–21] mainly worked short-term or even frame-level attention recognition, while our findings can inspire the upcoming research to work on the direction of attention span by showing potential for holistic attention recognition in instances with a longer attention span.

Rich cues for attention analysis. In this work, we only presented some preliminary baselines using attention regulator behaviors and self-reports as cues and ground truths. However, rich cues provided in WEDAR, such as knowledge gains and reaction time, can offer more opportunities for a more holistic and long-term attention analysis.

3.4.2 Limitations

Differentiating spontaneous behavior vs. voluntary behavior. In this work, we focused on finding regulatory behavior that helps learners sustain their attention. We primarily focused on voluntary or semi-voluntary behaviors from learners with consciousness. However, it was often challenging to differentiate voluntary behaviors from spontaneous behaviors through human observation, which might have affected our labeling and prediction results.

Lack of categorical frameworks for attention regulator behaviors in e-learning. We strived to classify learner behaviors based on existing theoretical and empirical works. Though our work is a categorical expansion of [96], we still miss the dedicated framework that could be applied in the exploration of attention regulator behaviors in e-reading.

3.5 Conclusion and future work

In this work, we applied the categorical framework of [96] to an e-reading scenario and identified attention regulator behaviors, which was the first attempt. We collected a novel dataset from 30 higher education learners containing various cognitive, emotional, and behavioral cues. We annotated 931,340 frames of video data second-to-second into six categories. We used various classical models to recognize attention regulator behaviors as a baseline with the highest accuracy of 75.70% (subject-dependent) and 68.43% (subject-independent) with CNN-RNN. Attentional state recognition has been further conducted

by leveraging the attention regulator behaviors with a promising performance of 89.41% accuracy with a leave-five-subject-out protocol. Our webcam-based dataset and framework for the attention analysis make it feasible to comply with primary computing devices without sophisticated sensor implementation, allowing real-world implementation. We hope our work contributes to the field by providing insights into attention regulator behaviors in e-reading. The future research includes the system extension with the feedback implementation, which will function as an interactive feedback loop for attentive e-reading.

3

4

Investigating Behavioral Indicators for Predicting Learners' Higher and Lower-Level Thinking Skills: An Explainable AI Approach

The use of machine learning technology in learning analytics is becoming increasingly prevalent. However, the black-box nature of machine learning models presents challenges in interpreting and explaining the model's decision, which is critical for understanding the reasoning behind the results. Low interpretability limits the next-round intervention based on the analysis result, which is often a fundamental goal of learning analytics. In this study, we utilize the WEDAR dataset, which contains second-to-second video annotation with learners' behaviors that are directly and indirectly related to learners' self-reported distractions. The WEDAR has various data layers related to learners' attention, such as reaction time to the screen blur at randomized timings and learners' attention regulation behaviors during their studies. We further extracted features from learner behaviors, such as the dominance and expressiveness of attention regulation behaviors, quartiles of the reaction time, and reading speed, that we hypothesized to have correlations with learners' utilization of Higher-Order Thinking Skills (HOTS) and Lower-Order Thinking Skills (LOTS) in their digital reading. By developing decision trees to predict learners' cognitive processing and leveraging the feature importance of the models, we identified core indicators for predicting learners' HOTS and LOTS, supported by machine reasoning. The result indicates that the dominance of attention regulation behaviors is a reliable indicator of low use of LOTS, achieving 79.33% of prediction accuracy, while reading speed is a valuable indicator for predicting the overall usage of HOTS and LOTS, ranging from 60.66% to 78.66% accuracies. On the other hand, individual reaction time has only helped predict the usage of HOTS. Our study demonstrates how various combinations of behavior-based features can inform the development of explainable AI models for learners' cognitive processes that are both accurate and interpretable, providing valuable insights for education research and learning analytics. It supports future research for learners' cognitive processing in e-reading based on machine reasoning.

Digital technologies have transformed how we engage with educational materials [29]. With the increasing use of digital texts in formal and informal education [138], assessing and evaluating learners' cognitive processings in e-reading has become more critical [139]. It is a foundation for learning analytics and designing timely and effective interventions for learners who engage in digital reading [20]. However, sensor-based laboratory experiments often used in learning analytics challenge understanding learners' natural cognitive processing by changing the nature of real-life e-reading and the ecosystems with intrusive sensor implementations [140] and experimental design. In this sense, our work aims to understand learner behaviors in real life leveraged by AI technologies, with a multimodal WEDAR dataset [141] that premises a real-life understanding of e-reading with webcam-based data collection.

The existing approaches to e-reading assessment on cognitive dimension have predominantly relied on eye trackers [14, 142]. It is because indicators, such as pupil dilation [143], fixation, and saccades [144], work as objective and solid cues for understanding learners' cognitive states. At the same time, e-reading has a straightforward task with regular eye movement patterns (e.g., character-level fixations [145], scanning and skimming [146], Area of Interest (AOI) [147], number of blinks [148], re-reading [149]), making it a solid indicator of evaluating the cognitive demands and processing in e-reading. Various multimodal indicators, such as video data (e.g., valence, arousal [20]) and multiple layers of log data (e.g., mouse dynamics [140]), have been combined for a more multi-dimensional understanding of learner states and learning. However, feature-based analysis has suffered from the limitation coming from lacking standards of defining ideal learner features, which is often the case for e-reading analytics, too [20].

Based on multimodal learning analytics, learners' cognitive states, such as mind-wandering [14, 142], switches of internal thoughts [15], working memory [140], and affects (e.g., valence, arousal [20]) have been the target of the previous analysis. As a means to compensate for traditional feature-based analysis, self-reported data showing learners' subjective perceptions about their learning and experts' observations have been used as ground truths for machine reasoning [150]. Also, different physiological patterns found from learners with and without successful learning outcomes have been predicted using machine learning [151]. Models aimed at critical features are automatically learned in the model training processes in optimal ways. At the same time, human experts can consider the domain knowledge in the first round feature selection and the model interpretation process while still leveraging various AI technologies [152], which contribute to building a hybrid intelligence, bridging advantages of human and machine intelligence. However, due to the non-explainable nature of black-box in AI [153], there is a growing need for eXplainable AI (XAI) in education to understand the reasoning behind the model's decision [154].

In this regard, our work aims to fill the gap by developing an XAI model for e-reading assessment with behaviors, which identifies and analyzes the features to predict learners' Higher-Order Thinking Skills (HOTS) and Lower-Order Thinking Skills (LOTS). As represented in revised Bloom's Taxonomy from David R Krathwohl [79], HOTS and LOTS are involved in learning as cognitive objectives. Learners utilize LOTS in remembering,

---

This chapter is partly based on 📄 *Y. Lee., G. Migut., M. Specht. Unveiling Learners' Higher-Order and Lower-Order Thinking Skills through Attention Regulation Behaviors in E-reading: An Explainable AI Approach, submitted to a peer-reviewed journal.*

understanding, and analyzing knowledge, while HOT is used for more in-depth cognitive processing in applying, evaluating, and creating knowledge [155]. As HOTS and LOTS involve different cognitive processing, we hypothesized that such differences could be captured through specific observable behavioral cues (i.e., attention regulation behaviors) that are found to correlate with learners' distractions and attention management [156]. By adopting an XAI approach with behavior-based prediction, we try to find critical behavioral learning features, such as dominance and expressiveness of attention regulation behaviors, reaction time, and reading speed, that can potentially work as observable cues for predicting HOTS and LOTS during e-reading.
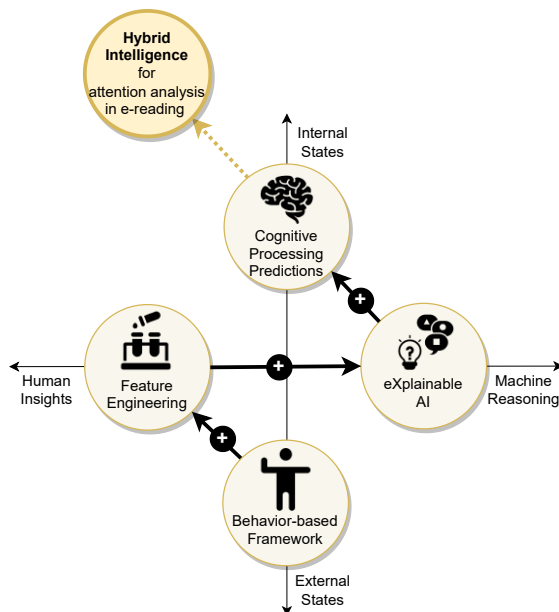
Figure 4.1: Our approach aimed at building a Hybrid Intelligence in e-reading by enhancing a behavior-based framework with added human expertise through feature engineering, machine reasoning via an explainable AI approach, and automatic cognitive processing prediction.

This work suggests a Hybrid Intelligence (HI) framework for human attention analysis in e-reading (see Figure 4.1). As suggested as an important challenge for the future AI application in education [157], we strived to balance the human insights from the experts and understand critical components for machine reasoning via the explainable AI approach. Such interpretability can greatly foster an integrated understanding of human attention, leveraged by machine and human intelligence. Our framework is 1) based on the behavior-based frameworks that are based on external states. 2) Using human insights, we select behavioral features that we hypothesize are correlated with human attention and can be used for machine reasoning. 3) We take the explainable AI approach that we can trace back which behavioral features contribute to machine reasoning to predict learners' cognitive processing. 4) Using the model, we evaluate the internal states of learners via external behavioral cues and make the prediction automatic at scale. All in all, our contributions

are listed below.

1) Simple understanding of learners' cognition through behavioral cues: Our study represents the first attempt to apply XAI to understand learners' cognitive processing in e-reading. By using our behavior-based HOTS and LOTS predictions and finding critical indicators in prediction, we can grasp complicated cognitive processing via interpretable combinations of learner behaviors. Traditionally, the learning analytics on learners' cognition has been done via dedicated biosensors, such as an eye tracker [151], which have challenged educational researchers with complicated hardware implementations and combinations of multimodal data streams with different granularity and thus not intuitively comprehensible. Also, such sensor layers have been criticized for bringing intrusiveness to the learning activity and hindering learning analytics in real-life e-reading. However, our webcam-based behavior analysis with XAI reveals the relationship between semantically understandable behavioral cues and learners' hidden cognitive processes in learning without obstructiveness.

2) Machine's behavior-based decision-making with interpretability and scalability: Though Human educators have invaluable expertise with domain knowledge and the ability to empathize with learners based on contextual understanding [158], the physical environment of e-learning brings constraints to them. E-reading environments allow learners and educators to communicate only through the interface, limiting common situational awareness of the educational context. Also, human educators have limitations in that they can only recognize a problem at a time with somewhat arbitrary criteria [152], which challenges the subsequent feedback with consistency. Also, different experiences and perceptions of human educators may lead to inconsistent feedback provisions. In this regard, our machine-based decision-making shows more straightforward reasoning based on behaviors that can also be semantically understandable to humans but with scalability.

3) Future extension of the framework to various e-learning scenarios and feedback agents: In this work, using the XAI approach, we strived for a semantic understanding of the influential features based on machine reasoning. Understanding prediction mechanisms related to different levels of HOTS and LOTS of specific individuals provides valuable insights to instructional designers for more concrete and adaptive intervention plans [159]. For instance, the framework can further be extended to various e-learning scenarios based on digital reading. Attention regulation behaviors in a specific e-learning scenario can also be expanded. Furthermore, our framework can connect with diverse instructional design strategies and different interfaces (e.g., conversational agents), closing the feedback loop with learning's behavioral, cognitive, and affective enhancements.

Below, we articulated three research questions that we focused on in our study:

- RQ1. What are critical behavioral indicators for predicting learners' HOTS and LOTS in e-reading for machine reasoning?

- RQ2. How can learners be grouped based on their different usages of HOTS and LOTS in e-reading?

- RQ3. How can an automatic evaluation of the learners' HOTS and LOTS in e-reading be achieved?

4.1 Related work

This section explores various theoretical frameworks that construct our overall framework. As illustrated in Figure 4.2, we utilize the framework of learners' attention regulation behaviors in e-reading [156]. In the process, various behavioral indicators known to be directly and indirectly correlated to learners' attention are used for the model training. We leveraged the Explainable Artificial Intelligence (XAI) approach in our work to understand learners' utilization of the HOTS and LOTS in their learning on various levels.
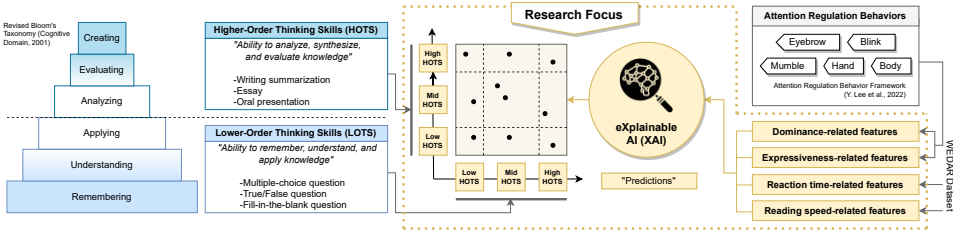
Figure 4.2: Our framework is based on the revised Bloom's taxonomy [79], which has HOTS and LOTS as components of learners' cognitive processing. Using decision trees, we strived to predict learners' HOTS and LOTS based on attention regulation behaviors [156]. We tried to understand the reasoning behind the model's decision to find the critical behavioral components for predicting different levels and combinations of learners' cognitive processing.

4.1.1 Current XAI approaches in education

XAI is an emerging topic with various applicability in areas where the reasoning behind decision-making is especially critical (e.g., healthcare, law, autonomous driving [154]). In education, Learning Analytics (LA) and Educational Data Mining (EDM) are two areas where AI-driven approaches commonly take place for various stakeholders (e.g., teachers, tutors, students, and managers [159]) throughout learning phases in collecting, processing, exploiting, and reporting the learning data [154]. However, while machine learning models can successfully perform tasks such as classification, regression, clustering, transferring, and optimization, researchers often cannot elaborate the reasoning behind the models' decisions due to the black-box nature of AI [160]. Therefore, understanding the specific task in feature engineering and result interpretation from human experts has been considered critical [159]. However, the limited explainability of models still raises ethical and trustworthiness issues for educational applications for lacking transparency, trust, and fairness in decision-making [153]. Therefore, various XAI approaches have been taken in educational research by revealing the feature dominance, correlation among features used for the training, the reasoning behind predictions [153], and sources of noise in the decision-making [159].

[159] has suggested a framework of XAI in Education (XAI-ED) that aligns the needs of stakeholders, interfaces, and AI models. Various XAI approaches in education, such as the Generalized Additive Model (GAM) with a linear relationship between independent and dependent variables [161], the decision tree model with a hierarchical structure, a rule-based model with conditional statements, the clustering method with specific data patterns, and natural language processing with data cross-validations among learning data

have been introduced. Through a survey, [153] introduced various XAI approaches and made baseline comparisons of different state-of-the-art methods with multiple modalities and features. XAI application has been divided into transparent methods (e.g., bayesian model, decision trees, linear regression, fuzzy inference systems) and post-hoc methods (e.g., LIME, perturbation, LRP, SHAP). While the former approaches are commonly used when simple relationships among features take place, the latter methods were found to be generally applied when high data complexity exists [153]. [154] has suggested the GUI web-based ExpliClas service, which provides text descriptions and a dashboard with data visualizations regarding the feature use and recommendations. [160] has implemented a decision tree to find critical features among learners' listening, watching, making, and speaking behaviors for predicting collaborative problem-solving competencies.

All in all, the general focus has been finding XAI implementation opportunities in education with model comparisons and platform suggestions. However, deep diving into features for predicting learners' cognitive processing has yet to be found. Especially according to our best knowledge, XAI in the behavior analysis for understanding the cognitive processing of learners nor XAI for digital reading applications has been attempted yet. It is essential for the rapidly growing necessity of learning analytics and feedback loop design for real-life digital reading, which we target to foster using hybrid human and machine intelligence.

4.1.2 Assessing learners' HOTS and LOTS in e-reading

Understanding the way that learners utilize types of thinking skills in learning is essential since the thinking skills affect the ability [162], speed, and effectiveness [163] of learning [164]. This work uses revised Bloom's taxonomy [79], which differentiated learners' six levels of thinking skills as remembering, understanding, applying, analyzing, evaluating, and creating. This work uses LOTS (i.e., remembering, understanding, applying) and HOTS (i.e., analyzing, evaluating, creating) for learning analytics and machine learning model training.

LOTS facilitates lower-level cognitive processing, such as comprehension and information recall. LOTS encompass fundamental cognitive abilities such as remembering, understanding, and applying knowledge during learning. These skills are closely associated with acquiring concepts, facts, and procedures [165]. Therefore, LOTS often utilize short-term memory, which relies on temporary memory retention [166] and has often been evaluated through multiple-choice, true or false [167], and fill-in-the-blank questions [168] for the reading assessment.

Conversely, HOTS supports more complex cognitive processing, including analysis, evaluation, and synthesis [169], which involves more proactive judgment and assessment from learners. These skills include creative and critical thinking, analysis, and knowledge synthesis [170]. Such cognitive processing enables learners to acquire and retain knowledge in the long term [171]. Unlike LOTS, which relies on memorization and superficial judgments, knowledge from HOTS is highly transferable to the new contexts [171] and involves interconnecting prior knowledge with further information and creating own judgments [155]. For evaluating HOTS for reading tasks, previous works utilized posthoc summarization [170], and essay writing [172].

Understanding learners' cognitive processing can be complicated since some learners may excel in LOTS but struggle with HOTS, while others may demonstrate the opposite pattern

[155, 167]. Recognizing the different usages of HOTS and LOTS can inform researchers and instructional designers of different cognitive processing, learning needs, and subsequent educational interventions, enabling personalized solutions for each group of learners. For further details for assessing and segmenting learners based on HOTS and LOTS, please refer to sections 4.2.2 and 4.2.2.

4.1.3 Behavior-based framework for evaluating learners' HOTS and LOTS in e-reading

In this work, we extract behavioral features from the WEDAR dataset that we hypothesized to correlate with learners' HOTS and LOTS, leveraged by human expertise. We mainly focus on learners' behaviors, such as 1) dominance and 2) expressiveness of attention regulation behaviors, 3) reaction time to the secondary blur stimuli, and 4) reading speed. Those indicators were hypothesized to, directly and indirectly, reflect learners' affective and cognitive states in e-reading practices. We use features in combinations to see the best prediction results and analyze critical elements for the model's decision. By doing so, we attempt to understand the most influential behavioral features in machine reasoning for predicting cognitive processing so our work can further assist the intervention design for various learners with different cognitive processing patterns. Below, we articulate how our behavioral indicators have been understood in previous research.

Attention regulation behaviors

In the previous work, several behavioral cues have been defined as attention regulation behaviors that indicate learners' own perceived attention loss during their e-reading [156]. In the framework, various movements in eyebrows (e.g., raising, bringing together), blinks (e.g., blink flurries, voluntary prolonged blink), mumble (e.g., mumble reading), hand (e.g., touching body and or face), and body (e.g., adjusting position and or angle of torso, arm) have been considered as voluntary and spontaneous actions learners engage in to regain attention during e-reading. In the previous study, real-time attention regulation behaviors were found to correlate with learners' self-reported distractions, leading to the development of a video-based distraction recognition based on the WEDAR dataset [156]. In this work, post-hoc features from the WEDAR are processed to capture behaviors directly or indirectly related to attention and further affect learners' cognitive processing. We hypothesized that such behaviors and the usage of HOTS and LOTS have a specific correlation that can also work as a foundation for automatic recognition of cognitive processing.

Dominance and expressiveness of attention regulation behaviors

Contextual features [95] (e.g., individual and cultural factors) are known to highly influence human behaviors' frequency and expressiveness [173]. Such individual differences in behaviors often pose challenges for the generalized behavior-based learning analytics [174]. In this study, we aimed to investigate the relationship between dominance and expressiveness of attention regulation behaviors and learners' cognitive processing, specifically HOTS and LOTS. By exploring the role of dominance and expressiveness in learners' cognitive processing, we can gain deeper insights into the role of attention regulation behaviors in e-reading. The analysis can work as a foundation for personalized learning and optimize instructional design strategies accordingly.

Reaction time to the screen blur at randomized timing during e-reading

Reaction time has long been recognized as a reliable indicator of learners' arousal and attention during task performances [15]. Fast reaction time has commonly been associated with efficient attentional control and vigilance [175], indicating the ability to maintain focus and allocate cognitive resources effectively. Conversely, slow reaction time has been suggested as disengagement from the task and challenges sustaining an optimal attentional state amidst distractions [15]. The influence of affective states, including arousal and engagement, is known to shape individuals' reaction time [176], representing its potential correlations to cognitive processing in learning. Given the suggested insights, we hypothesized that reaction time to the screen blur could work as a feature that robustly predicts the utilization of learners' HOTS and LOTS during their e-reading.

Reading speed

Reading speed is known to provide valuable insights into learners' cognitive load and information processing capabilities [177]. Though faster reading does not guarantee better learning, it is found to be associated with more rapid information gain and reduced cognitive load compared to that from slower readers [178]. Moreover, fast readers on screen-based reading are known to experience fewer distractions [179], which supports our attempt to predict cognitive processes based on various reading behaviors, including attention regulation behaviors. Regarding the reading speed, we hypothesized that higher attention and faster reading speed would contribute to enhanced HOTS and LOTS.

4.2 Methods

In this chapter, we introduce how we preprocessed the multimodal WEDAR dataset and trained them to predict the HOTS and LOTS.

4.2.1 Multimodal WEDAR dataset

Figure 4.3: Our work utilized the knowledge gained for assessing HOTS & LOTS, distraction self-reports, reaction time to screen blur stimuli, and attention regulation behaviors of learners from a multimodal WEDAR dataset.

In this study, we utilized the public WEDAR dataset, which includes assessments for representing LOTS from the multiple choice questions, HOTS from the text summarization

task, moment-to-moment self-reported distractions, learners' reaction time to the random-ized screen blur, and attention regulation behaviors annotated every second, which has approximately 8.7 hours long. The dataset was collected from 30 higher education learners during screen-based e-reading. Please note that this study only used post-hoc features since HOTS and LOTS were not collected in real-time; thus, predicting the post-hoc targets (i.e., HOTS and LOTS) with real-time behavior features can be misleading.

Table 4.1: Pe-processed multimodal WEDAR dataset, which has been used for the XAI model training.

Feature categories	#	Feature names	Feature description	Categorical / Nominal
Dominance-related	F1	behavior_eyebrow	number of eyebrow behaviors/total number of attention regulation behaviors	continuous (0-1)
(Attention regulation behavior)	F2	behavior_blink	number of blink behaviors/total number of attention regulation behaviors	continuous (0-1)
	F3	behavior_mumble	number of mumble behaviors/total number of attention regulation behaviors	continuous (0-1)
	F4	behavior_hand	number of hand behaviors/total number of attention regulation behaviors	continuous (0-10
	F5	behavior_body	number of body behaviors/total number of attention regulation behaviors	continuous (0-1)
	F6	first_behavior (one-hot encoded)	occurrences of having the most dominant attention regulation behaviors	5-classes (0, 1)
	F7	second_behavior (one-hot encoded)	occurrences of having the second dominant attention regulation behaviors	5-classes (0, 1)
	F8	third_behavior (one-hot encoded)	occurrences of having the third dominant attention regulation behaviors	5-classes (0, 1)
Expressiveness-related	F9	expressiveness	number of attention regulation behaviors/duration of the video	continuous (0-1)
(Attention regulation behavior)	F10	exp_level (one-hot encoded)	low (Q1), mid (Q2), high (Q3) expressiveness levels of each participant	3-classes (0, 1)
Reaction time-related	F11	indiv_reaction_average	reaction time average of each participant	continuous values
	F12	reaction_time (one-hot encoded)	fast (Q1), mid (Q2), slow (Q3) reaction time levels of each participant	3-classes
Reading speed-related	F13	individ_reading_speed	reading speed average of each participant (word/duration of reading)	continuous values
	F14	reading_speed (one-hot encoded)	fast (Q1), mid (Q2), slow (Q3) reading speed levels of each participant	3-classes (0, 1)
HOTS & LOTS	F15	LOTS	post-test score-pre-test score (multiple choice, full score:10)	continuous (0-10)
	F16	HOTS	BERTScore calculated based on written summarizations	continuous (0-1)
	F17	LOTS_level (one-hot encoded)	low (Q1), mid (Q2), high (Q3) LOTS of each participant	3-classes (0, 1)
	F18	HOTS_level (one-hot encoded)	low (Q1), mid (Q2), high (Q3) HOTS of each participant	3-classes (0, 1)
	F19	thinking_skills_clusters	3 clusters derived from K-means with LOTS and HOTS as feature vectors	3-classes (0, 1)

4.2.2 Feature Engineering of the WEDAR for the model training

As we aimed at the prediction of combinations of HOTS and LOTS, we identified four behavioral categories of features from the WEDAR: dominance-related, expressiveness-related, reaction time-related, and reading speed-related features as depicted in Table 4.1. Figure 4.3 provides an overview of the features from the WEDAR used for our study.
1) *Dominance-related features* are extracted based on the frequency of a specific attention regulation behavior that occurred, compared to the whole attention regulation behaviors that occurred. Also, the feature category includes the number of each attention regulation behavior found as each individual's first, second, and third frequently used attention regulation behaviors. 2) *Expressiveness-related features* indicate the number of attention regulation behaviors that occurred compared to the duration of reading that has taken place. The feature category also includes data on whether the individual belongs to a group with high, mid, and low behavioral expressiveness. 3) *Reaction time-related features* concern how long it took for individuals to react to the screen blur stimuli that were given at random timings. The feature category also includes information about participants considered fast, mid, or slow learners in terms of reaction time. 4) *Reading speed-related features* include the cues concerning the individual reading speed and where each learner belongs to groups of fast, mid, or slow readers. Lastly, 5) *HOTS & LOTS feature categories* include various features that were used as the targets of the predictions to understand behavioral features affecting different combinations of cognitive processing of individuals. LOTS was derived from the multiple choice question scores, while the HOTS was gauged by the BERT scores evaluated on learners' summarization. The features LOTS_level *(F17)* and HOTS_level *(F18)* represent each individual as high, mid, and low performers compared to all learners for their multiple choice and summarization, respectively. The feature thinking_skills_clusters *(F19)* are derived by the k-means clustering performed on individuals' HOTS and LOTS when k

has been decided as three in our case from the elbow method applied to the preprocessed WEDAR dataset. Please see the following section for the details of calculating the HOTS, LOTS, and groups with different levels of HOTS and LOTS combinations.

Pre-post multiple choice questions for evaluating LOTS
In the WEDAR, 10 multiple-choice questions related to the reading materials were given before and after the reading. Such pre-post questionnaires are often used for evaluating LOTS, focusing on the factual cognitive processing of learners. We calculated the LOTS by subtracting the pre-test score from the post-test score, making the final LOTS range from a scale of 0 to 10.

$$Score_{LOTS} = \sum_{i=1}^{N post} S_i^{post} - \sum_{i=1}^{N pre} S_i^{pre}, \tag{4.1}$$

where S_i^{post} is the post-test score (0 or 1) for question i, while S_i^{pre} is the pre-test score (0 or 1). Note that the pre-test and post-test questionnaire content were the same, making the LOTS range from 0 to 10.

BERT applied to the text summarization for evaluating HOTS
Evaluating text summarization by human evaluators can be subjective; thus, we utilized the automatic evaluation technique. We employed the Bidirectional Encoder Representations from Transformers (BERT), natural language processing model [180], to assess the HOTS from learners' summarization. BERT is a widely used language model because it can handle various language tasks under the consideration of contexts. It is especially relevant to our aim of understanding the similarity of learners' summarization and the original text in understanding learners' ability to reconstruct the contents that they read. Based on BERT, we evaluated participants' summaries, resulting in precision, recall, and F-1 scores ranging from 0 to 1. We used the entire reading content as a ground truth of the BERT model and each learner's overall summarization as inputs for evaluation. Recall, precision, and F-1 scores have been evaluated as R_BERT, P_BERT, and F_BERT, respectively, as below:

$$R_{BERT} = \frac{1}{|X|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \boldsymbol{x}_i^\top \hat{\boldsymbol{x}}_j, \tag{4.2}$$

$$P_{BERT} = \frac{1}{|\hat{X}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \boldsymbol{x}_i^\top \boldsymbol{x}_j, \tag{4.3}$$

$$F_{BERT} = 2 \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}, \tag{4.4}$$

where the ground truth that we extracted from the reading content has been x, while the summarizations from the participants have been $\hat{x}$. While X is the set of vectors representing the tokens from ground truth, $\hat{X}$ is the set of vectors representing the tokens from the summarizations given by the participants. x_i and $\hat{x}_j$ represent a vector in the X and $\hat{X}$, respectively. $\boldsymbol{x}_i^\top \hat{\boldsymbol{x}}_j$ and $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ are dot products between vectors from x_i to $\hat{x}_j$ and from x_i to x_j, respectively, calculating the average of the maximum similarity between each token from the ground truth and given answers from participants. Note that we used the F_{BERT} as the HOTS of each learner since the F-1 score considers both precision and recall,

which provides balanced perspectives of cognitive processing with higher-order thinking skills in reading.

Learner segmentation based on combinations of HOTS and LOTS

To explore different levels and combinations of learners' HOTS and LOTS, we first defined the target combinations of HOTS and LOTS: 1) with k-means clustering method for automatic clustering and 2) quartile analysis to define thresholds for high (1st quartile: Q1), mid (2nd quartile: Q2), and low (3rd quartile: Q3) ranges of HOTS and LOTS. These categorizations provide insight into how learners can be divided based on the use of HOTS and LOTS.
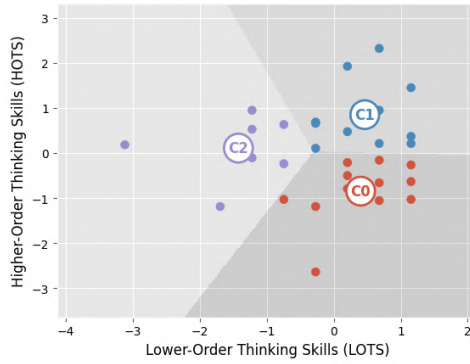
Figure 4.4: Learner segmentation based on k-means clustering.
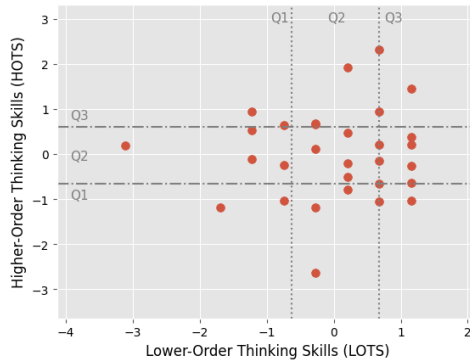
Figure 4.5: Learner segmentation based on quartile analysis.

Figure 4.6: K-means clustering and quartile analysis have been used for segmenting learners based on their HOTS and LOTS.

Automatic unsupervised clustering: k-means clustering

As can be seen from Figure 4.6, we performed k-means clustering [181] using HOTS and LOTS as feature vectors to segment the learners automatically. Based on the elbow method, we determined $k = 3$ and obtained three clusters. The clustering results helped to define one

group of learners (C2) with a relatively low LOTS range and two groups (C1 and C0) with a comparatively higher LOTS range. One of the groups with high LOTS (C1) demonstrated a higher HOTS, while the other (C0) exhibited a lower HOTS. The sample consisted of 12 learners in C0, 11 in C1, and 7 in C2, respectively. Note that we standardized HOTS and LOTS by mean-max scaling for both segmentation, subtracting and scaling the mean to unit variance [182] to ensure a fair comparison of HOTS and LOTS with different data ranges.

Quartile analysis: defining the high, mid, and low ranges of HOTS and LOTS
We also conducted quartiles analysis [183] to define thresholds for high, mid, and low ranges of HOTS and LOTS. For both HOTS and LOTS, learners in the top 25% (Q1) were considered high, the middle 25% to 75% (Q2) were considered mid, and those in the 75% (Q3) were considered learners with low HOTS and LOTS. This resulted in 9 (3 HOTS * 3 LOTS) thinking skills and combinations from learners with high, mid, and low HOTS and high, mid, and low LOTS, respectively.

To understand the relationship of the groups of learners derived from the first (i.e., k-means) and the second (i.e., quartile) segmentation methods, we analyzed the groups of learners assigned based on different segmentation criteria as shown in Figure 4.7. As represented in the figure and the literature study, HOTS and LOTS do not necessarily correlate, and the first and second methods do not result in consistent learner segmentation. It supports our attempt to predict different combinations of HOTS and LOTS based on individual behavioral features and understand the critical components in each decision: it can further help identify which cognitive processes can be used as targets of associated behavior analysis and assist instructional design for future application.
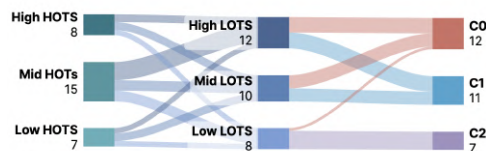
Figure 4.7: The usage of high, mid, and low ranges of HOTS & LOTS and their relations to clusters derived from k-means clustering.

4.3 Results
4.3.1 Model training protocols
Based on the decision tree, our model has been trained for accurate and explainable model development. The decision tree also has the advantage of automatically excluding irrelevant features and including only influential features by calculating the Gini impurity in its training process. The prediction target has been: 1) multi-class prediction of automatically-generated clusters derived clusters from k-means, and 2) binary prediction of high, mid, and low usage of HOTS and LOTS, based on the thresholds derived from the quartile analysis. It is important to point out that we used different target levels in prediction (i.e., multi-class, binary) because these approaches are best suited to the distinct characteristics of each target.

The clusters derived from the k-means method can be dynamic with added data points. Therefore, it is more logical to understand features that make such clusters distinctive than to comprehend the clusters themselves, which makes multi-level classification a more valid approach.

However, the quartiles represent fixed, interpretable, and distinctive data segments into low, mid, and high. This stable structure makes the binary classification approach more valuable, enabling a clear and straightforward analysis of whether specific traits contribute to the predictions of established thresholds. Therefore, Our methodological choices are tailored to each data type's unique properties and aimed at harnessing each classification approach's strengths.

Due to the limited sample size, we employed the SMOTE oversampling method to generate extra samples and match the number of samples to the most extensive label. We followed the traditional sampling method, dividing the training and testing sets into 80% and 20%, respectively. We used 5-fold cross-validation to train and test the data, which we subsequently averaged to compensate for the limited sample size.

From the features that we listed in Table 7.1, we used one-hot-encoded *F17* and *F18* as the training targets to achieve the quartile prediction. For cluster prediction, we set *F19* as the training target. We used dominance-related *(F1-F8)*, expressiveness-related *(F9-F10)*, reaction time-related *(F11-F12)*, and reading speed-related *(F13-F14)* features as predictors in combinations and independently. Please refer to Table 4.2 for the accuracy of predictions. We further conducted the feature importance analysis in the later section to understand critical behavior components that are used for the cognitive processing prediction by machines.

Table 4.2: Accuracies for predicting clusters and quartiles based on the decision tree.

Features	Prediction objectives						
	thinking_skills_clusters (k-means) (F19)	LOTS_level_high (F17)	LOTS_level_mid (F17)	LOTS_level_low (F17)	HOTS_level_high (F18)	HOTS_level_mid (F18)	HOTS_level_low (F18)
Random Guess	33.33	50.00	50.00	50.00	50.00	50.00	50.00
All *(F1-F14)*	**72.00**	67.33	49.33	67.33	**70.66**	66.66	**72.00**
Dominance-related *(F1-F8)*	62.00	64.66	63.99	**79.33**	52.66	58.00	**72.66**
Expressiveness-related *(F9-F10)*	36.66	71.33	52.66	71.33	38.66	49.33	40.00
Reaction time-related *(F11-F12)*	27.99	67.99	44.00	75.33	56.66	57.33	64.00
Reading speed-related *(F13-F14)*	42.66	**78.66**	65.33	78.66	60.66	**72.00**	70.66

[1] The best performances are **bolded** The second best performances are underlined.

4.3.2 Accuracy of the model prediction with different feature categories

We implemented a decision tree in our training to achieve a simple implementation and straightforward interpretation of the prediction. We set the maximum depth of the decision tree model into 10 to ensure simpler interpretability and prevent possible overfitting.

As shown in Table 4.2, using all features led to the best prediction performances for predicting the three thinking skill clusters derived from k-means, achieving an accuracy of 72.00%, highly surpassing the random guess that can be made by 33.33%. Considering the prediction results coming from different feature categories, the accuracy ranges from 27.99% with reaction-time-related features to 62.00% with dominance-related features; the overall prediction seems to depend heavily on dominance-related features. Also, dominance-related features achieved the highest accuracy for the quartiles of HOTS and LOTS, and dominance-related features achieved the highest accuracy, at 79.33% for predicting the

low LOTS and 72.66% for predicting the low HOTS. The result indicates that learners' dominance-related attention regulation behaviors work as robust cues of learners' low HOTS and LOTS.

Reading speed-related features also worked as robust predictors for predicting overall quartiles of HOTS and LOTS, with accuracy ranging from 60.66% to 78.66%. On the other hand, expressiveness-related features from attention regulation behaviors were only valuable for predicting low and high LOTS levels, with an accuracy of 71.33% each, while showing limitations in predicting HOTS. Similarly, reaction time-related features have shown sound prediction results for high (67.99%) and low (75.33%) LOTS while having comparatively poor prediction results for HOTS and learners with mid-range HOTS and LOTS. We assume that learners with low and high ranges of HOTS and LOTS show specific behavior patterns that can inform the machine reasoning, while learners with mid-range HOTS and LOTS did not show consistent behavior patterns, especially in expressiveness of attention regulation behaviors and reaction time to the screen blur.

4.3.3 Model interpretation for identifying significant predictors for cognitive processing

In this section, we examine plot trees and feature importance of models to identify the essential behavioral features for predicting learners' cognitive processing.
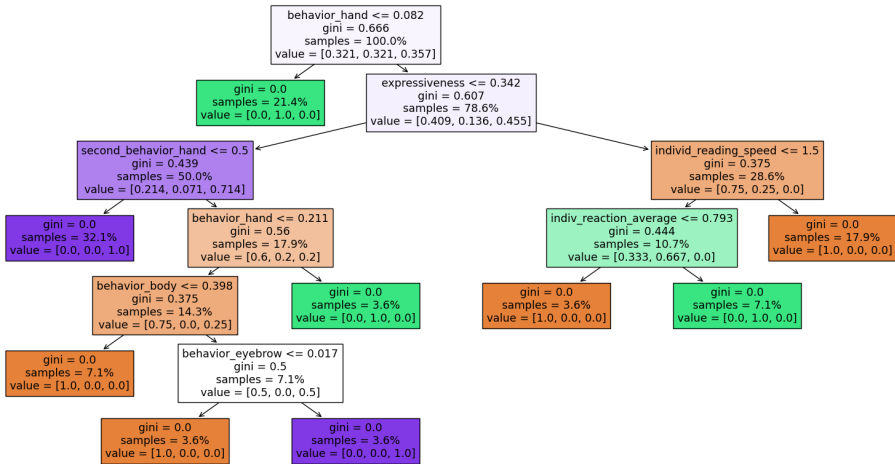
Figure 4.8: A plot tree to explain the model built upon the decision tree for predicting the thinking skill clusters.

Predicting three thinking skill clusters derived from k-means (F19)

Decision tree models provide great interpretability with the plot tree. Figure 4.8 illustrates the model's depth-by-depth decision-making process for predicting the three-level thinking skill clusters. The tree uses Gini impurity to understand the quality of the split of groups based on the condition, having 0 as the best purity with the best distinctions in the decision, while 1 indicates the impurity, which requires another round of decision-making. Values
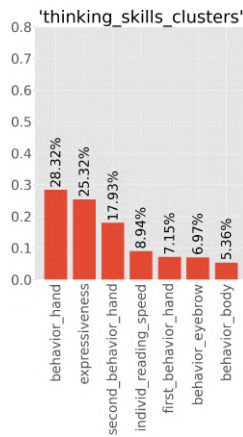
Figure 4.9: The feature importance for predicting thinking skill clusters has been investigated.

in the bracket indicate the possibility that each condition is classified as C0, C1, and C2, respectively.

Plot tree analysis for predicting thinking skill clusters derived from k-means (F19)
As can be seen from Figure 4.8, the first depth of the model considers the dominance of hand behavior *(F4)* as the most influential feature in the decision-making: It informs that if the feature marks less than 0.082, samples are classified as C1, making decisions for 21.4% of the samples. For the remaining 78.6% samples, the condition in the second depth, expressiveness of the attention regulation behavior *(F9)* of 0.342, has been used. On the right branch, 17.9% of the samples were classified as C0, having less than or the same as 1.5 as individual reading speed *(F13)* as the condition. The following condition of an individual reaction time average *(F11)* of less than or equal to 0.793 classified 3.6% of the samples as C0. Another 7.1% of the samples were classified as C1, with an individual reaction time average *(F11)* of more than 0.793. From the left branch, 3.6% of the samples were classified as C1, with a dominant hand behavior *(F4)* of more than 0.211. In the left branch, a second dominant hand behavior *(F7)* of 0.5 was the following condition, and 32.1% of the samples were classified as C2. The following condition of having dominance of body behavior *(F5)* of less than 0.398 classified 7.1% of the samples as C0. Finally, the last condition classified 3.6% of the samples as C0, having less than or equal to 0.017 as eye behavior dominance *(F1)*. In contrast, 3.6% were classified as C2, with more than 0.017 as eye behavior dominance. All in all, by conducting the feature analysis, we aimed to grasp how the model made the decision. Having those procedures aligned is especially insightful for education researchers and instructional designers, who work with the same sets of indicators and parameters. By having such standards, they can take a more systematic approach to learning analytics and subsequent intervention design, especially with learning behaviors.

Feature importance analysis for predicting thinking skill clusters derived from k-means (F19) In Figure 4.9, we listed features that were used for the model training and

ranked their feature importances derived from the tree model. The result shows that hand behaviors *(F4)* are usually the dominant feature for predicting the thinking skill clusters *(F19)*. Not only the dominance of the hand behavior (*F4*, 28.32%) but also hand behaviors as the first dominant (*F4*, 7.15%) and the second dominant (*F7*, 17.93%) behavioral features contributed to making decisions for thinking skill clusters *(F19)*. The expressiveness of the learner's behavior (*F9*, 25.32%) was the second most significant feature used to predict the thinking skill clusters *(F19)*. Additionally, individual reading speed (*F13*, 8.94%), the dominance of eyebrow movements (*F1*, 6.97%), and body movements (*F5*, 5.36%) among attention regulation behaviors were also used as indicators for predicting thinking skill clusters *(F19)*.

Predicting high, mid, low HOTS and LOTS (F17, F18)
Plot tree analysis for high, mid, low HOTS and LOTS (F17, F18)   In our comparative analysis, three decision tree models were developed to predict the high, mid, and low levels of LOTS (Figure 4.10, 4.11, 4.12) and HOTS (Figure 4.13, 4.14, 4.15), respectively.

The three trees for predicting LOTS share a consistent set of predictors, utilizing a variety of dominant hand *(F4)*, mumble *(F3)* behaviors, and behavioral expressiveness *(F9)* play a significant role in the prediction across all LOTS levels. Figure 4.10 for predicting the high level of LOTS initiates the split with dominant hand behaviors *(F4)*, suggesting its strong influence. Subsequent splits on dominant mumble behaviors and behavioral expressiveness *(F9)* illustrate a focus on nuanced behaviors to refine the prediction. The tree presents a balanced path with splits occurring at both the left and right nodes, indicating diverse sample distributions. In Figure 4.11, aiming at the mid-level LOTS prediction, individual reading speed *(F13)* extends to greater depths, signaling a more complex decision-making process with multiple behavioral and reaction time-related features such as dominant blink behaviors *(F2)* and reaction time quartiles *(F12)*, reflecting the intricate nature of predicting mid-range outcomes. Figure 4.12 predicts the low LOTS level, revealing a notable difference by starting with behavioral expressiveness *(F9)* as the primary split. It indicates that expressive behaviors determine lower learning outcomes in LOT evaluation. Unlike the previous models, Figure 4.12 simplifies the decision process with fewer splits, potentially revealing more apparent distinctions among lower LOTS levels based on expressiveness alone.

On the other hand, all models for predicting HOTS (Figure 4.13, 4.14, 4.15) have commonly used the dominance of hand *(F4)*, body *(F5)*, and mumble *(5)* as critical features for prediction, indicating universal applicability of such features to different HOTS levels. To predict high levels of HOTS 4.13, body behaviors as the most common attention regulation behaviors (F6) have been used as the root node, suggesting that initial body language plays a significant role in predicting higher cognitive skills. Figure 4.13 is less complex, with fewer splits, showing a more straightforward relationship between observable behaviors and high HOTS. Figure 4.14 focuses on describing the mid-level HOTS, starting with individual reaction average *(F11)*, indicating that mid-level HOTS may be more closely linked to the arousal levels of each individual. This model branches out into more levels of depth, requiring a deeper analysis to achieve accurate predictions. To describe low HOTS levels (Figure 4.15), dominant eyebrow behaviors *(F1)* was used as the root node, having an intermediate complexity between the high and mid-level models, showing a balance between behaviors and individual traits in determining lower HOTS.

Both LOTS and HOTS models utilized features related to the dominance of hand *(F4)*, body *(F5)*, and mumble *(F3)* behaviors. It represents the dominance of attention regulation behaviors as predictors of thinking skill levels *(F19)*. The LOTS models often use dominant hand behaviors *(F4)* as the root node, while the HOTS models vary, with the root node being body behaviors as the most dominant attention regulation behaviors *(F6)* for predicting high HOTS, individual reaction time average *(F11)* for mid HOTS, and dominant eyebrow behaviors *(F1)* for low HOTS. It suggests that different aspects of behavior and individual traits are considered for predicting different thinking skill levels *(F19)*. The HOTS models exhibited varying complexities, indicating that the prediction of HOTS levels may be more complex and require a deeper understanding of the interplay between different predictors. In contrast, the LOTS models appear more balanced, suggesting a more uniform distribution of features across varying levels of LOTS. The mid-HOTS model stands out, using an individual cognitive metric as the root, whereas the LOTS and other HOTS models tend to prioritize behavioral indicators. It implies that individual cognitive metrics are more predictive of mid-level HOTS, while observable behaviors are more indicative of the extreme levels of both LOTS and HOTS.
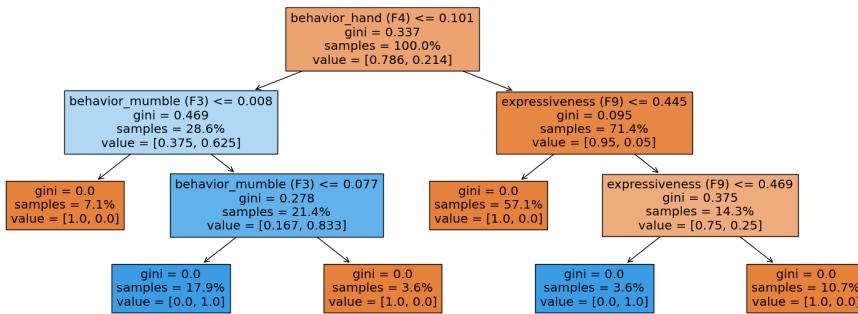
Figure 4.10: A plot tree to explain the model built upon the decision tree for predicting the high LOTS.
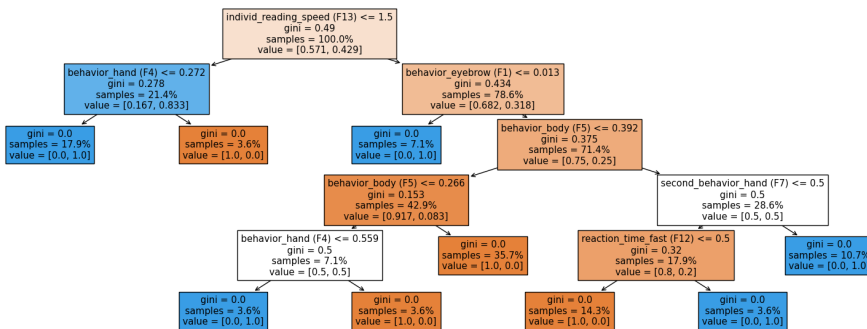
Figure 4.11: A plot tree to explain the model built upon the decision tree for predicting the mid LOTS.
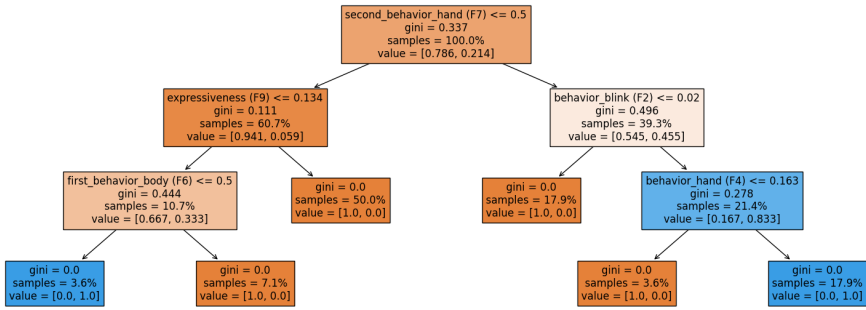
Figure 4.12: A plot tree to explain the model built upon the decision tree for predicting the low LOTS.
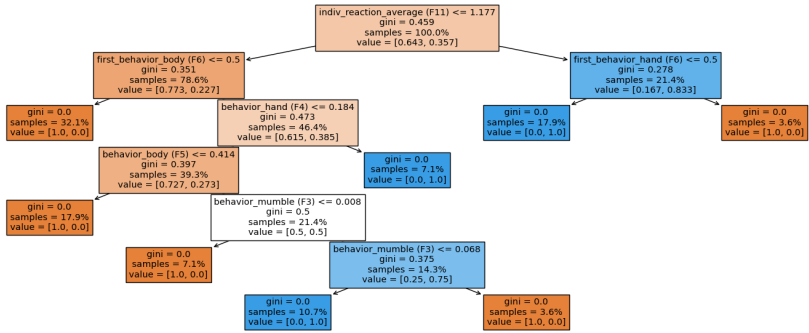
Figure 4.13: A plot tree to explain the model built upon the decision tree for predicting the high HOTS.
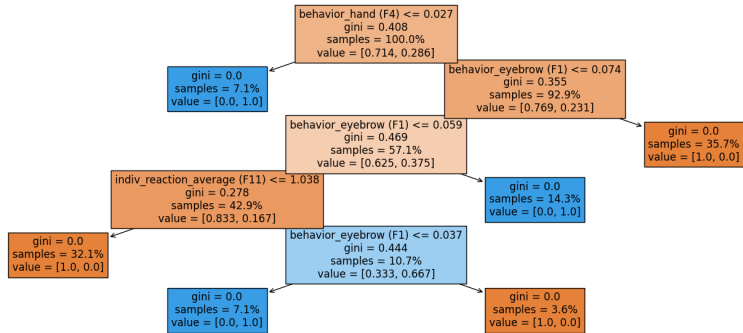
Figure 4.14: A plot tree to explain the model built upon the decision tree for predicting the mid HOTS.

Feature importance analysis for high, mid, low HOTS and LOTS (F17, F18)   Figure 4.16 shows that different behavioral features are essential in predicting LOTS and HOTS. The dominance of hand behaviors (*F4*, 40.08%), mumble reading (*F3*, 39.77%), and behavioral expressiveness *(F9)* have been identified as the most critical features for predicting high
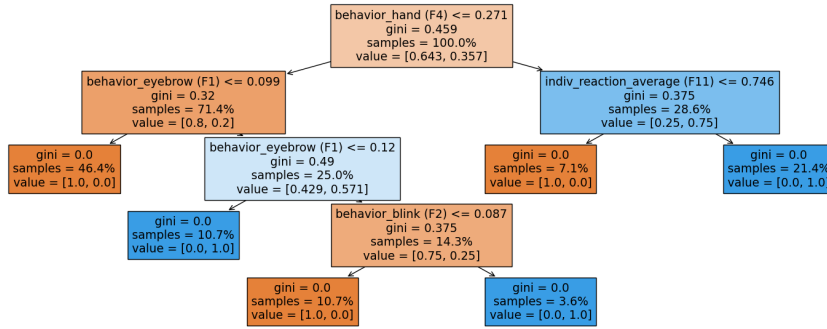
Figure 4.15: A plot tree to explain the model built upon the decision tree for predicting the low HOTS.
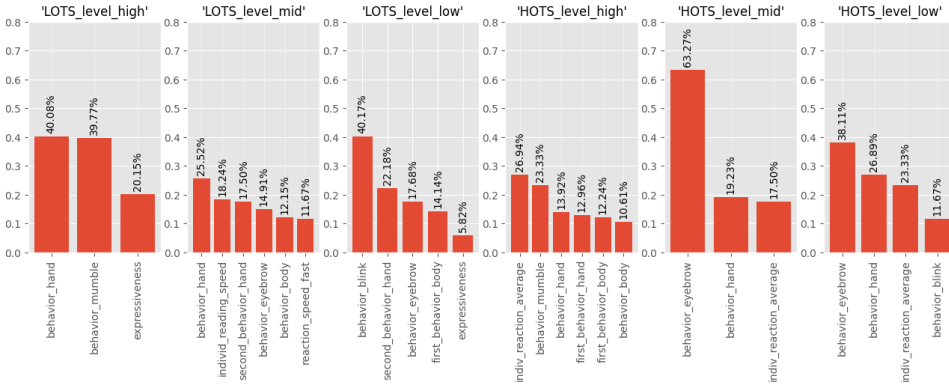
Figure 4.16: The feature importance for predicting high, mid, and low HOTS&LOTS have been investigated.

LOTS. For predicting mid LOTS, the dominance of movements in hand (*F4*, 25.52%), eyebrow (*F1*, 14.91%), and body (*F5*, 11.67%) have been used as indicators. Low LOTS have been predicted through features such as the dominance of blink behaviors (*F2*, 40.17%) and movements in eyebrows (*F1*, 14.91%). The dominance of body movements (*F5*, 12.15%) as the first dominant behavior and hand movements (*F4*, 22.18%) as the second dominant behavior *(F7)* have also been considered significant. Behavioral expressiveness (*F9*, 5.82%) has been identified as a predictor of low LOTS. In general, the dominance of diverse attention regulation behaviors *(F1-F5)* and behavioral expressiveness *(F9)* have been considered for predicting LOTS.

For high HOTS, individual reaction time average (*F11*, 26.94%) towards the screen blur stimuli has been identified as the most critical feature. The dominance of mumbling (*F3*, 23.33%), movements from the body (*F5*, 10.61%), and hand (*F4*, 13.92%) have been considered essential for predicting high HOTS. For learners with mid-HOTS, behavioral dominance of eyebrow movements (*F1*, 63.27%) and hand movements (*F4*, 19.23%), as well as individual reaction time average (*F11*, 17.50%), have been identified as critical features. For predicting low HOTS, the dominance of the eyebrow (*F1*, 38.11%), hand (*F4*, 26.89%), and blink (*F2*,

11.67%) have been used, along with individual reaction time average *(F11)* toward the secondary blur stimuli (*F12*, 23.33%).

All in all, the dominant movements from the eyes (i.e., eyebrows *(F1)*, blinks *(F2)*) were commonly used for predicting both low LOTS and HOTS. Expressiveness *(F9)* has been used for predicting LOTS. At the same time, learners' reaction time (*F11*, i.e., arousal) has been identified as a critical feature for predicting HOTS. Contrary to our hypothesis, reading speed *(F13)* was not considered more important than other behavioral feature categories for predicting overall LOTS and HOTS.

4.4 Discussion

Reaction time as a predictor of HOTS and expressiveness as a predictor of thinking skill clusters and low and high LOTS   The result indicates that reaction time *(F11)* has only been used for predicting HOTS, while it has not been considered for making judgments for LOTS and thinking skill clusters *(F19)*. It might indicate that more arousal, observed from fast reaction time, is likely related to learners' HOTS. On the other hand, the expressiveness *(F9)* is used for predicting high and low LOTS *(F17)* and thinking skill clusters *(F19)*. As higher expressiveness indicates more attention regulation behavior during their reading, more distractions likely led to low LOTS *(F17)*, while fewer attention regulation behaviors have been interpreted to high LOTS *(F17)*. All in all, we assume that learners' arousal has been targeted for predicting HOTS *(F18)*, while more self-aware distractions (i.e., attention regulation behaviors) have been used for predicting LOTS *(F17)*, which needs further validation.

Expansion of model is necessary with cognitive frameworks and more sample collection   Our work emphasizes the formation of an automated prediction system based on XAI that helps with learning analytics and feature interpretations in e-reading. However, due to the limited sample instances collected from 30 learners, more data inputs, and following distribution changes in targeted thinking skill clusters, HOTS and LOTS might change further feature interpretations and the feature importance analysis. This work is meaningful in paving the way for an automatic machine learning feature analysis for XAI, leveraged by hybrid human and machine intelligence. However, it can still be nurtured by expanding behavioral and cognitive frameworks and more sample collections for more generalized results.

Hierarchy of feature importances among behavior categories   This work has implemented four categories of features as predictors of learners' HOTS and LOTS in e-reading. Regarding accuracy, we found reading speed to be a reliable indicator of predicting HOTS and LOTS, with an accuracy range between 60.66% to 78.66%, sometimes working as a better indicator than attention regulation behavior-related features. However, when all features have been taken for the prediction, reading speed has only been used to predict the mid-range of LOTS, not for other targets with levels of HOTS and LOTS. In this sense, we need further investigation into how reading speed is located among other features in the model building with more extensive samples.

4.5 Conclusion

This study focused on developing behavior-based XAI models in e-reading to predict learners' cognitive processing based on learners' utilization of HOTS and LOTS. Using our multimodal WEDAR dataset, we extracted behavioral features related to learners' attention, including dominance and expressiveness of attention regulation behaviors, reaction time to secondary blur stimuli, and reading speed. We hypothesized that these features could serve as predictors of HOTS and LOTS. We adopted an unsupervised clustering method (i.e., k-means clustering) and statistical quartile analysis to define targeted learners' cognitive processing in various levels and combinations. To achieve better explainability, we employed decision tree models with maximum depths of 10 suitable for small datasets with fewer feature categories.

The prediction results for thinking skill clusters and each high-mid-low level of HOTS and LOTS demonstrate robust accuracies ranging from 65.33% to 78.66% across different behavioral features and their combinations. The feature importance analysis reveals that attention regulation behavior is consistently a strong predictor for all types of HOTS and LOTS. According to the following critical component analysis of training features, individual reading speed was found to be relevant only in predicting thinking skill clusters, while behavioral expressiveness played an essential role in predicting thinking skill clusters and LOTS. Individual reaction time to secondary stimuli was utilized only in predicting HOTS.

In conclusion, our study successfully developed XAI models for behavior-based prediction of learners' cognitive processing with HOTS and LOTS in e-reading, leveraged by the hybrid approach of combining human intelligence and machine reasoning. The findings highlight the significance of attention regulation behavior as a consistent predictor across different cognitive processing with levels of HOTS and LOTS. At the same time, we found that expressiveness exclusively predicted thinking skill clusters and high and low LOTS, which seems to be related to learners' self-aware distractions shown by behavioral cues. On the other hand, reaction time was used for predicting HOTS, which we found to be related to learners' arousal, which needs further validation. The results contribute to understanding behavioral factors to predict learners' HOTS and LOTS in e-reading and provide valuable insights for educators and instructional designers.

5

Data-Driven Persona Development and Automatic Recognition for Real-Time Applications: An Unsupervised Machine Learning Approach

Different individual features of the learner data often work as essential indicators of learning and intervention needs. This work exploits the personas in the design thinking process as the theoretical basis to analyze and cluster learners' learning behavior patterns as groups. To adapt to the learning practice, we develop data-driven personas by clustering learners' features based on factual learning outcomes (i.e., knowledge gain, perceived learning experience, perceived social presence) based on unsupervised learning, a more accessible and objective intervention design strategy for e-reading practices. Using the Chi-square test, we quantitatively evaluate different clusters driven by various unsupervised learning methods on the multimodal SKEP dataset. Furthermore, for a more practical real-life application, we achieved automatic persona prediction based on the attention regulation behaviors of learners. The subject-independent evaluation results indicate the best classification accuracy of 70% for the four-level classification task, differentiating three personas of learners with needs and another without feedback needs. It also shows that time-based sampling on both independent and cumulative learner behaviors works as robust predictors of learner personas, achieving a stable accuracy range of 65%-70% throughout the e-reading with the SVM classifier. Our work inspires the design of a real-time feedback loop for e-learning based on conversational agents.

Understanding users is an essential system design requirement for usability and better-perceived services [185]. It is especially well-emphasized for digital product (e.g., software, online courses, eBooks) design since poor user requirement engineering causes a perception gap between users and the practitioners, while users are often veiled with unknown varieties [186]. Likewise, understanding learners' traits and needs has been a critical challenge in e-learning intervention design. Especially, learning and learner necessities in e-learning tend to be more specific due to the physical absence of human educators and peers, while keeping close attention and engagement remains a challenge compared to the traditional on-site learning environment [156]. E-learning is becoming a mainstream education with recent social changes (e.g., COVID-19 [187]) with widespread e-learning platforms, digital devices, various forms of learning interventions, feedback agents, and modalities (e.g., social robot [158]). Those are designed for diverse learning objectives (e.g., formal and informal learning [188]) in e-learning and hybrid education settings [189]. However, e-reading system development approaches for better engagement seem scarce compared to the field's rapid growth and necessities [190].

As a user-centric decision-making tool [185, 186], the concept of "personas" has taken place in various domains, such as healthcare, knowledge management, social media, software development, and games [191] since its first appearance in 1999 [192]. Persona was devised as a practical and iterative [185] interaction design tool [192] in the design thinking process [193]. Persona has been further elaborated as hypothetical "archetypal" representatives [194] with specific needs, goals [195], attitudes, skills, roles, and expected behaviors [185, 196]. Those imaginary presences are believed to deliver certain behavioral traits, perceptions, and beliefs of specific segments of people in the real-world [191]. Persona is meaningful in providing a shared understanding of target users, their needs, and system usage [197]. Recent advancements in big data, data science algorithms, and data infrastructures have made data-driven persona development and analytics more accessible than ever [191], that has been traditionally done by several dozens of experts [198] for months and years [199]. Even though feedback personalization in education has become more accessible with more accurate predictions available through sensors, algorithms, and computing resources, according to our best knowledge, data-driven persona developments and the following learning analytics in education for feedback system development, especially in e-reading, have yet to be attempted.

In this regard, we develop data-driven personas using user modeling techniques based on unsupervised learning and its analysis [185, 194]. Instead of designing the feedback first and fitting learners with somewhat arbitrary criteria, we utilize the factual learning outcomes (i.e., knowledge gain, perceived learning experience, perceived social presence) collected from learners and use them as features for clustering learners, serving as the objective ground truths for analysis. Our data-driven persona approach is especially valuable for instructional designers and practitioners who lack standardized methods for analyzing learners as groups for further learning analytics and intervention design. Even with the same set of learner data and analytical objectives, it is nearly impossible to share the same criteria when developing a persona with somewhat manual and qualitative methods, with different perceptions and experiences of evaluators. Such deviations in decision-making inevitably lead to subjective and inconsistent learner clustering, which hinders timely and adequate intervention provision for learners.

Not merely working on the quantification and diversification of clusters, which has been a focus of early development of quantitative persona [191], this paper strives for deeper insights into learner analysis for e-reading intervention design by connecting the quantified persona model to statistical analysis. We explore utilizing data-driven learner persona to provide valuable insights into who learners are in terms of their categorical divisions, feature compositions, and their needs as a group in one grasp with statistical interpretations [200] and recognize them with classical machine learning classifiers.

From the perspectives of instructional designers, it is also more practical and feasible to understand the semantic and statistical meaning of the core features of groups and design interventions for them than making specific rules for individual features that deliver fragmentary and linear information. Feature-based learner divisions often end up deriving hypothetical learners with flat and stereotypical characteristics, which limits deeper insights about learners. To compensate for the limitation, we suggest the intervention design based on the data-driven persona using learners' factual learning outcomes as major dimensions of learner clustering.

Furthermore, we address a core issue of the utility of the above automatically generated persona categories for the following intervention: predicting the learners' persona categories for robust and timely learning interventions. To this end, we utilize human-labeled video samples from the SKEP dataset [201] to train machine-learning models to achieve the prediction of learners' persona categories based on their real-time and accumulated behaviors. The methods are validated via subject-independent protocols to ensure the generalizability of our method. Our automatic data-driven persona development framework and its prediction can assist in forming a feedback loop for better learning outcomes and experiences [195].

This work follows the procedure of 1) feature engineering on various types and levels of factual learning outcomes, 2) implementation of various unsupervised learning techniques and validation, 3) archetype extraction and data-driven persona development based on quartile analysis, and 4) learner persona prediction based on attention regulation behaviors (see Fig.5.1). We first utilize the multimodal SKEP dataset with the 25 multimodal features that include various matrices (e.g., pre-post test, Attrikdiff questionnaire, Social Presence questionnaire, and human annotation of six attention regulation behaviors for every second on approximately 40 hours of video data) to understand diverse perspectives of factual learning outcomes, collected from 60 higher education learners. It is a dataset that has been carefully designed and collected to understand learner behaviors and internal attributes in e-reading with emphatic and metacognitive feedback prompts from conversational agents. See [158] for the experimental details.

As suggested in the recent review of [191], we implement and compare various clustering methods on the dataset, such as k-means clustering, hierarchical clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and spectral clustering, that represent various modeling methods (i.e., centroid, hierarchy, density, graph) with various hyperparameters, which have further been cross-validated via Chi-square test. Subsequently, we conduct statistical analysis on each cluster to find distinctive and significant clusters features and draw our insights based on it [197]. Using classical machine learning models, such as AdaBoost, SVM, kNN, and Random Forest classifiers, we develop the behavior-based prediction model for personas on multi-levels as a part of the potential

1. Feature engineering of multimodal SKEP dataset	2. Learner clustering based on unsupervised learning	3. Archetype extraction and data-driven persona development	4. Persona prediction based on attention regulation behaviors
RQ1	**RQ2**	**RQ3**	**RQ4**
-Knowledge gain -Perceived learning experiences -Perceived social presence -Attention regulation behaviors	-K-means clustering -Hierarchical clustering -Density-based Spacial clustering -Spacial clustering	-Statistical distribution -Chi-square test -Quartile analysis -Archetype extraction	-Multiclass (six-level, four-level) & Binary classification (two-level) -Learning phase-based & Time based prediction

Figure 5.1: Our work covers multimodal data processing, learner categorization based on unsupervised learning methods, archetype extraction based on statistical analysis, and personas prediction based on learner behaviors.

feedback loop for e-reading. All in all, our research questions are as listed as followings.

- RQ1. How can learner features best reflect learners' performances, experiences, and perceptions of conversational agents' interventions in e-reading?

- RQ2. How can unsupervised learning methods be used for learner pattern clustering and validation?

- RQ3. How can we extract valuable archetypes of learners from different clusters and develop data-driven personas based on them?

- RQ4. Can we predict learner personas based on attention regulation behaviors?

To summarize, our contributions are listed below.
1) To our best knowledge, it is the first attempt to extend the data-driven persona development framework to e-reading with conversational agents. Personas provide learner clusters with a more concrete, multi-dimensional synthesis of learner features that represent learner categories differently from the cumbersome manual divisions of learners. Our feature engineering and the clustering result can provide the foundation for future data-driven persona-based learning analytics and intervention design for learners and instructors.
2) Despite its necessity, an extensive comparison among various clustering methods with learner data has yet to be attempted. We implement four unsupervised models with various modeling methods: k-means, hierarchical, DBSCAN, and spatial clustering. We conduct a Chi-square test to find the similarity among clusters derived by different modeling methods to validate clusters suggested by each other. It is a valuable attempt for future researchers' model implementation decisions for unsupervised learning-based clustering.
3) We explore the application of proposed data-driven personas: predicting the learners' persona categories for robust and timely learning interventions. We train machine-learning models to predict learners' persona categories based on their real-time and accumulated attention regulation behaviors. It will provide a foundation for a solid HRI feedback loop design in e-reading, promoting knowledge gain, perceived learning experiences, and perceived social presence of learners. It is beneficial for the following learning analytics and instructional design in e-learning for adaptive feedback implementation.

5.1 Related work
In recent years, more technology-enhanced learning and machine learning approaches have taken roles to reveal hidden patterns in learning and help with the intervention design

for the education administration and instructional design [202]. This section introduces previous approaches using unsupervised machine learning methods in diverse learning scenarios. The topic will be more specified with the review of data-driven persona development, which will be the focus of this work. At the same time, the section on learning analytics indicators on e-reading will help us derive important learner features and further analysis. Lastly, we develop behavior-based machine learning models to bridge learning analytics and data-driven persona prediction.

5.1.1 Unsupervised learning method for education

In this section, we focus on input features, objectives, and validation methods that have been applied to unsupervised methods in education. [75] focused on individual factors (e.g., gender, age, region, highest education, Index of Multiple Deprivation (IMD) bands, disability) and the data from the previous course (e.g., studied credit, number of clicks), to gauge the student involvement and their achievements in online-learning, using k-means clustering. [203] segmented the students' learning behaviors, utilizing data layers of 22 features (e.g., in information acquisition, solution construction, and solution assessment). It applied a t-test to represent the significance of particular features and a sparse k-means clustering for the feature selection and the final segmentation of learners, respectively. [204] has used k-means clustering with multimodal indicators, such as eye-tracking, physiological, and motion-sensing data, to automatically identify learners' productivity states (e.g., neutral, collaborative, non-collaborative) in collaborative learning. The model has been evaluated through correlation analysis between learner states, task performances, and learning gains. [74] has utilized student posts (i.e., textual dialogues) in MOOC for K-12 education to understand functional similarities of discourses (e.g., questioning, statements, reflections, scaffolding, references) via the k-means clustering, combined with bayesian information criterion. For validation, machine-generated clusters have been compared with human-coded clusters. [205] has divided learners based on their answers to system questions, comparing clusters from hierarchical (i.e., hierarchical clustering) and non-hierarchical (i.e., k-means clustering) clusters. For validation purposes, the within-group and between-group squared sum have been evaluated, indicating that the non-hierarchical method enables more detailed clustering results than the hierarchical method. [206] has used 12 engagement metrics (e.g., number of logins, number of forum reads, number of forum posts, quiz reviews, assignment lateness, assignment submission) to cluster higher education learners with k-means clustering method, aiming at personalized online education. Various values of k have been applied to draw multi-levels of learner engagement clusters. [207] has segmented higher education learners' using the k-means clustering method based on learners' academic performance (e.g., students' entry mode, residential category, scores of courses, age, post-UTME scores, GPA, gender, class of degree) and validated the clusters with a self-organizing map.

All in all, 1) from available implementation cases, it has been observed that the *k-means clustering method has been dominantly applied.* The only exception was [205], which has applied a hierarchical modeling method (i.e., hierarchical clustering) and a non-hierarchical modeling method (i.e., k-means clustering) to cross-validate each cluster. 2) *Large datasets from online education platforms have often been used* as input for modeling due to the easily accessible data. However, because such a dataset often only conveys rather superficial

quantitative log data (e.g., demographics of learners, number of clicks), result analysis has shown its limitation without in-depth insights on the specific topics. It differentiates the application of our SKEP dataset, which has been exclusively designed to understand learner behaviors (i.e., attention regulation behaviors), performances (i.e., knowledge gain), and internal states (i.e., perceived learning experiences, perceived social presence) with metacognitive feedback prompts from conversational agents in e-reading. 3) *There has yet to be a fixed validation method for modeling results* due to the nature of the unsupervised machine learning method, which relies on practitioners' further interpretations of results. Thus, various validation methods (e.g., within-group squared sum, t-test) have been applied based on researchers' needs on model implementations. 4) Though all works have represented learning analytics as outcomes to certain degrees, *there has yet to be an attempt to directly analyze the effect of feedback prompts of conversational agents and connect them with intervention loops.* It supports our attempt to develop an automatic data-driven persona and behavior-based prediction model that expands the feedback loop in e-reading with conversational agents.

5.1.2 Data-driven Persona Development approaches

Personas have been developed as representative figures that carry diverse user roles (e.g., users' characteristics, needs, and behaviors), profiles (e.g., demographic characteristics, motivation, goals, and personalities of users), segments (e.g., user relationship to the system, fundamental needs, characteristics of groups), and extreme characters (e.g., radical personalities of users), that delivers personal, technical, relationship, opinion information of users [185]. It started as a somewhat manual and qualitative analytics tool until recent years' proliferation of data, computing resources, and machine learning techniques [191]. The data-driven persona has been developed to compensate for the limitations of manual and qualitative persona: 1) high cost with long development duration with high monetary investments, 2) lack of objectivity and rigor due to the subjective criteria, 3) lack of scaling, which often leads to poor adaptation in big-scale data, 4) misrepresentation of clusters due to different insights and expertise of practitioners, and 5) expiration of validity with sample updates [191, 194]. The persona has evolved from the 1) qualitative method and 2) qualitative method with further quantitative validation in the early development. 3) Quantitative personas [208] have taken place with the implementation of unsupervised machine learning techniques, which is often further supported by the qualitative interpretations of practitioners on input indicators and clusters. Thus, the recent challenges of data-driven persona development have mostly come from data quality as the model input and interpretations of unsupervised models (e.g., data quality, data availability, method-specific weaknesses, human and machine biases [191]). The inputs of recent work of data-driven persona have ranged from accessible mouse-click log data to pricey data from surveys, self-reports, interviews, and user observations [196]. Regarding model implementation, a recent review has represented k-means clustering as the most used algorithm, followed by non-negative matrix factorization and hierarchical clustering. Various methods, such as latent semantic analysis [209], principal component analysis [191], and Cohen's Kappa [209], have been applied to best describe the distinctive cluster features and cluster validation using the clusters and new sample sets, respectively. Though no standardized methods exist for cluster validations, the most common data-driven persona validation methods

have been calculating the Euclidean distance between the different variables or testing the Chi-square. At the same time, subject experts validated the cluster by reviewing the clusters in a few pieces of literature [194].

To conclude, 1) the framework of *data-driven persona development has yet to be applied to the field of education*, which seems to be especially valuable for instructional design practitioners and researchers by representing learner groups with the synthesis of learner features. 2) *Comparative research among various unsupervised methods has been suggested but did not take place* in the field of data-driven persona development [191], which encourages our attempt to compare modeling methods (i.e., centroid, hierarchy, density, graph-based) and use each other for the cluster validation.

5.1.3 Indicators and measures of attentive e-reading

This section investigates various indicators to evaluate learners' e-reading with emphatic and metacognitive feedback prompts with conversational agents, especially based on Human-Robot Interaction (HRI). Analytics4Action Evaluation Framework (A4AEF) [210], an evidence-based learning analytics intervention evaluation protocol for online learning, has been applied, that has empathized teaching presence, cognitive presence, emotional presence, and social presence as core components of learning interventions. In the sub-section of *knowledge gain*, various feedback strategies from human educators and the existing systems are studied for insights into the feedback for better learning performances [211]. In the subsections of *perceived learning experiences* and *perceived social presence*, we investigate how multimodal feedback from systems is utilized and perceived by learners. In the subsection of *attention regulation behaviors in e-reading*, we investigate observable behavioral cues of learners that can be collectively used with other learning analytics measures to understand learners' attentional states during e-reading practices.

Knowledge gain

Knowledge gain is the primary goal of e-reading activities and vice versa; reading has been one of the most fundamental forms of knowledge gain in higher education [158]. In recent years, e-reading has become more commonplace with the rapid digitalization of education and the widely-used smart devices [156]. Reading comprehension, reducing reading times, and increasing meta-cognition have been considered the primary learning objectives in e-reading, based on the ability to sift vital information from others [139]. The knowledge gain evaluation has been conducted diagnostic, formative, and summative [212], with questions about finding global or local information, text organization, identifying main ideas, matching the sequence of events, and conclusions [139]. Several e-reading strategies have been suggested for better knowledge gain: exploring, finding, analyzing, and evaluating the reading material [213]. Furthermore, specific behavioral instructions have been suggested, such as oral reading and revisiting mistakes [214].

Setting up the short-term goal related to the result (i.e., product goal) and the process (i.e., process goal) has also been suggested, known to improve learners' self-efficacy, which positively affects the choice of activities, effort, persistence, and achievement of learners [215]. Observing the process goal, such as correct answers, test scores, and grades, was suggested [215]. As known to negatively affect student motivation, learning capabilities, and skill acquisition [215], resolving self-doubts in the learning process has also been

suggested as a relevant feedback role. Regarding human educators' feedback provision pattern in reading, more self-corrections were expected from high performers, while more frequent feedback was given to learners with lower learning achievements [214]. Some human instructors focused more on contextual cues that are more relevant to our work, aiming at overall comprehension. In contrast, some focused on specific cues that are more relevant to the proficiency of certain skills [214].

Perceived learning experiences

Perceived experience is often interpreted and evaluated as User Experience (UX) in diverse domains. One commonly referred definition of UX is users' perceptions and responses toward specific products and services based on users' usage and anticipations [78]. The increasing roles of conversational agents in everyday activities make the consideration of UX in HRI more important [211], which affects the overall system acceptance [216]. In this section, we look into the UX of the HRI, which is our focus as an intervention medium. The recent HRI evaluation has been criticized for its questionable validity and reliability of measures [217]. It seems to be partial because that UX can only be understood subjectively through perceived users' internal states [217], which makes the evaluation validation more critical. Another comes from the fact that UX design implementation and evaluation of intervention can only be understood through context, which requires the whole iteration as a package but often takes place separately in most practices [211]. However, since HRI is a relatively young research field, we still need the common theories, methods, models, and tools [211] and dedicated studies for HRI design for specified objectives.

Though HRI evaluation can be partially inspired by the field of Human-Computer Interaction (HCI), HRI needs more specific evaluation methods as opposed to comparatively traditional, passive, and computer-based facets of HCI [78]. In the same line, [211] indicated the necessity of a systematic approach in HRI evaluation to guarantee a positive user experience regarding the system's acceptance, usability, learnability, safety, trust, and credibility. The presence of robot agents also makes the understanding of robots more essential, such as contact with humans (e.g., physical robot, virtual robot), robot functionality (e.g., adaptive function), robot roles (e.g., assistant, companion, partner), and social skills of the robot (e.g., desirable to fundamental level) [217]. Understanding the functions of conversational agents' characteristics (e.g., speaking style, personality) and interaction properties (e.g., human-likeness [216]) is also emphasized, built upon users' interaction needs and their profiles [217]. [211] has focused on the roles of the HRI (i.e., do-goals, be-goals), looking into the psychological need fulfillment, positive affect, and product perception of the robot interaction [77]. [77] suggested pragmatic, hedonic-identity, hedonic-stimulation, and attractiveness as primary qualities of UX evaluation, while [211] suggested diverse qualities, such as relatedness, meaning, stimulation, competence, security, and popularity, as means to measure needs in various activities (e.g., watching, listening, playing). Attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty have been suggested as HRI UX evaluation measurements by [218], while users' reactions and feelings were focused by understanding perceived humanness, eeriness, and attractiveness for the robot acceptance in [219].

Perceived social presence

Social presence has been defined as the *sense of being in the company of another living being*, which has been widely investigated in social robot studies [220]. The social presence of HRI has often been understood as part of UX, as forms of dependability [218], perceived humanness [219], relatedness, security [218], and acceptance [216]. Perceived social presence is a significant aspect differentiating HRI from other systems with artificially embodied entities [220], including the HCI systems. As revealed in [158], learners perceive that they recognize, understand, and communicate better with the HRI system with the humanoid compared to the HCI system, leading to knowledge gain, even though both feedback conditions were the same other than the assistance of a robot. Through a meta-review, [221] has revealed that the in-person HRI poses positive effects on the combined outcomes, efficacy, perceptions, and attitudes toward systems, compared to the remote HRI, indicating the significant effects of the "physicality" of in-person HRI interfaces on the learner perceptions. In this regard, we understand the perceived social presence of physical humanoids as our focus in this section, separately from the perceived UX. Perceived presence is known to enhance learner participation, satisfaction, [222], cognition, and critical thinking [223]. Also, the sense of social presence is known to aid learners' physical, emotional, and cognitive health in remote education, which seems especially relevant to the recent online education in the post-pandemic era [220].

As means to evaluate the social presence of social robots, the following measures have been investigated: perceived robot appearances [224], rapport building and relationship dynamics [225], immersion, parasocial interaction, parasocial relationships, physiological responses, social reality, and general social richness [226], salience, perceived actor-hood, co-location/non-mediation, understanding, association, involvement, and medium sociability [227]. Not merely focusing on perception towards the interaction medium itself (i.e., robot), the perception toward the message has also empathized that are relevant to the overall conversational agents [227]: attentional allocation, perceived message understanding, perceived affective understanding, perceived emotional interdependence, and perceived behavioral interdependence [228]. [229] suggested the different effects come from learner groups with varied consciousness, indicating that the higher social presence is associated with the perceived learning and satisfaction in learners with low consciousness. In contrast, the social presence did not affect the perceived learning or satisfaction in the highly-conscious learners. Studies have suggested enhancing the social presence of learners: Using scaffolded and self-reflective topics for better self-disclosure, [230], facilitating small group discussions [231], utilizing the storytelling [232], and providing personalized features in implementation, such as personal profiles, text messages, individualized video feedback, and one-on-one email communication [229].

Attention regulation behaviors in e-reading

Physical reading behaviors have been used as measurements to understand learners' engagement and visual attention in e-reading, having various sensors, such as eye tracker [14, 233, 234], motion sensors [23, 24], webcam [19, 156, 158], 3D-camera [116] and log data layers [235], implemented. However, webcam-based attention feature extraction has rarely been attempted, which could significantly assist the real-world feedback loop design without complicated sensor implementations in various learning scenarios. This work implements the webcam-based e-reading attention recognition framework of [156] for
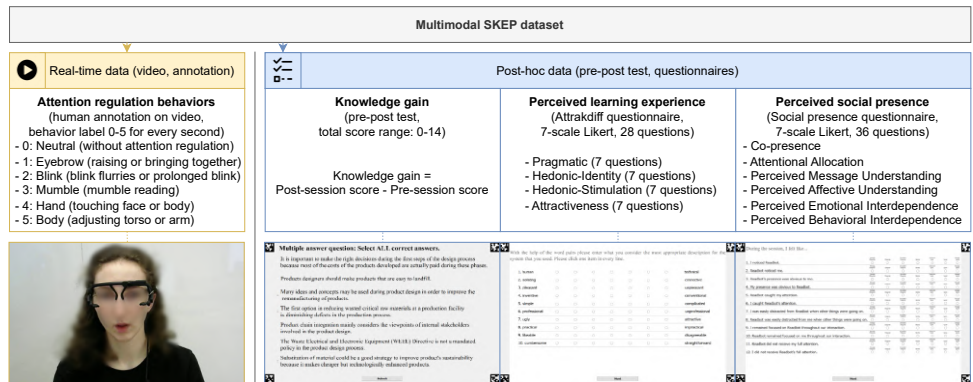
Figure 5.2: Multimodal SKEP dataset for attention regulation behaviors, knowledge gain, perceived learning experience, and perceived social presence in e-learning with a conversational agent.

attention regulation behavior annotation. [156] suggested attention regulation behavior as a critical cue where learners are aware of their attention loss and try to regain their focus in e-reading. The behaviors have been movements from eyebrow (e.g., eyebrow raise, bring together), blinks (e.g., blink flurries, prolonged voluntary blinks), mumble (e.g., mumble reading), hand (e.g., touch body, face), and body (e.g., adjust torso, arm, head), as opposed to neutral state without movements mentioned. Such behaviors have been revealed to correlate significantly with self-reported distractions of learners, indicating the behaviors as signs of attention loss and the following attention self-regulation. Video-based deep learning models have been implemented as a good predictor of self-reported distractions [156], knowledge gain, perceived learning experiences, and perceived social presence [158], respectively in the real-world setting [156], as well as in the laboratory-based-setting with the HRI system implemented [158].

5.2 Data pre-processing and unsupervised clustering based on factual learning outcomes

This section introduces the dataset and performs data pre-processing to construct features used for unsupervised learning. We represent feature constructions and further conduct feature engineering by comparing the silhouette scores of manually and automatically selected sets of features. We conduct unsupervised training based on various modeling methods (e.g., k-means, hierarchical, DBSCAN, spectral clustering) and validate clusters via the Chi-square test. In the process, we expect to tackle the following research question:

- RQ1. How can learner features best reflect learners' performances, experiences, and perceptions of conversational agents' interventions in e-reading?

5.2.1 Multimodal SKEP Dataset

We utilize a multimodal SKEP Dataset (see Fig.5.2) collected from 60 higher-education learners who use the English language for their daily education [201]. Participants were recruited for an e-reading task on the screen (Age: M= 24.9, SD: 3.92; Gender: 37 males, 23

females). They were given an e-reading system with emphatic and metacognitive feedback from conversational agents through pop-ups and speech from a Furhat Robot [236], a conversational agent in a humanoid robot form.

Before the e-reading, participants were given a pre-test with 14 questions to measure their prior knowledge about the topic as a diagnostic knowledge measurement tool. The e-reading content has had seven subsections with 4,750 words concerning "Waste management and critical raw materials". In the process, learners' self-reports from the pre-post test (e.g., knowledge gain) and questionnaires (e.g., perceived learning experience and perceived social presence) were collected. At the end of every subsection of the screen-based e-reading material, pop-up questions were given as formative measurements. At the end of all subsections, another seven questions were given as a summative measurement tool. Additionally, two post-session questionnaires took place to understand learners' perceptions of the learning experiences and their perception of the system as a social presence, respectively: the Attrakdiff questionnaire with 28 questions, which have pragmatic, hedonic-identity, hedonic-stimulation, and attractiveness as its subdimensions, and the Social Presence questionnaire with 36 questions that concerns co-presence, attention allocation, perceived message understanding perceived affective understanding, perceived emotional interdependence, and personal behavioral interdependence.

Also, throughout the experiment, the video data were collected through a webcam, and multiple annotators later annotated learners' behaviors for attention regulation. The video data contains a total of 2,339 minutes, reaching 40 hours. The video samples were segmented every second, and 140340 frames were annotated into five attention regulation behaviors (e.g., movements from eyebrow, blink, mumble, hand, body) and one neutral label that was further cross-validated. Note that learner data from GUI-based or HRI-based conversational agents from the SKEP dataset were not considered differently in this work. It is because our data-driven persona aims to see learners' perceptions and responses toward the feedback system regardless of the specific type of feedback.

5.2.2 Manual vs. Automatic Feature Selection

Feature vectors representing the best subset of variables are often scrutinized in two different ways: manually and automatically [237]. Manual feature selection is conducted based on a good understanding of the domain and dataset, often criticized for human bias and having deviated results from different evaluators. Automatic feature selection is especially beneficial in high-dimensional data where dimension reduction of data is essential and manual selection cannot achieve the utmost efficiency. However, automatic methods also have limitations, such as information loss and low interpretability in results. To achieve both semantically and scientifically sound results, we conducted the feature selection 1) by manually dividing categorical features into three semantic levels and 2) by conducting the automatic Principal Component Analysis (PCA) based on the percentage of consensus in generalized Procrustes analysis. We compared the silhouette scores of different sets of features, which offers the best distinctions among clusters. We found the best-performing method, which helped us find the optimal feature vectors with the best consistency of data clusters. The silhouette analysis and further applied elbow method are used to understand the number of optimal clusters for future unsupervised training. Note that mean-max normalization was applied to the SKEP dataset to make the subsets

Table 5.1: SKEP dataset with low-dimensional, mid-dimensional, and high-dimensional features.

Low-dimensional Features	Mid-dimensional Features	High-dimensional Features
Knowledge Gain	Knowledge Gain	Knowledge Gain
Perceived Learning Experience	-	-
	Pragmatic Quality	7 Sub-questions
	Hedonic-Identity	7 sub-questions
	Hedonic-Stimulation	7 sub-questions
	Attractiveness	7 sub-questions
Perceived Social Presence	-	-
	Co-presence	6 sub-questions
	Attentional Allocation	6 sub-questions
	Perceived Message Understanding	6 sub-questions
	Perceived Affective Understanding	6 sub-questions
	Perceived Emotional Interdependence	6 sub-questions
	Perceived Behavioral Interdependence	6 sub-questions

identical to avoid potential bias from the different data ranges.

Manual Feature Selection

As can be seen from Table 5.1 and equation 5.1, 5.2, and 5.3, the SKEP dataset has data with three layers: 1) *low-dimensional features* with three components (e.g., knowledge gain, perceived learning experience, perceived social presence), 2) *mid-dimensional features* with 11 components (e.g., knowledge gain, pragmatic, hedonic-identity, hedonic-stimulation, attractiveness, co-presence, attentional allocation, perceived message understanding, perceived affective understanding, perceived emotional interdependence, perceived behavioral interdependence measures), and 3) *high-dimensional features* with 65 components (e.g., knowledge gain, seven sub-questions of pragmatic, hedonic-identity, hedonic-stimulation, attractiveness measures, six sub-questions of co-presence, attentional allocation, perceived message understanding, perceived affective understanding, perceived emotional interdependence, perceived behavioral interdependence measures). Those are three levels of features with various dimensionality that make semantic sense to most human evaluators based on the information hierarchy. Thus, we used those three levels of dimensional features as manually selected features, which are listed in Table 6.1.

$$KnowledgeGain = \sum_{i=1}^{N=7} Score_i^{PreSession} + \sum_{i=1}^{N=7} Score_i^{InSession} - \sum_{i=1}^{N=14} Score_i^{PostSession} \tag{5.1}$$

$$PerceivedLearningExperience = \frac{\sum_{i=1}^{N=7} Score_i^{PragmaticQuality} + \sum_{i=1}^{N=7} Score_i^{Hedonic-Identity} + \sum_{i=1}^{N=7} Score_i^{Hedonic-Simulation} + \sum_{i=1}^{N=7} Score_i^{Attractiveness}}{28} \tag{5.2}$$

$$PerceivedSocialPresence = \frac{\sum_{i=1}^{N=6} Score_i^{Co-presence} + \sum_{i=1}^{N=6} Score_i^{AttentionalAllocation} + \sum_{i=1}^{N=6} Score_i^{PerceivedMessageUnderstanding} + \sum_{i=1}^{N=6} Score_i^{PerceivedAffectiveUnderstanding} + \sum_{i=1}^{N=6} Score_i^{PerceivedEmotionalInterdependence} + \sum_{i=1}^{N=6} Score_i^{PerceivedBehavioralInterdependence}}{36} \tag{5.3}$$

Automatic Feature Selection Based on PCA

We conducted the PCA to achieve an automatic feature selection. PCA is often used for unsupervised learning to reduce the data complexity by reducing the noise and the
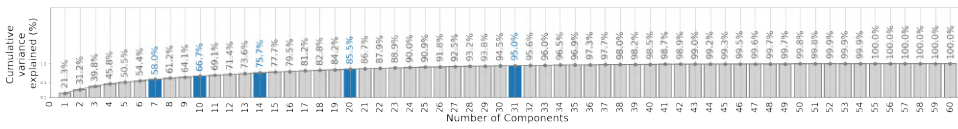
Figure 5.3: Number of principal components for explaining variance. 7, 10, 14, 20, and 31 components were required to explain the 55%, 65%, 75%, 85%, and 95% variances, respectively.

dimensionality of data. By only selecting the principal components that explain the greatest amount of variance, the computation becomes lighter with better clarity in convoluted and multi-directional factors with minimal information loss. The equation below shows that the PCA produces a linear composition of the original components until the d dimensions, from the highest variance in the first element to the lowest variance in the last element. The newly created k is called the principal component, which decides the new dimension of subsets. Note that $k<d$.

$$PC_i = a_1 X_1 + a_2 X_2 + \cdots + a_d X_{d'}, \tag{5.4}$$

where X_j is the initial function a_j. j is the ith PC, while a_j is X_j number coefficient.

As [238] indicated, 70% of explained variance is common, while [239] applied a total variance ratio greater than 80% to reveal the most critical variables through the PCA. Below, we applied variously explained variances to find the number of features required to achieve specific proportions of explained variance. Note that 55%, 65%, 75%, 85%, and 95% have been applied as the proportions of explained variance (see Fig.5.3). Feature numbers derived from each proportion of explained variance have been applied for the silhouette analysis in the next section (see Table 5.2). Note that the number of components in Fig.5.3 is 60, equivalent to the sample number since $sample numbers < feature numbers$ in this dataset. In such a case, the PCA automatically takes the sample numbers as the feature numbers.

Feature Selection Methods Comparision: Silhouette Analysis

We have conducted silhouette analysis on manually selected features and automatically selected features to find the best-distinguished clusters from the feature selection. Note that the silhouette coefficient ranges between -1 and 1, and a score close to 1 indicates the best performance.

In this study, silhouette analysis has been applied for two purposes: 1) measuring the quality of the clusters based on different feature vectors as a part of the feature selection process and 2) getting the first indication of the optimal number of clusters. See Table 5.2 for the silhouette coefficients derived from manually and automatically selected sets of features. Various cluster numbers have been applied in the analysis process for further insights. The result shows the best silhouette score has been achieved when manually selected low-dimensional data has been applied, indicating the optimal number of clusters as 6. Thus, this work uses knowledge gain, perceived learning experiences, and perceived social presence as three feature vectors for unsupervised model training.

We assume that the PCA did not improve the performance of silhouette analysis, seemingly because the PCA is based on the noise and the corresponding dimension reduction in the dataset. In the PCA process, some essential data structures or features might have been

Table 5.2: Silhouette analysis conducted on manually selected features and automatically selected features as a part of feature selection.

		Manual Feature Selection			Automatic Feature Selection				
		Low-dimensional (3 Features)	Mid-dimensional (11 Features)	High-dimensional (65 Features)	PCA (31 Features)	PCA (20 Features)	PCA (14 Features)	PCA (10 Features)	PCA (7 Features)
	2	0.679	**0.451**	**0.282**	**0.297**	**0.276**	**0.261**	**0.265**	**0.324**
	3	0.604	0.397	0.194	0.171	0.188	0.201	0.217	0.232
Number	4	0.641	0.406	0.174	0.114	0.126	0.118	0.129	0.159
of	5	0.647	**0.451**	0.149	0.094	0.083	0.101	0.118	0.148
Clusters	6	**0.729**	0.367	0.143	0.077	0.076	0.103	0.131	0.125
	7	0.670	0.338	0.116	0.026	0.059	0.085	0.076	0.099
	8	0.531	0.250	0.115	0.020	0.042	0.090	0.070	0.097

(a) k-means.   (b) Hierarchical.   (c) DBSCAN.   (d) Spectral.
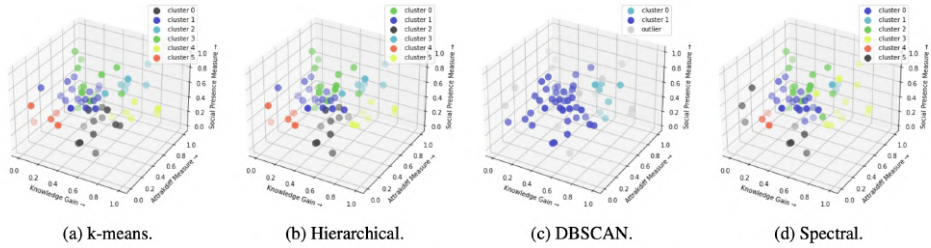
Figure 5.4: Visualized results of different unsupervised learning methods: k-means, hierarchical, DBSCAN, spectral clustering methods.

damaged, while all features were restructured as linear data and de-noised. In some cases, the neighboring clusters might have been too close when feature selection was made based on the PCA.

5.3 Unsupervised Learning for Learner Pattern Clustering and Comparative Analysis

In this section, we implement different unsupervised learning methods for further comparative analysis, suggested in previous review [191], but has yet to be attempted in data-driven persona development studies. We compare four unsupervised methods with various hyperparameters to evaluate the result consistency among methods as cross-validation. In the process, we tackle the following research question:

- RQ2. How can unsupervised learning methods be used for learner pattern clustering and validation?

Specifically, we implemented k-means clustering, agglomerative hierarchical clustering, DBSCAN clustering, and spectral clustering methods that represent centroid, hierarchy, density, and graph-based methods, respectively (see Fig.5.4 for the 3-D visualization of the clusters).

5.3.1 Cross-validating clusters from various modeling methods via Chi-square test

In this section, we apply the Chi-square test to validate the cluster distributions derived from different modeling methods. The chi-square test is a frequently applied method

to determine the statistical differences and homogeneity in one or more categories of a contingency. We premised that the clusters are well-defined if homogeneity is found among the models. As can be seen from Table 5.3, significant p values are observed, indicating a significant relationship between categorical variables built upon different models. Thus, from the following section, we use the cluster distribution derived from the k-means clustering method, considering the model capability of being applied to larger datasets for future research reproduction with increased samples. Note that the Chi-square test has been conducted based on independent categorical cluster inputs (e.g., A, B, C, etc.).

Table 5.3: Chi Squared Test Applied to Unsupervised Models with Different Parameters.

		k-means			Hierarchical												Ward			DBSCAN						Spectral		
					Single			Average			Complete			Ward			Min Outlier			Max eps								
		Value	df	p	Value	df	p	Value	df	p	Value	df	p	Value	df	p	Value	df	p	Value	df	p	Value	df	p
k-means		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Hierarchical	Single	163	25	<.001*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Average	163	25	<.001*	300	25	<.001*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Complete	163	25	<.001*	300	25	<.001*	300	25	<.001*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Ward	163	25	<.001*	300	25	<.001*	300	25	<.001*	300	25	<.001*	-	-	-	-	-	-	-	-	-	-	-	-
DBSCAN	Min Outlier	37.9	10	<.001*	52.9	10	<.001*	52.9	10	<.001*	52.9	10	<.001*	52.9	10	<.001*	-	-	-	-	-	-	-	-	-
	Max eps	77.6	15	<.001*	128	15	<.001*	128	15	<.001*	128	15	<.001*	128	15	<.001*	73.4	6	<.001*	-	-	-	-	-	-
Spectral		163	25	<.001*	300	25	<.001*	300	25	<.001*	300	25	<.001*	300	25	<.001*	52.9	10	<.001*	128	15	<.001*	-	-	-

The *single* linkage combines clusters with the nearest data pair. The *average* linkage combines clusters with the closest average distance between clusters. The *complete* linkage combines clusters, which have the maximum distance between any two points of different clusters. The *ward* linkage combines clusters with minimized within-cluster variances (=sum-of-squared distance) of all clusters. We apply the Eps of 0.205, where the result shows the least outlier, while the $K > 1$. We apply the Eps of as 0.19, where the Eps has the smallest value before the exponential increase in the number of clusters at 3.

5.4 Data-driven Persona Development and Statistical Interpretation of Each Cluster

In this section, we conduct the statistical analysis on each cluster derived from the k-means clustering method. In this section, unique features from each cluster are derived based on quartile analysis to tackle the following research question:

- RQ3. How can we extract valuable archetypes of learners from different clusters and develop data-driven personas based on them?

5.4.1 Archetype Extraction Based on Quartile Analysis using the Low-dimensional Data

As shown in Table 5.4, we conducted the statistical analysis, built upon clusters derived from the k-means clustering on factual learning outcomes. We first find the average feature of all learners from all clusters. We further conduct the quartile analysis on each cluster and see where each cluster is located from the whole set by comparing the *mean* value of each cluster, which represents the most typical learner in the cluster, and the *quartile* ranges from all learners.

Quartile analysis provides statistically critical information about the center point and the spreads of the data [240]. It shows where a specific learner cluster is located from the overall learner data points. Based on the quartile, we interpreted the learners into three levels: if *mean of cluster<1st quartile (25%) of all learners*, we interpreted it as *low*, which means that learners in the cluster show less tendency of having the specific feature than the average learners. If *1st quartile (25%) of all learners<mean of cluster<2nd quartile (50%) of all*

Table 5.4: Statistical analysis conducted on clusters derived from k-means clustering result. Factual learner features, such as knowledge gain, perceived learning experience, and perceived social presence, have mainly been investigated through quartile analysis.

		Counts	M	SD	Min	25%	50%	75%	Max	Interpretation
					Statistical analysis (k-means)					
	Overall	60	0.431	0.249	0.000	**0.250**	**0.416**	**0.604**	1.000	
	Persona A	19	**0.438**	0.123	0.250	0.375	0.416	0.500	0.666	**Mid**
Knowledge	Persona B	12	**0.201**	0.164	0.000	0.083	0.116	0.333	0.500	**Low**
Gain	Persona C	8	**0.760**	0.150	0.500	0.729	0.750	0.791	1.000	**High**
	Persona D	8	**0.135**	0.098	0.000	0.062	0.166	0.187	0.250	**Low**
	Persona E	6	**0.736**	0.081	0.666	0.666	0.708	0.812	0.833	**High**
	Persona F	7	**0.511**	0.089	0.333	0.500	0.500	0.583	0.583	**Mid**
	Overall	60	0.491	0.244	0.000	**0.310**	**0.444**	**0.652**	1.000	
	Persona A	19	**0.386**	0.120	0.129	0.296	0.388	0.444	0.611	**Mid**
Perceived	Persona B	12	**0.819**	0.128	0.592	0.787	0.824	0.898	1.000	**High**
Learning	Persona C	8	**0.581**	0.151	0.314	0.541	0.629	0.657	0.759	**Mid**
Experience	Persona D	8	**0.236**	0.007	0.148	0.185	0.194	0.300	0.370	**Low**
	Persona E	6	**0.635**	0.171	0.370	0.574	0.648	0.694	0.888	**Mid**
	Persona F	7	**0.280**	0.189	0.000	0.166	0.259	0.425	0.518	**Low**
	Overall	60	0.482	0.206	0.000	**0.327**	**0.523**	**0.606**	1.000	
	Persona A	19	**0.590**	0.106	0.396	0.551	0.584	0.617	0.811	**Mid**
Perceived	Persona B	12	**0.438**	0.208	0.047	0.341	0.433	0.603	0.745	**Mid**
Social	Persona C	8	**0.358**	0.114	0.160	0.308	0.334	0.433	0.518	**Mid**
Presence	Persona D	8	**0.435**	0.162	0.235	0.320	0.415	0.530	0.726	**Mid**
	Persona E	6	**0.768**	0.132	0.632	0.693	0.726	0.816	1.000	**High**
	Persona F	7	**0.215**	0.122	0.000	0.155	0.264	0.301	0.330	**Low**

learners, we interpreted it as *Mid*, which means that learners in the cluster are located in the average range of the particular feature (e.g., knowledge gain, perceived learning experience, perceived social presence). If *2nd quartile (50%) of all learners<mean of cluster<3rd quartile (75%) of all learners*, we interpreted it as *high*, which means that learners in the cluster show the strong tendency of having the particular feature than the average learners.

5.4.2 Archetype Extraction Based on Quartile Analysis using the Mid-dimensional and High-dimensional Data: Top-down Approach

We applied the quartile analysis to the mid and high-dimensional data to understand learners based on more detailed artifacts. While the quartile analysis on the low-dimensional data provides a general understanding of learner clusters, the top-down approach based on the mid and high-dimensional data lets a vivid understanding of learners based on more detailed features. See Fig.5.5 for the visualized archetype based on the mid-dimensional data. See Table 5.5 for the detailed archetype descriptions based on the high-dimensional data.

5.4.3 Data-driven Personas built upon archetypes of different clusters

Persona A: archetypes derived from cluster 0

Persona A has been the most common type among all (60 participants), having 19 participants (31.67%) in the same cluster. *Persona A* has shown no significant knowledge gain and perceived learning experiences. In the perceived social presence measure, *persona A* did not show significant variances from the average learners, aside from one sub-measure from perceived message understanding, that "it was easy to understand Readbot (i.e., feedback system with conversational agents)". All in all, *persona A* is the most average type of learner among all participants.

(a) Persona A.   (b) Persona B.   (c) Persona C.

(d) Persona D.   (e) Persona E.   (f) Persona F.

Figure 5.5: Archetype extraction and visualization based on the mid-dimensional data.

Table 5.5: Archetype extraction and quartile analysis based on the high-dimensional data.

Low-dimensional	Mid-dimensional	High-dimensional	Persona A High (Q3)	Persona A Low (Q1)	Persona B High (Q3)	Persona B Low (Q1)	Persona C High (Q3)	Persona C Low (Q1)	Persona D High (Q3)	Persona D Low (Q1)	Persona E High (Q3)	Persona E Low (Q1)	Persona F High (Q3)	Persona F Low (Q1)
Knowledge Gain			·	·	✓	·	✓	·	·	✓	✓	·	·	·
Perceived Learning Experience	Pragmatic	Q1. Technical (Low)-Human (High)	·	·	✓	·	·	·	·	·	·	·	·	·
		Q8. Impractical (Low)-Practical (High)	·	·	·	·	·	·	·	·	✓	·	·	·
		Q12. Unpredictable (Low)-Predictable (High)	·	·	·	·	·	·	·	·	✓	·	·	·
	Hedonic-Identity	Q13. Cheap (Low)-Premium (High)	·	·	·	·	·	·	·	·	·	·	·	✓
		Q15. Separates me from people (Low)-Brings me closer to people (High)	·	·	·	·	·	·	·	·	·	✓	·	·
	Hedonic-Stimulation	Q4. Conventional (Low)-Inventive (High)	·	·	·	·	·	·	✓	·	·	·	·	·
		Q24. Dull (Low)-Captivating (High)	·	·	·	·	·	·	·	·	·	·	·	✓
		Q27. Ordinary (Low)-Novel (High)	·	·	·	·	·	·	·	·	·	·	·	✓
	Attractiveness	Q3. Unpleasant (Low)-Pleasant (High)	✓	·	·	·	·	·	·	·	·	·	·	·
		Q9. Disagreeable (Low)-Likable (High)	✓	·	·	·	·	·	✓	·	·	·	·	·
		Q21. Repelling (Low)-Appealing (High)	✓	·	·	·	✓	·	✓	·	✓	·	·	·
		Q26. Discouraging (Low)-Motivating (High)	✓	·	·	·	·	·	✓	·	·	·	·	✓
Perceived Social Presence	Co-presence	Q1. I noticed Readbot.	·	·	·	·	·	·	·	·	·	·	·	✓
		Q2. Readbot noticed me.	·	·	·	·	·	·	·	·	·	·	·	✓
		Q3. Readbot's presence was obvious to me.	·	·	·	·	✓	·	·	·	✓	·	·	✓
		Q4. My presence was obvious to Readbot.	·	·	·	·	·	·	·	·	·	·	·	✓
		Q5. Readbot caught my attention.	·	·	·	·	✓	·	·	·	✓	·	·	✓
		Q6. I caught Readbot's attention.	·	·	·	·	·	·	·	·	·	·	·	✓
	Attention Allocation	Q10. Readbot remained focused on me throughout our interaction.	·	·	·	·	·	·	·	·	·	·	·	✓
	Perceived Message Understanding	Q14. Readbot was clear to me.	·	·	·	·	·	·	·	·	✓	·	·	·
		Q15. It was easy to understand Readbot.	·	✓	·	·	·	·	·	·	·	·	·	✓
		Q17. Understanding Readbot was difficult.	·	·	·	·	·	·	✓	·	·	·	·	·
	Perceived Affective Understanding	Q19. I could tell how Readbot felt.	·	·	·	·	·	·	·	·	✓	·	·	·
		Q20. Readbot could tell how I felt.	·	·	·	·	·	·	✓	·	·	·	·	·
		Q22. My emotions were not clear to Readbot.	·	·	·	·	·	·	·	·	·	✓	·	·
		Q23. I could describe Readbot's feelings accurately.	·	·	·	·	✓	·	✓	·	·	·	·	·
		Q24. Readbot could describe my feelings accurately.	·	·	·	·	·	·	·	·	✓	·	·	·
	Perceived Emotional Interdependence	Q25. I was sometimes influenced by Readbot's moods.	·	·	·	·	·	·	✓	·	✓	·	·	·
		Q26. Readbot was sometimes influenced by my moods.	·	·	·	·	·	·	·	·	✓	·	·	✓
		Q27. Readbot's feelings influenced the mood of our interaction.	·	·	·	·	·	·	✓	·	✓	·	·	✓
		Q29. Readbot's attitudes influenced how I felt.	·	·	·	·	·	·	·	·	✓	·	·	✓
		Q30. My attitudes influenced how Readbot felt.	·	·	·	·	·	·	·	·	✓	·	·	✓
	Perceived Behavioral Interdependence	Q31. My behavior was often in direct response to Readbot's behavior.	·	·	·	·	·	·	·	·	·	·	·	✓
		Q32. The behavior of Readbot was often in direct response to my behavior.	·	·	·	·	·	·	·	·	·	·	·	✓
		Q33. I gave and took Readbot's actions mutually.	·	·	·	·	·	·	·	·	·	·	·	✓
		Q34. Readbot's gave and took my actions mutually.	·	·	·	·	·	·	·	·	✓	·	·	✓
		Q35. Readbot's behavior was closely tied to my behavior.	·	·	·	·	·	·	·	·	✓	·	·	✓
		Q36. My behavior was closely tied to Readbot's behavior.	·	·	·	·	·	·	·	·	✓	·	·	✓

Persona B: archetypes derived from cluster 1

Persona B has been the second most common type of learner group among all participants, having 12 learners (20.0%) in the same segment. *Persona B* has achieved the second lowest knowledge gain compared to other groups. However, *Persona B* has evaluated the system as partially pragmatic and most attractive among all groups. The feedback from conversational agents has been evaluated as "human", "pleasant", "likable", "appealing", and "motivating" by *persona B*. *Persona B* did not show any significant perceived social presence. *Persona B* is the learner type that perceives the system positively and has a good learning experience. However, it did not lead to good knowledge gain, which is against of notion that the quality of the learning experience somewhat leads to positive learning outcomes.

Persona C: archetypes derived from cluster 2

Persona C was derived from eight learners (13.33%). *Persona C* has shown high knowledge gain among all participants and found the system "appealing" in the attractiveness of learning experience evaluation. However, *persona C* has responded generally negatively to the social presence measures, especially in co-presence, perceived emotional interdependence, and perceived affective understanding. *Persona C* has evaluated that "Readbot did not notice me." and "Readbot did not catch my attention.", showing low sense of co-presence. Furthermore, regarding perceived emotional interdependence, *persona C* answered that "I could not describe Readbot's feeling accurately.". *Persona C* has responded that "I was not influenced by Readbot's moods." and "Readbot's mood did not influence the mood of our interaction.", showing the lower perceived emotional interdependence in two sub-measures. All in all, *persona C* is the learner type that performs highly in knowledge gain, regardless of mediocre learning experience and mediocre to low perceived social presence of the system. Person C is a learner type that has trouble relating to conversational agents due to his or her low co-presence with the system. However, the knowledge gain has been achieved highest among all learners groups.

Persona D: archetypes derived from cluster 3

Persona D has derived from eight learners (13.33%). *Persona D* has achieved the lowest knowledge gain among all participant groups. Also, *persona D* has evaluated the perceived learning experience among all participant groups, especially in attractiveness and pragmatic value of the system, perceiving the system as "disagreeable", "repelling", "discouraging", and "conventional", respectively. In perceived social presence, *persona D* has provided answers within the mid-range. However, in some perceived affective understanding sub-measures, indicating that "Understanding Readbot was difficult." and "Readbot could not tell how I felt.", while perceiving that, " I could describe Readbot's feelings accurately.". Overall, *persona D* is regarded as the learner type who performs poorly in knowledge gain based on a poor perceived learning experience with the system. *Persona D* seemed discouraged and repelled by the system that did not understand him or herself, likely in awareness that the feedback was not based on their responses (i.e., intelligent system), having no difference from the conventional one-way feedback system. In that regard, It seems that a better interaction design based on an intelligent system might bring a better-perceived learning experience and subsequent improvements in knowledge gain for *persona D*.

Persona E: archetypes derived from cluster 4

Persona E has been built based on data from six learners (10.0%). *Persona E* has recorded the second-highest knowledge gain among all learner groups. *Persona E*'s evaluation of his or her learning experience was average. However, *Persona E* has evaluated the pragmatic value of the conversational agents poorly, perceiving the feedback as 'impractical" and "unpredictable". However, *persona E* evaluated the system as "appealing". The most distinctive feature of *persona E* came from its generally high perceived social presence, which has not been found in other groups. The tendency has shown more obvious in assessing perceived emotional interdependence, reporting their perceptions as "I was sometimes influenced by Readbot's mood.", "Readbot was sometimes influenced by my mood.", "Readbot's feelings influenced the mood of our interaction.", and "Readbot's attitudes influenced how I felt.". Accordingly, *persona E* has shown high perceived behavioral interdependence, perceiving that "Readbot's gave and took my actions mutually.", "Readbot's behavior was closely tied to my behavior.", and "My behavior was closely tied to Readbot's behavior.". Moreover, in the co-presence sub-measures, *persona E* responded that "Readbot noticed me." and "Readbot caught my attention.". *Persona E* has also reported that "I could tell how Readbot tells.", "Readbot could describe my feelings accurately.", and "Readbot was clear to me.", showing high perceived message understanding and per affective understanding compared to other groups of participants.

Persona F: archetypes derived from cluster 5

Persona F has been developed based on seven learners (11.67%). *Persona F* did now show any significant knowledge gain compared to other groups of participants. The general perceived learning experience and social presence have been the lowest. *Persona F* has evaluated the system as "cheap", "dull", and "ordinary", in Hedonic-Identity and Hedonic-Stimulation measures. In the attractiveness sub-measures, *Persona F* found the system "discouraging".

In terms of perceived social presence, *Persona F*'s responses toward co-presence and perceived behavioral interdependence were all negative, indicating that "I did not notice Readbot.", "Readbot did not notice me.", "Readbot's presence was not obvious to me.", "My presence was not obvious to Readbot.", "Readbot did not catch my attention.", "I did not catch Readbot's attention.", and "My behavior was not in direct response to Readbot's behavior.", "The behavior of Readbot was not in direct response to my behavior.", "I did not give and take Readbot's actions mutually.", "Readbot's did not give and take my actions mutually.", "Readbot's behavior was not closely tied to my behavior.", and "My behavior was not closely tied to Readbot's behavior.", respectively. Also, *persona F*'s perceived emotional interdependence was also low, responding that "Readbot was not influenced by my mood.", "Readbot's feelings did not influence the mood of our interaction.", "Readbot's feelings did not influence the mood of our interaction.", "Readbot's attitudes did not influence how I felt.", and "My attitudes did not influence how Readbot felt.". Low attention allocation and perceived message understanding sub-measures from *persona F* have shown that "Readbot did not remain focused on me throughout our interaction." and "It was not easy to understand Readbot.".

Overall, *Persona F* did not consider conversational agents as beings with identity or being good hedonic stimuli to e-reading. At the same time, poorly perceived co-presence seemed to lead to *persona F*'s low perceived emotional interdependence and behavioral interdepen-

dence, subsequently. Interestingly, low perceived learning experience and social presence did not negatively impact the knowledge gain of *persona F*. However, it also indicates room for improvement in knowledge gain if guaranteed a better-perceived learning experience and social presence of conversational agents.

5.5 Automatic persona predictions based on attention regulation behaviors

Attention regulation behaviors are proven to be a robust predictor of learners' attention in e-reading [156]. In this section, we study if persona prediction can also be achieved using attention regulation behaviors. We implement multiple classification models to classify different patterns of personas via the attention regulation behaviors of learners. We utilized the cross-subject evaluation protocol in all classification tasks. We chose the classical 70-30 protocol of dividing 60 samples into 40 for training and 20 for testing. We also scrutinize which part of the video sample can provide the best clues for persona prediction by introducing comparative learning phase-based and time-based prediction approaches. Also, we compared two different sampling methods of instant and cumulative learner behavior labels. In the process, the research question below is answered:

- RQ4. Can we predict learner personas based on attention regulation behaviors?

5.5.1 Learning Phased-based & Time Duration-based Persona Prediction

This section implements four classical machine learning classifiers: AdaBoost, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and Random Forest. We predict learner personas based on their behaviors shown during 1) various phases of learning and time points based on 2) instant and cumulative behavioral data points. 3) we train on instant and cumulative data so that our work can contribute to the real-time feedback loop by investigating behavioral clues for predicting various personas.

5.5.2 Six-class Cluster Prediction (Multiclass Classification Task)

Six-class persona prediction via attention regulation behavior has been conducted to differentiate all six personas (A-F) derived in the previous data-driven persona development section. As seen from Table 5.6, The best performances have been 45% of accuracy, using SVM and kNN applied to cumulative behaviors shown in various learning phases; the same performance has been achieved in the kNN and Random Forest, using the time duration-based method in 25%~50% of reading duration. It is a significantly higher performance than the random guess of 17%. There has been a general tendency that behavior from instant behavior data from subtopic 5 data has achieved better accuracy than the other part of instant data. Likewise, cumulative behavior data shown throughout subtopics 1-6 has derived the best result, with 45% as the best accuracy.

5.5.3 Four-class persona Prediction (Multiclass Classification Task)

We further conducted the four-class persona prediction (see Table 5.7). We selected three personas that we found to have the feedback necessity among six personas: two personas with low knowledge gain with a low and high perceived learning experience, respectively

Table 5.6: Six-class persona prediction based on the learning phased-based & time duration-based learner behavior data.

Instant		Learning phase-based							Time duration-based			
		Subtopic 1	Subtopic 2	Subtopic 3	Subtopic 4	Subtopic 5	Subtopic 6	Subtopic 7	~25%	25%-50%	50%-70%	75%~
	Random Guess	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
	AdaBoost	0.15	<u>0.30</u>	0.25	0.20	**0.35**	0.20	<u>0.30</u>	<u>0.30</u>	<u>0.30</u>	<u>0.30</u>	**0.35**
	SVM	0.15	0.25	0.25	0.15	**0.40**	0.20	0.25	0.30	**0.40**	0.20	<u>0.35</u>
	kNN	0.30	0.25	0.30	0.20	**0.40**	0.20	0.30	0.30	<u>0.35</u>	0.30	0.30
	Random Forest	<u>0.35</u>	0.25	0.30	0.20	**0.40**	0.20	0.30	<u>0.35</u>	<u>0.35</u>	0.20	0.30
Cumulative		Subtopic 1	Subtopic 1-2	Subtopic 1-3	Subtopic 1-4	Subtopic 1-5	Subtopic 1-6	Subtopic 1-7	~25%	~50%	~75%	~100%
	Random Guess	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
	AdaBoost	<u>0.30</u>	0.20	0.20	<u>0.30</u>	**0.40**	**0.40**	0.15	<u>0.30</u>	<u>0.30</u>	<u>0.30</u>	0.25
	SVM	0.15	0.20	0.30	0.30	0.25	**0.45**	0.10	0.35	<u>0.40</u>	0.20	0.35
	kNN	0.15	<u>0.30</u>	0.10	<u>0.30</u>	0.25	**0.45**	0.10	<u>0.30</u>	**0.45**	0.25	<u>0.30</u>
	Random Forest	0.15	0.25	0.15	0.30	0.30	0.30	0.10	0.30	**0.45**	0.30	<u>0.35</u>

The best and the second best performances are **bolded** and <u>underlined</u>, respectively.

Table 5.7: Four-class persona prediction based on the learning phased-based & time duration-based learner behavior data.

Instant		Learning phase-based							Time duration-based			
		Subtopic 1	Subtopic 2	Subtopic 3	Subtopic 4	Subtopic 5	Subtopic 6	Subtopic 7	~25%	25%-50%	50%-70%	75%~
	Random Guess	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
	AdaBoost	0.20	0.20	0.30	0.30	0.30	0.15	0.20	0.50	<u>0.55</u>	**0.60**	<u>0.55</u>
	SVM	0.35	0.40	0.35	0.35	0.35	0.40	0.40	**0.70**	<u>0.65</u>	**0.70**	**0.70**
	kNN	0.35	0.25	0.35	0.45	0.45	0.60	0.25	<u>0.65</u>	**0.70**	0.55	0.65
	Random Forest	0.20	0.25	0.35	0.35	0.4	0.35	0.20	**0.60**	0.45	**0.60**	**0.60**
Cumulative		Subtopic 1	Subtopic 1-2	Subtopic 1-3	Subtopic 1-4	Subtopic 1-5	Subtopic 1-6	Subtopic 1-7	~25%	~50%	~75%	~100%
	Random Guess	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
	AdaBoost	0.20	0.40	0.10	0.40	0.10	0.20	0.35	<u>0.50</u>	**0.55**	<u>0.50</u>	**0.55**
	SVM	0.35	0.40	0.40	0.40	0.40	0.40	0.40	**0.70**	<u>0.65</u>	**0.70**	**0.70**
	kNN	0.35	0.40	0.30	0.35	0.30	0.35	0.35	<u>0.65</u>	<u>0.60</u>	0.65	0.65
	Random Forest	0.30	0.35	0.25	0.25	0.30	0.30	0.30	<u>0.60</u>	**0.65**	0.55	0.55

The best and the second best performances are **bolded** and <u>underlined</u>, respectively.

(i.e., Persona B, Persona D), which indicates low learning performances. Another cluster was with the low learning experiences and social presence (Persona F), which suggests a potential to improve system perceptions and the following knowledge gain improvements with the future feedback loop implementation. We made the task to classify those three personas from others (Persona A, Persona C, Persona E), which have been combined as one label in the training process.

A four-class persona prediction is an economical approach to classify learners with learning needs, compared to the six-level persona prediction for all learner personas. As seen from Table 5.7, the time duration-based model has achieved the best accuracy both with learners' instant and cumulative behavioral data points. The result shows the highest classification accuracy of 70% via the SVM, kNN, and Random Forest classifiers with the instant behavior data and the SVM with the cumulative behavior data. It is a considerable improvement from the 6-class persona prediction of 45% as the best prediction result and observers' random guesses, which has an accuracy of 25%. The SVM classifier has shown relatively stable and robust performances in both instant and cumulative data in the time duration-based method, proving the most appropriate classifier for real-time feedback loop development. Model training on instant behavior data has shown generally higher accuracy than training on cumulative data.

Once the model is implemented as part of the real-time feedback loop, the time duration-based model using the SVM model based on both instant and cumulative video samples can work as a stable and robust method among all combinations from attempted cases, achieving the lowest 65% and the highest 70% accuracy.

5.6 Limitations and future work

5.6.1 Feature engineering still requires high-level human judgments

As revealed in the feature selection process, evaluating multi-dimensional learner features requires a deep understanding of the domain and the specific dataset. Especially learning analytics and cluster classification greatly depend on how feature vectors and structures are designed. Therefore, feature engineering for different learning domains and tasks in future employment requires expertise with a deep understanding of the field and the data. It emphasizes the importance of more iterative and context-based data collection and learning analytics in a loop.

5.6.2 Combining expert annotation and k-means clustering might provide more valuable insights

The k-means clustering method first chooses a random point and forms a cluster from that point until the last sample. Thus, the quality of the randomly-chosen first data point affects the clustering result, which might affect subsequent statistical analysis results. To overcome such methodological limitations, we suggest involving experts in deciding the centroids of each cluster for k-means clustering. By specifying the centroids rather than starting from random data points, the model can significantly reduce the possibility of selecting an outlier as the first centroid point and having misleading clusters that do not appropriately represent the learner groups.

5.6.3 Feedback implementation for different cluster needs remains a challenge

We aimed at the data-driven persona development to build a foundation for a feedback loop in e-reading. Though we built up an architecture for automatic cluster generation, analysis, and persona prediction based on learners' behavior labels, we still need to implement specific interventions for personas at needs and close the feedback loop. Thus, intervention design and implementation in e-reading is our next research focus for the multimodal feedback "loop" design in e-reading.

5.7 Conclusion

In this work, we implemented a framework of data-driven persona to a multimodal SKEP dataset, which contains various data layers that reflect learners' attention and perception of their e-reading with feedback from conversational agents. We clustered learners based on their knowledge gain, perceived learning experience, and social presence using various unsupervised learning methods to find the feedback necessities of different learner segments. The Chi-square test has compared and validated machine-generated personas from different modeling methods. In the process, feature selection methods (e.g., manual, automatic) and different hyperparameters have been compared. We conducted the statistical quartile analysis on each cluster based on clusters derived from the k-means clustering method. We extracted each cluster's archetypes that make the cluster distinctive from each other and defined six personas. Furthermore, learners' different attention regulation behaviors were used to predict learner personas. In the process, diverse data points, such as instant and cumulative learner behavior labels, have been explored as one dimension while having the

learning phase and time duration as another. Various classical classification models, such as AdaBoost, SVM, kNN, and Random Forest, have been applied to perform the 6-level and 4-level classification tasks. The result indicates that 4-level classification for finding personas with feedback needs, achieving 65-70% accuracy based on the SVM classifier on the time duration sampling method, showing the potential for the real-time feedback loop design. Overall, we aimed to build the architecture for further feedback prompts in e-reading. Our automatic data-driven persona development and prediction can contribute as a practical and effective learning analytics tool for real-time intervention design, greatly assisting researchers and instructional designers in the field.

5

6

Feedback Design Strategies: The Impact of Conversational Agents and Empathetic & Metacognitive Feedback

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.

This chapter is partly based on 📄 *Y. Lee., M. Specht. Can We Empower Attentive E-reading with a Social Robot? An Introductory Study with a Novel Multimodal Dataset and Deep Learning Approaches, 13th International Learning Analytics and Knowledge Conference (LAK)'22* [31].

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 look into self-regulation support for learners in online environments [22, 241]. 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 [139]. Additionally, young readers in the previous years have suffered from attention span reduction by using social media and primarily video-based content [139]. On the one hand, low attention of learners in e-reading leads to less effective and efficient learning [83]. On the other hand, it can also form a negative loop resulting in learners losing interest and engaging less in reading activities [242]. 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 [188]). Educational support has been implemented for various learning objectives (e.g., vocational training [243]) and different target groups (e.g., elementary school students [244]), taking different roles in the educational dialogue as educators, co-learner, and companions [245] in and outside the classroom [246].

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 gestures. Educators' feedback with empathy and meta-cognition prompts have been directly related to learners' cognitive, affective, and behavioral development in learning, leading to positive experiences and effective learning outcomes [247, 248]. 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 [12, 13]. 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 learning outcomes, perceived experience, and perceived social presence of the social robot be predicted through the self-regulatory behaviors of learners?

---

[1] https://furhatrobotics.com/

- How can we develop a data-driven system to automatically conduct an attention analysis in e-reading by intertwining multimodal data streams?

6.1 Background and related works
6.1.1 Attention theories and indicators
Human attention has been defined and interpreted diversely at an intersection of education, psychology, neuroscience, and affective computing. [249] 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 [250], form different internal associations [251], and affect the levels of working memory involved [252]. [253] 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 [83]. [254] 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 [255], 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, knowledge gain, perceived interaction experience, and perceived social presence of the HRI.

6.1.2 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 [210]. 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 the learning practice, and 4) *perceived social presence* of a social robot as a learning companion.

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 [256]. [241] 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 dependence on learners' perception and awareness [257]. 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 [241]. For instance, diverse parameters from the eyes, such as pupil diameter [233], fixations [14], and the number of blinks [234] have been investigated as cues of attention with the implementation of dedicated eye trackers and computer-vision approaches. Learner emotion and arousal, which are known to be critical elements for attention changes, have been interpreted through facial expression changes combined with various data points [235]. Gestural cues from the hands and body have been studied for individual, and group level attention [258]. In this work, we implement a framework of [38] 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 design. Model building for attention regulation behavior recognition could also help to develop real-time loops for further research.

Knowledge gain
Perceived interaction experience
Attention span is known to be highly associated with the motivations, and emotional arousal of learners [259]. 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 [260]. In this regard, the concept of User Experience (UX) and interaction experience [261] has often been adopted to understand learners' emotions, beliefs, preferences, perceptions, and accomplishments and applied to HRI and social robot evaluation, too [262]. 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 [263]. [264] 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) [76], while Attrakdiff measurement [77] 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* from [77] in our study since it has been a measurement developed especially for evaluating the interaction quality and focused more on users' affective perceptions, which is our focus of interaction experience analysis.

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 [222]. Furthermore, it is known to encourage the cognitive actions of learners, and their critical thinking in learning processes [223]. 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 [254], perception of the robot appearance [224], rapport building, and relationship dynamics [225]. Immersion, parasocial interaction, parasocial relationships, physiological responses, social reality, and general social richness have been found as crucial factors of media as presence [226], while it has been explicitly applied as a measurement for robot interaction in comparison with animated characters as social presence. The framework of Social Presence [265] 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 [228] the iCat, a companion robot for chess play. 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.

6.1.3 Behavior-based attention prediction

To our best knowledge, very little behavior-based attention prediction research has been conducted in e-reading. [140] developed an attention prediction model in e-reading based on multimodal cues, such as eyebrow, lip, head movements, and mouse orientation. [20] 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). [38] 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* to identify learning behavior differences in HRI. As we found that behavior patterns and analysis should vary based on a specific scenario [95], we collect 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 a comprehensive analysis of HRI for e-reading.

• 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). This webcam-based approach makes our work easy to reproduce and applicable to diverse reading-based e-learning scenarios and can further be used to design and assess feedback.

6.2 A NOVEL DATASET FOR HRI-based E-reading ANALYTICS
6.2.1 Apparatus
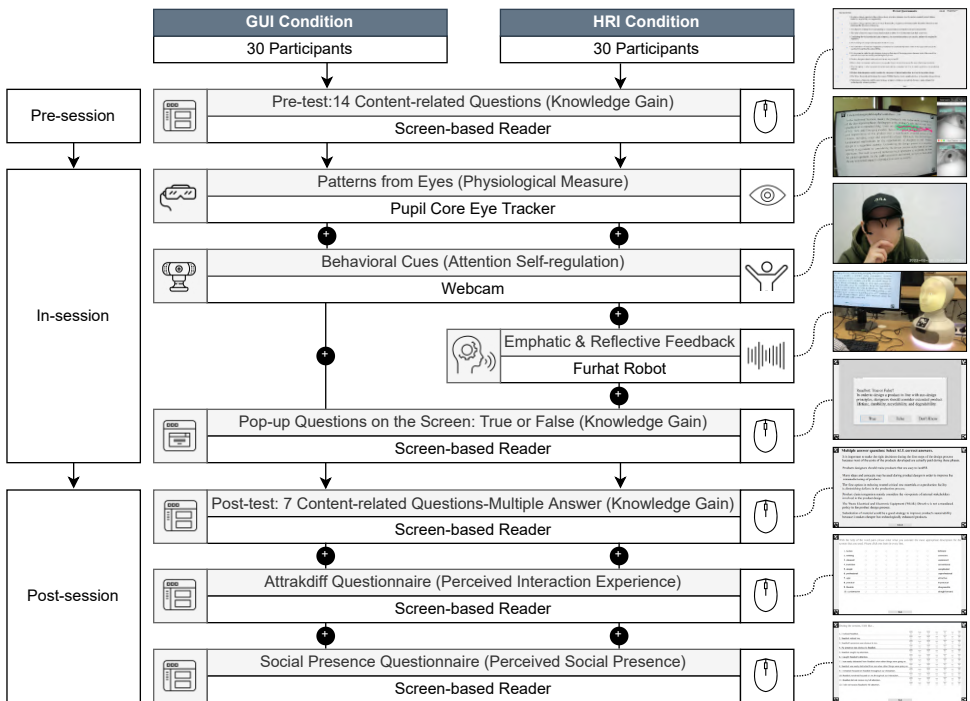
Figure 6.1: Overview of the procedural steps of the GUI and HRI conditions.

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] with two infrared cameras and one head-mounted camera and Logitech C505 HD Webcam[3], that were implemented. For both conditions, an informative

---

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

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 contained 4,750 words and has been divided into 29 pages that cover seven subtopics. The text has been implemented with 47pt on a 27-inch monitor, having 2560×1440 resolution. The setting has been optimized for the eye tracker implementation, which requires a bigger font size than the usual PDF readers for high-resolution data collection. See Figure 6.1 for a procedural summary of two experimental settings.
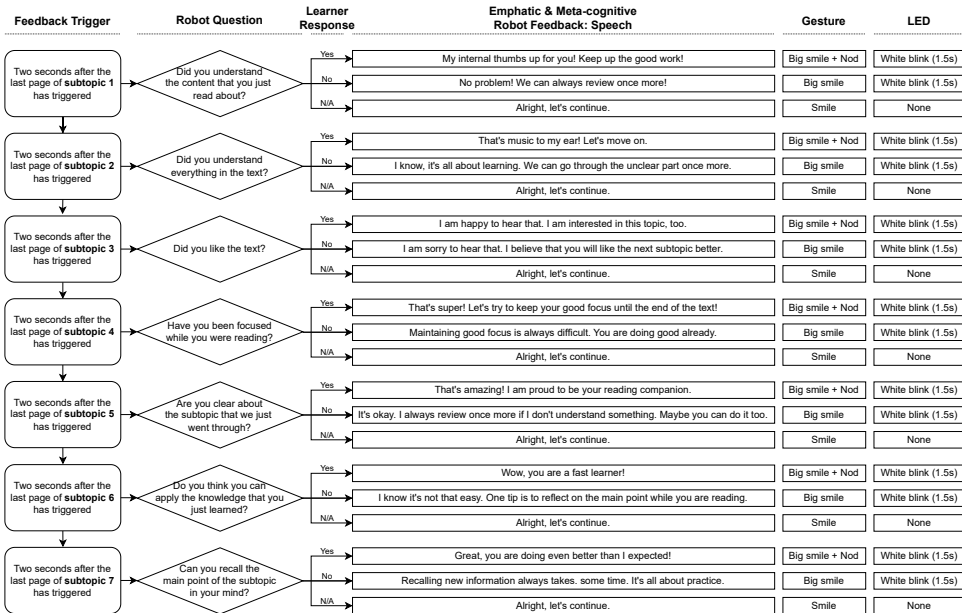
Figure 6.2: Emphatic & meta-cognitive HRI feedback protocol.

6.2.2 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.

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 [38] 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.

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 without contemplating. We followed the formula below to reduce the complication in calculating *knowledge gain*

$$Score_{pre} = \sum_{i=1}^{Npre} S_i^{pre}, \tag{6.1}$$

$$Score_{post} = \sum_{i=1}^{Nin} S_i^{in} + \sum_{i=1}^{Npost} S_i^{post}, \tag{6.2}$$

$$KnowledgeGain = Score_{post} - Score_{pre}, \tag{6.3}$$

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.

Perceived interaction experience

Attrakdiff measurement [77] 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).

Perceived social presence

Social presence measurement [265] 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 by each other between the user and the interface.

6.2.3 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 6.2) was given two seconds after the last page of each subtopic was triggered, only in the HRI condition. Once the reading session had finished, 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.

6.2.4 Dataset construction

Table 6.1: Summary of our novel attention self-regulation, knowledge gain, perceived interaction experience, and perceived social presence with HRI in 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

As illustrated in Table 6.1, 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.

6.2.5 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 6.3 for an overview of the data processing and annotation criteria.
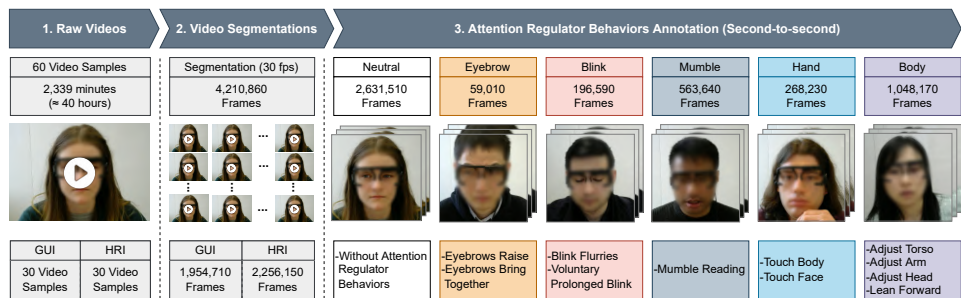
Figure 6.3: Data processing and annotation criteria.

6.3 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 sub-dimensions has been derived to get the overall Attrakdiff and Social Presence values. Furthermore, a one-way ANOVA (Welch's) analysis has been conducted to find the statistically significant differences between GUI and HRI conditions.

Table 6.2: Attention self-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 6.3: 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 6.4: 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 6.5: 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

6.3.1 Attention self-regulation

We labeled five attention regulation behaviors, which are sound indicators of learners' perceived distractions [38], 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 6.2, 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, frequent usage of particular behaviors, and significant behaviors as attentional cues, have been derived more from individual differences than conditions. In this regard, further model training does not differentiate attention regulation behavior labels by experimental conditions but combines both conditions as a whole to achieve attention regulation behavior recognition and further predict other attentional cues.

6.3.2 Knowledge gain

Table 6.3 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.

6.3.3 Perceived interaction experience

Overall Attrakdiff. As seen from Table 6.4, 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). How-

ever, our ANOVA analysis has shown a significance only in Pragmatic Quality measurement between two conditions.

Pragmatic Quality.   Table 6.4 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 has 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 measurement.

6.3.4 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 6.5). 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.

6.4 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 implement vision-based behavior recognition of the subjects with computer vision technologies. In the latter stage of high-level analysis, we utilize recognized subjects' behaviors as feature vectors to achieve the attentive e-reading analysis with machine learning models in a holistic way.

6.4.1 Recognizing attention regulation behaviors with computer vision techniques

Recent years, the deep learning and computer vision fields have made remarkable achievements in various vision tasks [266]. Inspired by those powerful AI models, we try to leverage them to enhance the HRI-based attentive e-reading. More precisely, we implement 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 and a testing group of 20 subjects. The training and testing sets have 94,519 and 45,843 samples, respectively. We use the six classes of annotated attention regulation behaviors as the ground truth to train and evaluate the models' performances. In Table 6.6, we present the performances of these baseline networks.

As listed in Table 6.6, our observations are listed as follows: 1) the best methods' accuracy can go 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 the highest performance 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

Table 6.6: 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	72.84%
	10	112	55.88%
		224	65.89%

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.

6.4.2 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 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 attentive patterns as shown in Table 6.7, 6.8 and 6.9.

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 regulator behaviors that happened within a given attention span as feature vectors with dimensions of $1 \times N$. N is the number of attention regulator behaviors, as six in practice. Then, we fed the feature vectors to classifiers to predict learners' knowledge gain. We present two evaluating 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 6.7, for the coarse(3-level) knowledge gain prediction, all the classifiers can achieve encouraging results (above 63.57% accuracy) and relatively lower accuracy (around 40%) on challenging fine-grained knowledge gain prediction, with the SVM classifier of the highest accuracy for both fine-grained and coarse 45.0% and 74.29%, respectively.

Table 6.7: 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**

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 (Attrakdiff overall and sub-dimensions > 4), neutral (Attrakdiff overall and sub-dimensions = 4), and negative (Attrakdiff overall and sub-dimensions < 4) based on the medium scale "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 [183] with the percentile of participants' scores (below 25%, 25-75%, and above 75%). As seen from Table 6.8, Random Forest provides the best performance for all sub-dimensions of Attrakdiff measurement, scoring the highest performance in the definitive score for the Pragmatic Quality prediction with 92.5% of accuracy. The best accuracy lies on the Hedonic-I prediction with 87.5%.

Table 6.8: 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

Perceived social presence prediction

Perceived social presence prediction has followed the protocol of perceived interaction experience prediction: using 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 6.9 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 for the Perceived Behavioral Interdependence measurement (PBI) prediction. 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 can go 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.

Table 6.9: Perceived social presence measurement prediction using attention regulation behaviors as a predictor. 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

Implementation details. In the above models, we set the following architecture hyper-parameters: CNN: ImageNet-pre-trained [267] InceptionV3 [128] 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 6.6. We used Tensorflow/2.8 platform [4] for deploying the deep learning models and scikit-learn Python [5] for machine learning models.

6.5 Conclusion

We comprehensively investigated the effect of social robots in e-reading by collecting a 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 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 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.

---

[4] https://www.tensorflow.org/
[5] https://scikit-learn.org/stable/

7

Designing Feedback Timing: Deep Learning-Based Attention Regulation Recognition and Real-Time Feedback Loop

This study is built upon a behavior-based framework for real-time attention evaluation of higher education learners in e-reading. Significant challenges in AI model developments for learning analytics have been 1) defining valid indicators and 2) connecting the analytics results to interventions, balancing the generalization and personalization needs. To address this, we utilized a public multimodal WEDAR dataset and trained a neural network model based on real-time features of learners, aiming at predicting learners' moment-to-moment distractions. Real-time features for model training include 30 learners' attention regulation behaviors annotated every second, reaction times to blur stimuli, and page numbers indicating various reading phases. Our preliminary model based on a neural network has achieved 66.26% accuracy in predicting self-reported distractions. Based on the model, we suggest a framework of a Behavior-based Feedback Loop for Attentive e-reading (BFLAe). It has text blur as feedback, a mechanism responsive to learners' distractions that also works as data for next-round feedback. The general feedback implementation rules are established on a statistical analysis conducted on all learners. In addition, we propose a strategy for personalizing feedback using a quartile analysis of individual data, promoting learner-specific feedback. Our framework addresses the high demand for an automated e-learning assistant with non-intrusive data collection based on real-world settings and intuitive feedback provision. The feedback system aims to help learners with longer attention spans and less frequent distractions, leading to more engaging e-reading.

With recent quantitative and qualitative growth in data and computing availability, machine learning approaches are becoming more prevalent in learning analytics and educational data mining [159]. Behavior-based learning analytics is one approach that utilizes cameras and wearable sensors (e.g., eye tracker [140, 158]) to investigate human needs and necessities from their lifestyle, habits, abnormal patterns, and conditions [269]. In learning analytics, machine learning models are often used to predict learning performances and specific internal states of learners from their affective (e.g., arousal, valence [20, 140]) and cognitive states (e.g., mind-wandering [14, 142], switches of internal thoughts [15]) that are associated with learners' performances and experiences. These approaches are applied to individual-level and group levels [22, 258] for various learning scenarios. Based on real-time action recognition and assessment, most systems aim to form an intervention loop and fundamentally aid learning [26, 189].

Regardless of their accurate prediction capabilities, sensor-based approaches are often criticized for being intrusive [26], changing the nature of learning experiences. Thus, various computer vision-based approaches [15, 23] have been suggested to make learning and system design more seamless for real-world applications. Especially behavior-based analytics is valuable in that particular behavior that machines recognize is also observable and semantically interpretable to humans to some extent [160, 256]. Common challenges in behavior-based machine learning applications in learning analytics have been 1) to find valuable features for model training [160] and 2) to specify the implementation conditions and parameters that best support the accurate recognition of targeted signals [32]. 3) Also, closing the feedback loop, considering generalization and personalization [26] in the analytics phases, and implementing the feedback has been difficult.

In this regard, our objective is to suggest a Behavior-Based Feedback Loop for Attentive e-reading (BFLAe) framework, which involves 1) webcam-based video data collection, 2) computer vision-based learning analytics, 3) blur feedback implementation in text, and 4) further cognitive&behavioral changes of learners as consequences of feedback loop implementation. The framework is built upon a multimodal WEDAR dataset, which provides valuable insight into learners' behavior during e-reading activities. Our approach involves training a neural network model on real-time features that reflect learner behavior, including attention regulation behaviors, reaction times to blur stimuli, and page numbers that reflect different reading phases from the public WEDAR dataset [141]. These features provide a basis for predicting learners' perceived distractions and form a foundation for implementing feedback mechanisms. By implementing the blur feedback on the screen-based e-reader, we aimed to close the feedback loop that enables the further loops, which is not obstructive to the primary reading task and is semantically intuitive. Feedback could potentially help learners reflect on their current state and strategize for future reading [159], which may not be subjectively noticeable to them. The objectives of the behavior-based real-time feedback loop have been 1) extending the overall attention span of learners and 2) reducing the frequency of distractions.

We believe that this personalized, behavior-based feedback loop offers a practical solution to the challenges faced by the fields of Technology-Enhanced Learning (TEL) and Multimodal Learning Analytics (MMLA), promoting more engaging, effective, and individually tailored learning experiences [270]. This article contributes to the ongoing discussion of how best to use technology and learning analytics to support learners. By presenting an innovative

framework for an attention regulation behavior-based feedback loop in e-reading, we hope to inspire further research and practical applications of behavior-based models in education.

Our contributions are as stated follows:

- According to our best knowledge, it is the first framework to introduce a real-time feedback loop for attentive e-reading. Our webcam-based behavioral framework is non-obstructive and applicable to diverse e-learning scenarios which involve e-reading as a major learning activity. Our BFLAe framework with increasing digital reading in formal and informal learning with prevalent digital technologies will be more valuable.

- It is a framework built upon WEDAR, a multimodal public dataset collected in an e-reading scenario. It offers more relevant data specified for attention measurement for e-reading. With the implementation details depicted in our framework, the work can be reproduced and further elaborated for specific scenarios based on different tasks and implementation requirements.

- By specifying the statistical values of different behavior labels that represent attentive (i.e., neutral) and distractive (i.e., attention regulation behaviors) learner states, we provide researchers and instructional designers with options to make choices on thresholds for the feedback trigger. As feedback necessities vary depending on the system goals, our analysis result can provide valuable ground for the feedback rules for different systems.

7.1 Behavior-based Analysis on Multimodal WEDAR dataset

In this section, we briefly analyze the multimodal WEDAR dataset. By doing so, we tried to understand the dataset's structure and attention regulation behaviors shown in e-reading and potential patterns that are shown together with the self-reported distractions.

7.1.1 Preliminary analysis on attention regulation behaviors

We used the multimodal WEDAR dataset in our investigation [141]. This dataset comprises human-labeled behavioral labels with five categories of attention regulation behaviors and a neutral behavior as the label, all annotated in every second of the video data. These videos were collected from 30 higher education learners. In particular, this study used real-time distraction reports as the ground truth for distraction instances [118]. As depicted in Figure 7.1, the distribution of attention regulation behaviors in the dataset is not even. The most common behaviors are body movements, which account for 18.5% of the behaviors, and hand movements, which contribute 12. 1% to the duration of the video. The remainder consists of eyebrow movements (3.1%), mumbling (2.6%), and blinking (2.1%). Furthermore, neutral labels, indicating states of attention, constitute 90.9% of the behavioral labels. It is important to note that multiple attention regulation behaviors can co-occur within the same second, so the total proportions do not add up to 100%.
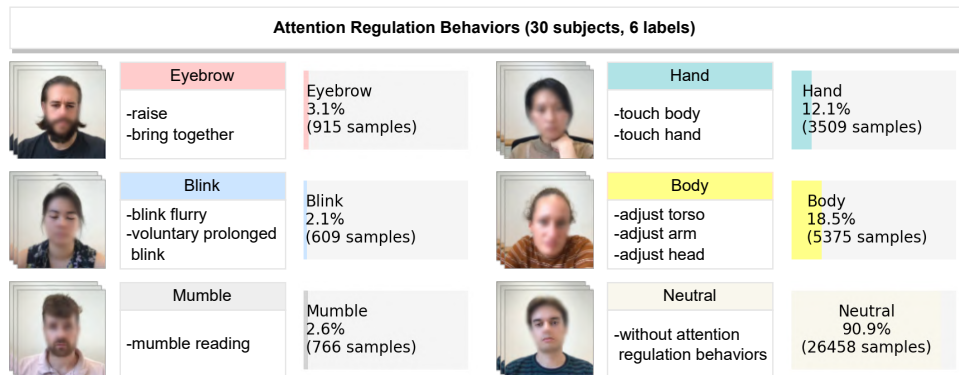
Figure 7.1: The multimodal WEDAR dataset contains second-to-second annotation labels of attention regulation behaviors and neutral behavior, consisting of varied label proportions.

7.1.2 Unobservable patterns between attention regulation behaviors and self-reported distractions

We graphically represented the five categories of attention regulation behaviors and neutral behaviors along with distraction reports to discern potential visual patterns between attention regulation behaviors and self-reported distractions. As is evident in Figure 7.2, participants exhibited a wide range of reading speeds, ranging from 461 seconds (7.7 minutes) to 1661 seconds (or 27.7 minutes). Moreover, we noticed substantial variation in the use of attention regulation behaviors, as well as in the patterns of perceived distractions and the reporting of these distractions. Given this unobservability, the integration of machine learning becomes crucial. It also represents the limitations of human educators in detecting complex patterns hidden within the behavioral patterns of learners.
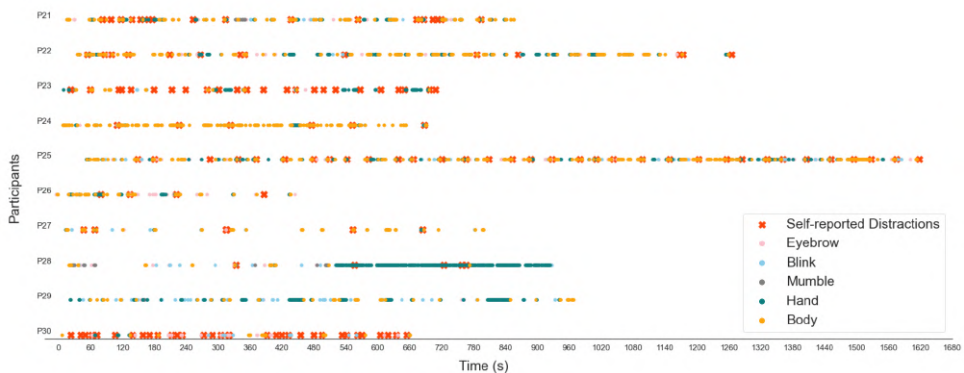
Figure 7.2: Self-reported distractions and five attention regulation behaviors visualized in time for one-third of all participants (P21 - P30)

7.2 Framework of Behavior-based Feedback Loop for Attentive E-reading (BFLAe) and its architecture
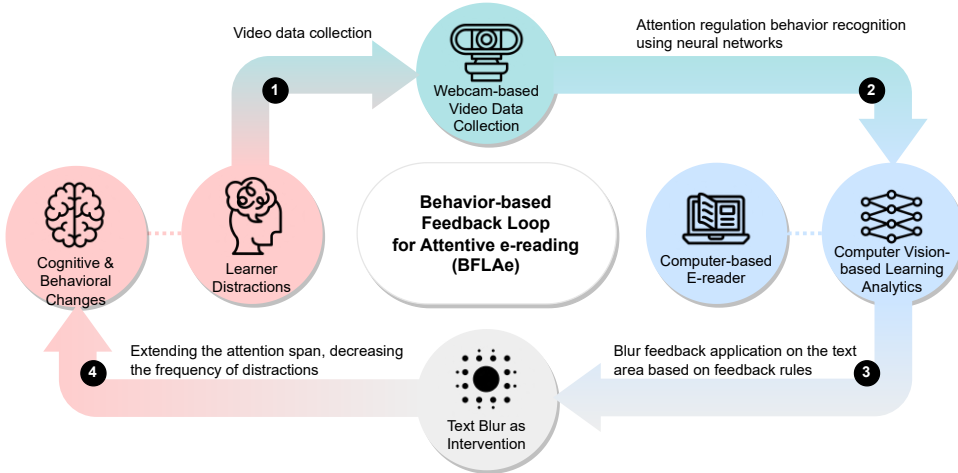
Figure 7.3: The overall architecture of the Behavior-based Feedback Loop for Attentive e-reading (BFLAe) framework includes four stages: 1) webcam-based video data collection, 2) computer vision-based learning analytics, 3) text blur as intervention, and 4) cognitive&behavioral changes aimed by the feedback.

This section presents the system's architecture, as shown in Figure 7.3. Drawing on previous research in the realm of multimodal learning analytics [26, 32], critical factors in forming a multimodal feedback loop for learning include 1) the alignment and integration of data streams, 2) the identification of learning requirements, 3) informed design decisions for multimodal feedback, and 4) the observation of implications within specific learning scenarios. Consequently, we propose a four-stage approach to BFLAe.

7.2.1 Framework of BFLAe: four stages in system architecture

In the first stage, webcam-based video data is collected during e-reading. This method offers an unobtrusive approach compared to other sensor-based strategies. The second stage involves learning analytics, which is based on a model developed from attention regulation behaviors and self-reported distractions. The following section will detail the specific features used in model training and the rules for triggering system feedback. In the third stage, a blur effect is applied to the reader's screen for the feedback generation condition, which was decided in the previous phase. The blur effect can be deactivated by the learner clicking on the reading area. This stage not only aids learners by increasing arousal but also serves as additional data for further learning analytics since the reaction time provides crucial cues about the learners' cognitive states. The final stage of the loop aims to induce cognitive and behavioral changes in learners. Specifically, the system's objectives are: 1) extending the attention span between distractions and 2) decreasing the frequencies of distractions, as measured by attention regulation behaviors, reaction speed to the blur stimuli, and self-reported distractions.

7.3 Behavior-based attention predictions based on Neural Network

This section introduces the features and computational model that we have established to predict attention levels: a prerequisite step integral to the subsequent feedback generation.

7.3.1 Feature engineering of real-time features

The WEDAR dataset provides behavioral attributes in real-time from 30 higher education learners engaged in e-reading. As referenced in Table 7.1, eight distinctive features have been harnessed for model training. Five attention regulation behaviors were used as binary features (feature 1) and independent features (features 2-6). Reaction times to secondary blur stimuli, activated at random intervals, have been implemented as another feature (feature 7). Reaction time is a classical measure used to assess learners' arousal levels [15, 175]: shorter reaction time is often interpreted as higher arousal, while a longer reaction time is often considered an indicator of more distractions. The last feature is the specific page number (ranging from 1 to 10) that the learners were on, which represents the reading phases of the learners. For feature engineering, this data was one-hot-encoded (feature 8). It is important to note that we have only extracted real-time features from the dataset. This decision aligns with the feedback loop's objective of a real-time approach.

Table 7.1: Real-time features have been pre-processed from the multimodal WEDAR Dataset.

#	Feature name	Feature description	Categorical / Nominal
1	Attention_regulation_behavior_binary	Occurrences of any of attention regulation behaviors	0,1
2	Eyebrow_occurence	Occurrences of movements from eyebrow as attention regulation behavior	0,1
3	Blink_occurence	Occurrences of movements from blink as attention regulation behavior	0,1
4	Mumble_occurence	Occurrences of movements from mumble as attention regulation behavior	0,1
5	Hand_occurence	Occurrences of movements from hand as attention regulation behavior	0,1
6	Body_occerence	Occurrences of movements from hand as attention regulation behavior	0,1
7	Reaction_time	Reaction time to randomly triggered blur stimuli	Continuous
8	Page_number (one hot encoded)	The page number that learners are currently on	1,2,3,4,5,6,7,8,9,10

7.3.2 Data pre-processing

We utilized eight real-time features described in Table 7.1 for our model training. We initially partitioned our dataset into training and testing sets, comprising 80% and 20% of the data, respectively. We balanced the data set, using the synthetic minority oversampling TEchnique (SMOTE) to prevent an imbalance between distracted and attentive states so that neither state would dominate the other in proportion and provide sufficient data points for the training. Subsequently, we applied min-max normalization to confine the data distribution between 0 and 1. This process was implemented to mitigate any potential bias from different data ranges. Furthermore, min-max normalization is acknowledged for its ability to accelerate training. It is particularly advantageous for our approach, which will have many data points from second-to-second recognition.

7.3.3 Model training using neural network

As shown in Figure 7.4, we employ a sequential neural network model with its linear stack of layers. Our network architecture comprises three hidden layers with a rectified linear unit (ReLU) activation function. To mitigate the risk of overfitting, we incorporated a dropout layer into our model, which is widely used for randomly nullifying a fraction of

the layer's output features during the training phase. In our case, the dropout layer is configured with a rate of 20%, omitting one-fifth of the input. The final layer of our model is a dense layer with a Sigmoid activation function, with an output range between 0 and 1. It is an optimal choice for our binary classification task. The loss function is designated as mean squared error (mse), the optimization algorithm is set as Adam, and the accuracy is selected as the metric for model evaluation during training. The model has reached an accuracy of 66.26%. This performance exceeds the 50.00% accuracy expected from random guess, which implies that the prediction capacity of the model is considerably better than the chance. The real-world implementation could be enhanced by integrating the feedback rules, which will be further elaborated on in the next section.
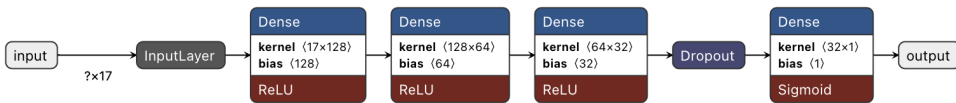
Figure 7.4: Our model structure, built upon a neural network, consists of one input layer, three hidden layers, one dropout layer, and one output layer.

7.4 Automatic feedback constructs with visual stimuli

This section introduces the rationale for implementing blur stimuli, feedback rules, and Human-Computer Interaction (HCI) architecture (Figure 7.5). See Figure 7.4 for descriptions of HCI, showing the functions of components and blur feedback applied in response to learners' distractions.

7.4.1 Type of feedback: blur stimuli

We suggest the implementation of blur on text area as automatic visual feedback (see Figure 7.6), which has also been used to measure reaction time in previous studies [15, 156]. In the following, we introduce the advantages of introducing blur stimuli as part of a feedback loop.

1) The blur stimuli serve a dual function: they trigger the learner's arousal and simultaneously work as data points for future feedback loops. Different reaction times, behavioral features, and self-distraction reports are incorporated into the screen-based reader as next-round feedback, enabling more precise predictions and personalized feedback.

2) Critics often suggest that feedback interrupts the primary task by adding secondary tasks to learners, inducing cognitive overload [158]. In this context, the interaction between the learner and the system is semantically intuitive and actionable by having a prominently placed deactivation button, where the learners naturally focus during the reading task.

7.4.2 Feedback implementation rules: statistical analysis on learner behaviors indicating different attentional states

The window size in machine learning refers to the number of data points that are considered to capture information and contexts at each step, which is especially crucial for sequential data processing [269]. We propose tailoring different window sizes to different attention regulation behaviors to enhance the prediction of self-reported distractions. As evidenced
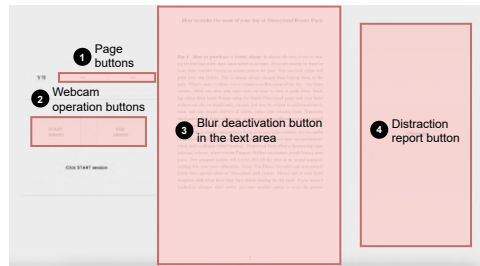
Figure 7.5: HCI components and functions: page, webcam operation, blur deactivation, and distraction report buttons.
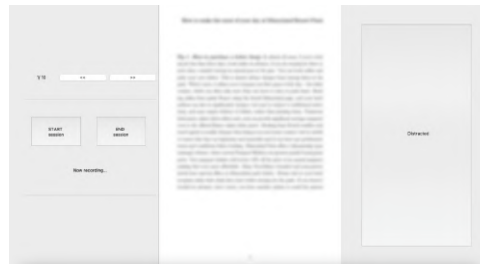
Figure 7.6: Blur feedback is applied to the text area as an intervention triggered by recognized distractions.

in Table 7.2, derived from the WEDAR dataset, the minimum, maximum, average, median, standard deviations, and quartiles of behaviors exhibit variability of the duration of each state. The current distraction prediction model was designed based on second-to-second labeling for all attention regulation behaviors. However, incorporating different behaviors and applying a range of sliding windows could potentially improve the accuracy of the learners' distraction predictions.

Table 7.2: Statistical analysis conducted on durations of each behavior label, collected from 30 participants.

[1]Behavior labels have been annotated second-to-second, making the minimum, maximum, median, Q1, Q2, and Q3 values integers.

Durations (s)[1]									
Attentional States	Behavior Labels	Min	Max	Mean	Median	SD	Q1	Q2	Q3
Attention	Neutral	1.0	124.0	9.44	5.0	12.73	2.0	5.0	12.0
	Eyebrow	1.0	5.0	1.20	1.0	0.52	1.0	1.0	1.0
	Blink	1.0	5.0	1.14	1.0	0.45	1.0	1.0	1.0
Distraction	Mumble	1.0	35.0	3.15	2.0	5.11	1.0	2.0	3.0
	Hand	1.0	62.0	3.58	2.0	4.88	1.0	2.0	4.0
	Body	1.0	44.0	3.20	2.0	2.85	1.0	2.0	4.0

The system's feedback mechanisms can be varied according to its specified objectives. For example, some may apply a window size spanning the third quartile to maximum values of specific behavior for attention prediction. On the contrary, those who require stricter self-regulation among learners may opt to utilize a window size between medium and maximum values for the same task. By establishing specific ranges that act as a foundation for feedback implementation, researchers and educational practitioners will benefit from devising their intervention rules, drawing on general learning behavior. Please note that our analysis has been performed on the WEDAR dataset. Thus, the predefined ranges

may undergo further refinement with the accumulation of additional sample data in future studies.

7.4.3 Considerations for Feedback Personalization: Quartile analysis in individual data

The creation of personalized models can be facilitated by conducting quartile analysis in individual data, considering individual differences in relation to their own unique behavioral status [183]. Quartile analysis offers a way to position specific learners within the broader learner population by distinguishing the first (0% to 25%), second (25% to 75%), and third (75% to 100%) quartiles. This study recommends applying quartile analysis to individual datasets for evaluating learner behaviors and performance. For example, in assessing the reaction time to blur stimuli, each reaction of a single individual can be classified as a fast (1st quartile), medium (2nd quartile), or slow (3rd quartile) response. These categories can also be correlated with high, medium, and low arousal states. Through the accumulation of such data as model features, we can enable the provision of more precise and personalized predictions and feedback provision.

7.5 Conclusion

We propose a framework of behavior-based feedback loops for attentive e-reading. As established in previous research, the challenge of closing the feedback loop has been a recurring issue in the fields of TEL and MMLA. We leverage the multimodal WEDAR dataset in this work, which aids in developing behavior-based predictions of learners' perceived distractions. Real-time features have been extracted to train a neural network that predicts learners' perceived distractions. These features encompass attention regulation behaviors, reaction time to blur stimuli, and reading phases derived from page numbers. Our approach involves the implementation of blur feedback in response to learners' distractions and establishing the foundation for feedback rules based on the statistical attention regulation behavior analysis derived from general data. Simultaneously, we propose a strategy for personalizing the feedback based on a quartile analysis of individual data. Our behavior-based model addresses the emerging need for an e-reader with automatic learning analytics and feedback mechanisms that can be applied to real-world scenarios.

7.6 Discussion and Future Work

Optimizing the window sizes of attention regulation behaviors for accurate distraction prediction   A statistical analysis of learners' data in e-reading has been performed in the current framework. Broad ranges of learners' attention regulation behaviors have been derived, indicating learners' states of attention and distraction. In future work, several ranges of different behavior recognition technologies will be applied and tested. Doing so will provide practical insights into real-time recognition and feedback generation that can best assist our feedback objectives.

Testing the effects of the automated feedback from an intelligent e-reading system
Though the overall behavior-based feedback loop framework has been suggested, the effects of implementing automated feedback still need to be tested: investigating the attention

span and frequencies of distractions. Our intelligent system can be further evaluated for subsequent effects, such as learning outcomes and perceived learning experiences, with various qualitative and quantitative measures. Our next step involves comparing the intelligent feedback loop based on the current BFLAe framework and time-based feedback.

Exploring the effects of feedback types and modalities   In this work, we suggested blur feedback due to its intuitive actionability and less cognitive load than other feedback. However, with the same feedback timing, we still need to validate whether different types and modalities (e.g., speech-based feedback from conversational agents) of feedback provide additional value in learning. We will further test the effects of varying feedback with various types and modalities built into our current attention recognition mechanisms.

8

Real-time AI-based Feedback Loop Implementation and Its Impacts on Learners' Attention Span, Learning Outcomes, and Perceived Learning Experiences

In higher education, real-time intervention for e-reading often remains unimplemented due to technical challenges and misalignment of theoretical frameworks. To address such challenges, we develop a real-time feedback loop to assist attentive e-reading, aligning affective computing, education, and Human-Computer Interaction (HCI) leveraged by AI technologies. The system aims to recognize learners' real-time distractions and intervenes with learners for fewer distractions and longer attention spans in e-reading. We trained on neural networks based on the MediaPipe framework to recognize learners' behavioral cues, named attention regulation behaviors, that are known to correlate with perceived distractions. Screen blur as feedback was triggered based on the hybrid neural networks and thresholds updating every page based on learners' current arousal and distractions, which was leveraged by k-means clustering. We investigate how AI-based real-time feedback can help learners manage attention on behavioral, cognitive, and affective levels. The result shows that the implemented system assists attention management, leading to fewer distractions and longer attention spans for learners. The explainability of AI-based automatic feedback is emphasized for affecting learners' perceptions about the system experience and its subsequent implications on learning outcomes. We further investigate which behavioral components best predict learners' knowledge gain using machine reasoning, such as logistic regression and a decision tree. Our work suggests a practical and empirical foundation for AI-based e-reading support with broad applicability, robust recognition, and feedback adaptation strategies.

Attention management has been a critical challenge for learners who engage in e-learning [156]. Recent drastic changes from traditional classrooms to online and hybrid settings [1] from the COVID-19 pandemic have made learners' attentiveness [158], connectivity, participation, and behavioral engagement [271] in e-learning even more critical. Technological advancements with widespread digital devices and learning platforms have accelerated the trend [2]. In transitions, various theoretical approaches have been made to fill the gap for e-learning from the perspectives of student participation [272], satisfaction [273], educational productivity [274] and learning outcomes [275]. Meanwhile, learners' self-regulation has been emphasized more than ever [10] with learners positioned in such a change regardless of their capability or readiness [1]. Also, real-time learning supports [187] have been scarce despite their importance for learners adapting to the new forms of daily education.

Especially, e-reading holds a unique position in higher education. Higher education necessitates a substantial amount of self-directed reading [3], information processing [4], knowledge comprehension [5], critical thinking [6], and knowledge reproduction and application [7] through reading, all of which are integral to regular higher education and, thus, closely tied to learners' self-efficacy [8, 9], learning effectiveness [1], and academic achievements [10]. However, the transition to e-reading has not been accompanied by a thorough exploration of specific learning strategies or supports [276], leaving learning management largely reliant on individual cognitive processes [277], motivations [278], meta-cognition [279], and self-regulation [280]. While various learning domains (e.g., health sciences [281], foreign language learning [282]) have been considered as target learning scenarios for real-time intervention based on multimodal data streams, the real-time e-reading intervention has yet to be implemented.

Traditionally, learning analytics on e-learning have been taken post-hoc, based on the large-scale log data collected from expansive educational platforms, such as MOOCs and edX [283], having learning analytics results visualized on dashboards. Previous studies sought meaningful insights about learners' decisions, goals, and self-regulation [284]. However, the post-hoc dashboard analysis has shown its limitation as an intervention tool because of its retrospective nature; thus, learners cannot act upon it at the exact moments of intervention needs [150]. Dashboard feedback only becomes meaningful to learners when the next round of education occurs [284], while keeping learners in e-learning for the next iteration of education is a common challenge with a high dropout rate [1]. Also, due to data regulations of e-learning platforms, learning data, primarily log-based, tends to deliver somewhat distant and superficial information (e.g., demographics of learners, number of clicks) about learners and learning [75], having approximately 60% [285] of learning analytics conducted primarily based on a single type of data, which does not support the multidimensional understanding of e-learning.

Sensor-based approaches have emerged in the last decade to address the limitations of log-based data, utilizing multimodal data from various data streams and feeding them to learning analytics [286]. With the current integration with various machine learning techniques, studies have aimed to predict learners' internal states regarding the cognitive load and perceptions during learning [287]. In the process, various learners' states have often been used to predict learning performance and results [288] to find essential recipes for learning success. However, due to the difficulties of integrating diverse data streams

with different modalities, formats, and granularity [150], data processing and training often require expertise and sizable computational resources [289]. Also, applying the real-time approach requires practitioners with an excellent understanding and skills in technological deployment, which lays another layer of the challenge in practice.

Furthermore, another challenge of utilizing the sensor-based approach in education has been its obstructiveness in data collection; having various physical sensor implementations brings obstructiveness in learning and changes the nature of the learning itself [290]. For instance, biosensors, in the form of wearables (e.g., eye tracker [158]), are often implemented to collect physiological information from learners and learning environments. In such cases, data collection changes the learning ecosystems and obstructs education, leading to bias in interpreting learning and learners [286]. In this context, remote detection, leveraged by computer vision, has taken place to mitigate the limitation of traditional sensor-based approaches [290]. Learners' facial expressions and postures, combining other data layers, such as discourse, have been used to evaluate learning on individual and group levels [22]. However, despite these advancements, several challenges persist in developing the feedback loop for e-reading, which we aim to address through our work:

- **Feedback loop alignment:** Although real-time intervention is often a fundamental goal of real-time recognition of learners' states and learning analytics, the real-time feedback loop has rarely been attempted. It is because learning analytics and feedback design are often conducted in isolation by different practitioners; the indicator design, data collection, feedback strategies, system implementation, and evaluation in practices are often not aligned, making the integration of frameworks even more difficult.

- **Generalizing and personalizing the feedback loop:** Balancing the generalization and personalization needs of learning has been a fundamental challenge of experimental indicator design, learning analytics [150], feedback design, and implementation. It is because the system should be general enough to accommodate the learning needs of general learners, while the system should be able to capture the specific needs and assist them as a personalized intervention. Various learning indicators, theories, and feedback strategies should be investigated and implemented in practice to achieve this objective.

- **Closing the "real-time" feedback loop:** Closing the real-time feedback loop entails complicated technological alignments. Since recognition and feedback generation should be run in parallel, algorithms should be designed and optimized for limited computational resources. At the same time, the model should be able to capture learning efficiently, together with the feedback algorithm, in a timely manner, which requires various considerations in deployment and validations over iterations.

In this work, we develop a feedback loop for higher education learners' attentive e-reading to address articulated challenges. We design a hybrid model with robust accuracy that can recognize learners' attention regulation behaviors based on the MediaPipe framework with the skeleton analysis [291]. When learners' attention regulation behaviors, correlated with self-reported distractions [156], are shown for longer than predefined thresholds [292], the screen blurs, gently reminding learners of their distractions. These thresholds

for the screen blur stimuli are updated on each page based on learners' reaction time and behavioral expressiveness, representing learners' arousal and distractions, respectively, adapting to learners based on their intervention needs [184]. Through our implementation, we investigate the effects of AI-based feedback loops on learners' distractions, knowledge gain, and perceived interaction experiences during their e-reading. Our novel contributions are as follows:

- **First real-time implementation of a feedback loop for e-reading:** According to our best knowledge, it is the first attempt to implement a real-time feedback loop designated for e-reading. We develop the AI-based automatic feedback loop based on attention regulation behavior recognition. We investigate the effects of AI-based feedback on learners' attention management from various angles (i.e., distractions, knowledge gain, and perceived interaction experience), which form a foundation for future real-time e-reading support design and implementation.

- **Low implementation requirements from webcam-based skeleton recognition:** Thanks to our skeleton-based recognition, our framework can easily be applied to various e-learning scenarios without environmental constraints or high computational requirements while keeping the robust recognition. Also, our webcam-based approach is not intrusive to the learning activities and thus can be further applied to diverse e-learning scenarios based on reading.

- **Adaptive feedback implementation:** An adaptive feedback strategy has been designed based on the public WEDAR dataset [141] collected in e-reading in higher education, which shares the same setting as our work. Often, adaptation strategies clash with generalization needs for learning analytics and feedback provision. Our approach segments learners based on attention-related behaviors (e.g., arousal, distractions). It recalibrates feedback frequencies on every page based on where they belong on a group level (e.g., highly aroused learners with mid-range distractions represented as C2 in our work), a new approach implemented in our work.

8.1 Related Work
This section articulates three practical questions for our feedback loop design and implementation. Previous works are reviewed from various angles and answered in subsections.

1. When and how to intervene with learners to inform them of their distractions?

2. Which machine learning approaches best suit real-time and robust attention regulation behavior recognition in practice?

3. What adaptation strategies can be implemented in our real-time feedback loop?

8.1.1 When and how to intervene with learners to inform them of their distractions?
This subsection investigates the behavioral framework of attentive e-reading that can be utilized in our feedback loop design. For the feedback design, we study vital principles for designing and implementing real-time multimodal feedback.

Target behaviors: attention regulation behavior recognition and feedback generation based on Self-Regulated Learning (SRL)

In the revised Cyclical phases model adapted from Zimmerman and Moylan (2009) [293], there are iterative phases of forethought, performance, and self-reflection for SRL. In this work, we focus on the performance phase, where metacognitive self-monitoring and self-control take an essential place where learners perform their learning tasks [293]. The framework suggests that proactive task strategies and help-seeking are made as SRL in the performance phase. Lai and Hwang (2020) further scoped the learning activity into e-learning in their model, focusing on the role of systems and technological supports that can actively assist learners with affective performances in e-learning [294].

The framework of attention regulation behavior from Lee et al. (2022) [156] represents observable behaviors as cues of SRL in e-reading. In the framework, attention regulation behaviors are defined as *"learners' earliest self-awareness of attention loss and following observable behavioral changes as self-regulation".* Such behaviors are emphasized because those are the *"moments that learners are willing to and are still able to re-engage in their learning tasks"*, where learners necessitate learning supports, and the intervention can maximize its effects. In the same framework, various movements from eyebrows, blinks, mumble, hand, and body are specified as attention regulation behaviors in e-reading and are proven to be correlated with perceived distractions [156], which is described in the "Measurement" section in our work. For the model development for the behavior recognition and designing feedback adaptation strategies based on k-means clustering, combined with logistic regression, we utilize the public WEDAR dataset. Attention regulation behaviors are used as target behaviors of recognition, while we further leverage the Behavior-Based Feedback Loop for Attentive e-reading (BFLAe) framework [292] for designing the feedback thresholds of each attention regulation behavior, which decides the timing of interventions.

Feedback rules: principles of real-time multimodal learning interventions

Though feedback plays an integral role in learning, the impacts of feedback on individuals are known to vary [295]. Therefore, this subsection discusses various feedback principles and conditions for e-learning interventions that can best support our real-time feedback design.

Cognitive load has been considered a primary element for instructional design, directly connected to the utilization of higher-order cognitive skills [296]. Several rules have been suggested for managing cognitive load in multimedia-based instructional design [297] in ways to prioritize and clarify the use of the multimodal feedback: offloading the cognitive load into split sensory channels, segmenting, and pertaining once sensory channels are working with high demands in working memory, weeding extraneous materials, and directly signaling for problem-solving, eliminating redundancy, and synchronizing the information for representational feedback through multisensory channels [297].

In the same vein, *Non-intrusiveness* to the primary task performance [298] has been emphasized, as interpreting and processing intervention can cause split attention and further hinder the streamline of the task performance [4]. The redundant intervention has also been known to cause split attention, which results in fewer working memory capacities [299].

Timeliness of the feedback has also been mentioned as another critical principle of multimodal feedback. This is because intervening at the wrong time affects not only the distrust

toward the AI-based automatic feedback but also the overall learning experience and subsequent learning outcomes. It has been noted that immediate feedback works better than post-hoc feedback [296], which supports our real-time approach. Timely feedback provision is intertwined with accurate recognition [300] and design in application since the feedback generation in AI-based recognition relies on the accuracy of the implemented model and the model deployment.

8.1.2 Which machine learning approaches best suit real-time and robust attention regulation behavior recognition in practice?

This section explores the effective recognition method with swift processing and robust accuracy for real-time applications.

Model training methods: image/video-based vs. skeleton-based recognition
Previously, behavior-based attention recognition in e-learning has been conducted using image [301] and video-based [156] models, supported by machine learning techniques. Compared with image-based methods, video-based recognition, bolstered by deep learning models, has been known to consider the *temporal layers* in the model, thus capturing learners' actions with movements better [156]. However, video-based methods have shown limitations in real-time implementation due to the complexity of the model and the processing speed, which often exceeds the available computational capacity and practical implementations. In this regard, we explore a skeleton-based method based on MediaPipe[1] [302] to compensate for such a problem to enable the computationally inexpensive real-time feedback loop. Skeleton-based methods analyze visual inputs and use them to identify and track various joints and landmarks of the human body, which enables real-time behavior analysis and multimodal interactions (e.g., behavior-based augmented reality [303]). The model training can be conducted based on the multi-dimensional matrices of arrays extracted from joints as data input, which enables the training with fewer resources. Also, the model can still consider temporal layers by feeding it with data extracted from subsequent frames, which is 30 frames in our work. Below, we articulated multiple advantages of utilizing the skeleton-based attention regulation behavior recognition in closing the real-time feedback loop, which is the rationale behind our model design decision.

- **Low computational requirements**: The skeleton-based method extracts only specific landmarks of the human face and body in the 3D space and can transform that information into the form of multi-dimensional matrices. Such processing from high-dimensional data (i.e., videos) to low-dimensional data (i.e., NumPy arrays) can streamline the model training and its real-time implementation.

- **Accurate behavior recognition with spatial and temporal understandings**: Since the MediaPipe framework has already been built for capturing particular target features (i.e., human poses and actions), it provides a solid foundation for a *spatial* understanding of human behaviors in 3D spaces. At the same time, it can also be used to understand *temporal* information of the movements, which is essential for recognizing attention regulation behaviors.

---

[1]https://developers.google.com/mediapipe

- **Applicability without environmental constraints**: The conventional problem in image and video-based approaches has been that the model can also learn irrelevant visual features of non-targeted objects in training. The problem is more evident with physical environmental changes (e.g., lighting changes, objects in the background) that highly hinder model applications in real-life settings. However, the skeleton-based method can be applied to various environments thanks to its landmark extraction from the human body, making the model highly adaptive to multiple application scenarios and users.

8.1.3 What adaptation strategies can be implemented in our real-time feedback loop?

State-of-the-art feedback personalization strategies in e-learning
As indicated in the review of personalized feedback in e-learning [304], theoretical and empirical feedback strategies for e-learning have yet to be established, and the general effect of adaptive feedback needs to be investigated. The previous approaches to adaptive/personalized feedback design have been done: 1) *rule-based* and 2) *feature-based*. The rule-based approach is based on simple if-then principles, thus more straightforward in the application and interpretation [305]. On the other hand, feature-based feedback, leveraged by AI, operates based on the black box method; thus, the rationale behind AI-based feedback sometimes brings doubts due to its low explainability. The previous personalization approaches in e-learning have been geared toward student performances, motivation & engagement, and SRL based on individual goals, knowledge states, learning progress, learning behaviors, emotional or motivational conditions, and various student traits [304]. The feedback messages delivered through such feedback have been evaluative and informative, and those were applied to university online courses, blended learning, and formative assessment [304]. Feedback personalization in real-time has yet to be attempted in educational intervention design. Especially, adjusting the *timing* of the feedback interval has yet to be used as a feedback adaptation strategy, which we would like to study in our work.

Feedback adaptation strategy: learner segmentation based on behavioral features
Providing personalized real-time feedback, especially to unknown users, is a common challenge since there is often insufficient data to determine their characteristics, goals, skills, and needs [185]. In such cases, user segmentation is an effective method to address such issues. It segments users into groups based on their individual features, such as preferences, behavioral patterns, and demographics, which are critical to discerning different groups with distinctive features [184].
In previous work, Lee et al. (2023) [184] chose critical components essential in understanding user interactions as learners' knowledge gain, perceived interaction experience, and perceived social presence with the feedback system and clustered learners based on k-means clustering. It effectively addressed the concerns of balancing generalization and personalization needs in feedback design. From the instructional designer's perspective, making design choices for multiple groups is more straightforward than making decisions for a myriad of people with unknown features [184], which makes the approach generally applicable. At the same time, by having multi-dimensional "personas" derived from clustering results, it is easy to specify learners' preferences, challenges, and needs, which can
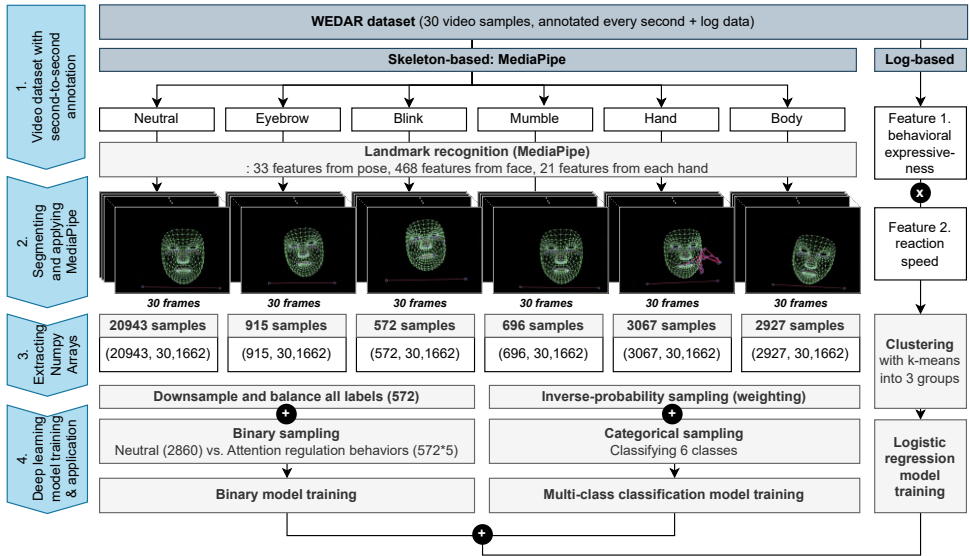
Figure 8.1: The overall framework: attention regulation behavior recognition based on MediaPipe framework and feedback adaptation strategies based on k-means clustering

be used for feedback personalization.

In this regard, we investigate critical components in understanding learners' attention management and distractions and use it as a means to make the feedback adaptive based on their learning needs through our work.

8.2 Methods

8.2.1 Overview of real-time feedback loop for attention management

This section represents the overall framework of our work (see Figure 8.1). Due to the context and domain-specific nature of the behavior analysis [156], we utilized the WEDAR dataset as a foundation of our work, which has the same target learning scenarios and settings: attention regulation behaviors of higher education learners in e-reading. The WEDAR dataset consists of video data with second-to-second behavior annotation on attention regulation and various log data, such as self-reported distractions, reaction time to the randomized screen blur, and knowledge gain evaluated from pre-post questionnaires. We applied the MediaPipe framework [291] to the video samples with 30 participants, and found landmarks on the pose, face, and hand. Each one-second clip has been divided into 30 frames, and the landmarks of each frame have been calculated with 1662 features, which consists of 126 landmarks in hands, 132 landmarks in pose, and 1404 landmarks in the face. Thirty multi-dimensional matrices from each frame, representing the spatial information of face and body, were used for the training, so the learners' "movements" with temporal features can also be taken into account for the model training. As a result, two behavior recognition models, a binary model for differentiating attention regulation

behavior vs. neutral and a multi-class classification model for finding each type of attention regulation behavior and neutral behaviors, have been developed for further model fusion in the deployment.

Aside from the behavior recognition models, we developed a model that can cluster learners according to their *reaction time to screen blur* and *behavioral expressiveness* using the WEDAR dataset. We first defined 3 clusters via k-means clustering with the elbow method to predict to which cluster the new instances from our newly collected dataset would be assigned. Based on the cluster membership and its recognition via logistic regression, we set up the feedback adaptation strategy: the personalized update of the feedback interval. The system provides more frequent feedback for learners who need more learning support than other learners with less feedback needs. More details on how we implemented it are available in the following sections.

8.2.2 System architecture and algorithm overview

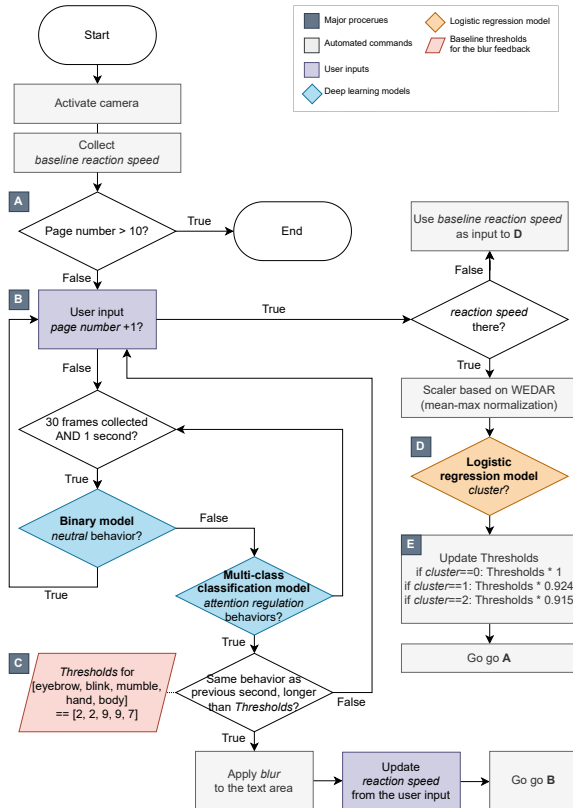Figure 8.2: Overall system architecture with the data streamline and model deployments.

This subsection explains our work's overall system architecture and algorithm deployments (see Figure 8.2). Our reading system contains ten pages of text with different lengths. The

system's primary function is to 1) provide screen blur feedback to the text area based on the automatic recognition of attention regulation behaviors for an extensive period of time and 2) collect various log data related to different learning behaviors that are used for the next round of feedback adaptation, which forms a loop.

Behavior recognition is conducted based on binary and multi-class classification model fusion. We took the hybrid approach since it often achieves more robust accuracy, compensating for each model's weaknesses [306]. By having two rounds of behavior recognition using a hybrid approach, we intended to minimize the false-positive case of recognizing attention regulation behaviors, especially mixed recognition between "body" and "neutral" labels, which leads to subsequent false positive feedback triggers that could greatly hinder learners' e-reading and lower the overall trust towards our system.

Once 30 frames are collected through the webcam within a second, the multi-dimensional matrices processed through MediaPipe go through the binary model to discern whether the behavior is "neutral" or one of "attention regulation behaviors". Please refer to the subsection of "Binary classification model design and configurations for training" for technical details. If the binary model recognizes the behavior as "neutral", it passes and starts the behavior recognition for the next frame set in the next second. If the behavior is recognized as "attention regulation behavior", the frames go to the multi-class classification model to classify the specific types of attention regulation behaviors. The multi-class classification model has been trained to differentiate six classes (eyebrow, blink, mumble, hand, body, and neutral). Please see the subsection of "Multi-class classification model design and configurations for training" for the details of the model training.

For personalizing the feedback, the thresholds update every page based on learners' behaviors on the previous page. From the clusters derived from the k-means clustering with the WEDAR dataset and logistics regression model to discern derived clusters, the feedback trigger thresholds are updated at the beginning of every new page. Learners' reaction time to blur stimuli that indicate their arousal and the learners' behavioral expressiveness, indicating more perceived distractions, are used as data inputs for predicting the clusters via the logistics regression. For learning groups requiring more behavioral corrections, the system shortens the feedback intervals. In contrast, the interval stays the same for learners with good behavioral performances, adapting to different feedback needs. Please refer to the subsection "Model design for feedback adaptation: k-means clustering combining with logistic regression" for details for k-means and logistic model training and coefficients applications for feedback interval updates.

8.2.3 Attention regulation behavior recognition model developments with neural networks

As indicated in the previous section, multi-dimensional matrices with 1662 behavioral features with 30 frames have been used for the model training. This section discusses how the binary and multi-class classification models were trained and validated for further implementation in the system.

Binary classification model design and configurations for training

For the binary model training, we first balanced the number of samples of all the attention regulation behaviors. We only have 572 samples of "blink" behavior, while examples of

other behaviors range from 696 to 3067 samples. We have randomly chosen 572 samples from each attention regulation behavior, making the sum of attention regulation behavior samples 2860 (572*5). It was to prevent label imbalance, which often affects the recognition accuracy. At the same time, we randomly chose 2860 out of 20943 samples from the "neutral" label. For the model training, we divided the training and testing set into a ratio of 80%-20%. Furthermore, we assigned 25% of the training set as the validation set, which made the ratio of training, testing, and the validation set 60%, 20%, and 20%, respectively. Please refer to the Table 8.1, in which we designed a 1D convolutional neural network (CNN) model to perform classification. Notably, the original WEDAR dataset exhibits a class imbalance, with neutral behaviors (20,943 instances) being approximately three times more prevalent than attention regulation behaviors (8,177 instances). In fine-tuning, we assigned a weight of 1.1 to neutral behaviors and 1 to attention regulation behaviors. Our binary classification model was trained over 1,000 epochs, employing binary cross-entropy as the loss function. This approach was designed to mitigate the risk of overpredicting attention regulation behaviors, thus reducing the likelihood of false-positive feedback. Upon evaluation using the test set, the model achieved an accuracy of 72.46%.

The model initiates with a 1D convolutional layer ('conv1d_16') of 64 filters, which allows the network to learn complex features from the input data of unspecified length, while another 1D convolutional layer ('conv1d_17') maintains a depth and significantly reducing the parameters. To stabilize and accelerate training, a batch normalization layer ('batch_normalization_4') has been employed, followed by a max-pooling layer ('max_pooling1d_12') that downsamples the spatial dimensions by half, mitigating the risk of overfitting while preserving the salient features. Furthermore, a dropout layer ('dropout_12') was inserted to prevent the co-adaptation of hidden units by randomly omitting a fraction of them during training. The following convolutional ('conv1d_18' and 'conv1d_19') and max-pooling layers ('max_pooling1d_13' and 'max_pooling1d_14') iterate the feature extraction and down-sampling processes. A second dropout layer ('dropout_13') was placed before the global max-pooling layer ('global_max_pooling1d_4'), summarizing the most significant features in the temporal dimensions. Two dense layers ('dense_8' and 'dense_9') progressively refined the extracted features into 32-dimensional and 2-dimensional spaces, respectively, the latter aligning with binary classification.

Multi-class classification model design and configurations for training
To train the multi-class classification model, we utilized the same number of samples for each class as in the binary classification task, comprising 2,860 instances each of attention regulation and neutral behaviors. The dataset was divided into training, testing, and validation sets with proportions of 60%, 20%, and 20%, respectively. The class weights were assigned inversely proportional to the original data distribution, resulting in weights of 2.190 for eyebrow, 3.501 for blink, 2.418 for mumble, 2 for hand, 0.995 for body, and 2 for neutral behaviors. This weighting strategy was fine-tuned to achieve better accuracy. The model underwent 2,000 epochs of training, achieving an accuracy of 72.74% on the test set. The confusion matrix providing detailed insights is depicted in Figure 8.4.

As can be seen in Table 8.2, the initial layer of the model, a 1D convolutional layer (conv1d_19) manages the initial feature extraction, while another 1D convolutional layer (conv1d_20) further refines the feature abstraction. In order to stabilize the learning process and to reduce internal covariate shifts, a batch normalization layer (batch_normalization_6) has

Table 8.1: Binary classification model development for recognizing attention regulation behaviors vs. neutral behaviors (2-class)

```
────────────────────────────────────────────────────────────────
Layer (type)                    Output Shape          Param #
================================================================
conv1d_16 (Conv1D)              (None, 30, 64)         319168

conv1d_17 (Conv1D)              (None, 30, 64)         12352

batch_normalization_4 (Bat      (None, 30, 64)         256
chNormalization)

max_pooling1d_12 (MaxPooli      (None, 15, 64)         0
ng1D)

dropout_12 (Dropout)            (None, 15, 64)         0

conv1d_18 (Conv1D)              (None, 15, 64)         12352

max_pooling1d_13 (MaxPooli      (None, 7, 64)          0
ng1D)

conv1d_19 (Conv1D)              (None, 7, 64)          12352

max_pooling1d_14 (MaxPooli      (None, 3, 64)          0
ng1D)

dropout_13 (Dropout)            (None, 3, 64)          0

global_max_pooling1d_4 (Gl      (None, 64)             0
obalMaxPooling1D)

dense_8 (Dense)                 (None, 32)             2080

dropout_14 (Dropout)            (None, 32)             0

dense_9 (Dense)                 (None, 2)              66

================================================================
Total params: 358626 (1.37 MB)
Trainable params: 358498 (1.37 MB)
Non-trainable params: 128 (512.00 Byte)
────────────────────────────────────────────────────────────────
```

Table 8.2: Multi-class classification model development for recognizing five types of attention regulation behaviors and neutral behaviors (6-class)

```
Layer (type)                    Output Shape              Param #
=================================================================
conv1d_19 (Conv1D)              (None, 30, 64)            319168

conv1d_20 (Conv1D)              (None, 30, 6              12352

batch_normalization_6 (Bat      (None, 30, 64)            256
chNormalization)

max_pooling1d_12 (MaxPooli      (None, 15, 64)            0
ng1D)

dropout_10 (Dropout)            (None, 15, 64)            0

conv1d_21 (Conv1D)              (None, 15, 64)            12352

max_pooling1d_13 (MaxPooli      (None, 7, 64)             0
ng1D)

conv1d_22 (Conv1D)              (None, 7, 64)             12352

max_pooling1d_14 (MaxPooli      (None, 3, 64)             0
ng1D)

dropout_11 (Dropout)            (None, 3, 64)             0

dense_8 (Dense)                 (None, 3, 32)             2080

dropout_12 (Dropout)            (None, 3, 32)             0

global_max_pooling1d (Glob      (None, 32)                0
alMaxPooling1D)

dense_9 (Dense)                 (None, 6)                 198

=================================================================
Total params: 358758 (1.37 MB)
Trainable params: 358630 (1.37 MB)
Non-trainable params: 128 (512.00 Byte)
```

been applied, normalizing the output from the convolutional layer. A max-pooling layer with size 2 (max_pooling1d_12) has been employed, reducing the spatial dimensions of the output by half without introducing additional parameters. Subsequently, a dropout layer with the rate of 0.25 (dropout_10) has been applied, randomly setting a proportion of input units to 0 at each update during training time, which helps to prevent overfitting. Two more convolutional layers (conv1d_21 and conv1d_22) have continued the hierarchical feature extraction, with max-pooling layers with size 2 (max_pooling1d_13 and max_pooling1d_14) that progressively down-sample the spatial dimensions of the extracted features, further enhancing the model's translational invariance. An additional dropout layer (dropout_11) with a rate of 0.25 and a densely connected layer (dense_8) facilitated a transition from feature extraction to classification. Subsequently, a dropout layer (dropout_14) with a rate of 0.5 has been applied. Finally, the penultimate layer employed global max pooling (global_max_pooling1d), reducing each feature map to a single value summarizing the presence of each learned feature in the input. Finally, a dense output layer (dense_9) extracted high-level features to the output space.

8.2.4 Hybrid approach to decrease the false-positive feedback trigger

Intervening through non-timely feedback with learners leads to distrust toward AI-based automatic feedback [307]. It negatively affects the overall learning experience and motiva-

tion in learning [159]. We adopted the model hybrid approach to *decrease the false positive feedback* [306], a strategic and practical approach suggested from the previous work in real-time educational feedback implementation [308]. Combining the confusion matrices from the binary (Figure 8.3) and multi-class (Figure 8.4) classification models, we calculated the False Positive Rate (FPR) of predicting the attention regulation behaviors that lead to false positive feedback. This process shows how the model fusion effectively decreases the false positive feedback trigger.

The FPR of the overall attention regulation behaviors in 1 second is 7.68% when solely implementing the multi-class classification model, having 2.98%, 0.35%, 2.25%, 1.92%, and 0.18% of $FP_{\text{MultiClass,Eyebrow}}$, $FP_{\text{MultiClass,Blink}}$, $FP_{\text{MultiClass,Mumble}}$, $FP_{\text{MultiClass,Hand}}$, and $FP_{\text{MultiClass,Body}}$, respectively. However, our hybrid approach of adding the binary model to the architecture dramatically reduces the FPR to 2.2%, having FPR of 0.85%, 0.11%, 0.65%, 0.54%, and 0.05% of FPR_{Eyebrow}, FPR_{Blink}, FPR_{Mumble}, FPR_{Hand}, and FPR_{Body}, respectively, as indicated in the equations articulated below. Once further implemented with the feedback threshold, ranging from 2 to 9 seconds, the chances of triggering false-positive feedback from a behavior drop down from nearly 0% since the rate takes square 2 to 9 square with the threshold implementation.
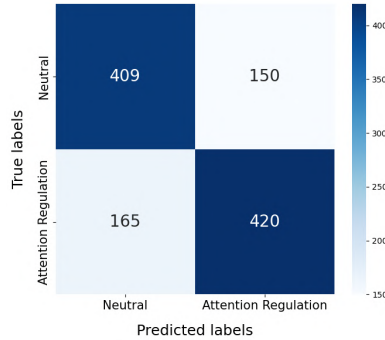
Figure 8.3: Confusion matrix from the binary classification model.

$$FPR_{\text{Eyebrow}} = FP_{\text{Binary,AttenReg}} \times FP_{\text{MultiClass,Eyebrow}} = 0.282 * 0.030 \approx 0.0085(0.85\%), \qquad (8.1)$$

$$FPR_{\text{Blink}} = FP_{\text{Binary,AttenReg}} \times FP_{\text{MultiClass,Blink}} = 0.282 * 0.004 \approx 0.0011(0.11\%), \qquad (8.2)$$

$$FPR_{\text{Mumble}} = FP_{\text{Binary,AttenReg}} \times FP_{\text{MultiClass,Mumble}} = 0.282 * 0.023 \approx 0.0065(0.65\%), \qquad (8.3)$$

$$FPR_{\text{Hand}} = FP_{\text{Binary,AttenReg}} \times FP_{\text{MultiClass,Hand}} = 0.282 * 0.019 \approx 0.0054(0.54\%), \qquad (8.4)$$

$$FPR_{\text{Body}} = FP_{\text{Binary,AttenReg}} \times FP_{\text{MultiClass,Body}} = 0.282 * 0.002 \approx 0.0005(0.05\%). \qquad (8.5)$$

Model design for feedback adaptation: k-means clustering combining with logistic regression

This section discusses how feedback adaptation was achieved in our real-time loop. Though the baseline feedback thresholds were designed to accommodate general users based on the previous research [292], our work also considers learners' behaviors *during* the reading practice in deciding the feedback intervals.
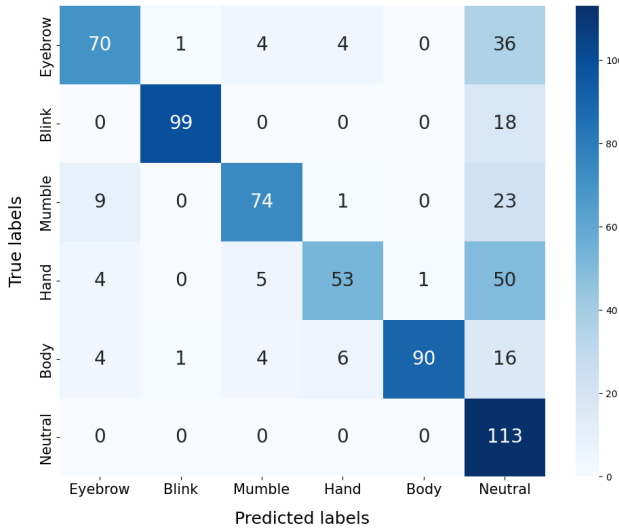
Figure 8.4: Confusion matrix from the multi-class classification model.
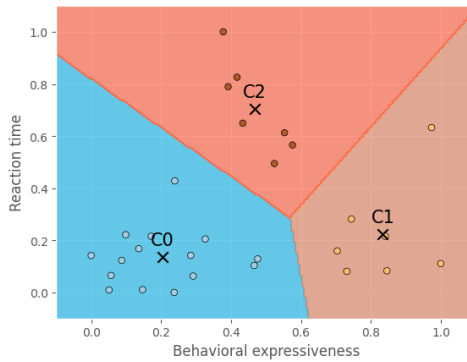
Figure 8.5: K-means clustering was conducted on the WEDAR dataset based on the reaction time to the screen blur and behavioral expressiveness. Three colored sections were derived from the logistic regression model training for predicting C0, C1, and C2 clusters.

First, from the WEDAR dataset, 30 learners' *reaction time to screen blur* and *behavioral expressiveness* have been used as feature vectors for k-means clustering. Instead of using all sample instances, we utilized averaged results of individuals because each individual might have a different baseline reaction time and behavioral expressiveness. Imbalances among different learners' data points might diminish the traits we can learn about individual differences. See the equations below to see how we calculated individuals' reaction time to blur stimuli as

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \tag{8.6}$$

where each individual i ($i = 1, 2, ..., 30$) has n_i data points from reaction time instances that are different based on individuals, while j (j=1, 2 ..., j_i) indicates the reaction time data point index for each individual. x_{ij} represents each reaction time data point for an individual and $\bar{x}_i$ shows an individual *reaction time to screen blur*.

Given the WEDAR dataset with behavioral data from 30 individuals, behavioral expressiveness has been calculated as

$$E_i = \frac{a_i}{b_i}, \tag{8.7}$$

where each individual i ($i = 1, 2, ..., 30$) exhibits a_i attention regulation behaviors and has b_i total behavioral data points, the *behavioral expressiveness*, E_i, for each individual i.

Using *reaction time to screen blur* and *behavioral expressiveness* as feature vectors, we conducted the k-means clustering. We applied the mean-max normalization to the data to prevent possible bias from different data ranges. The elbow method, a commonly applied statistical method for finding an optimal k for clustering, was applied and derived 3 as k in our case. As seen from Figure 8.5, three clusters have been derived: cluster 0 (C0) with the fast reaction time and low behavioral expressiveness, which we see as an ideal group of learners with more arousal and fewer distractions. Cluster 1 (C1) has been defined as learners with comparatively fast reaction time and high behavioral expressiveness, requiring more feedback than learners assigned to C0. Cluster 2 (C2) is a group with relatively slow reaction time and mid-range behavioral expressiveness, requiring more feedback. The specific coefficient has been calculated for the centroids of each cluster, which has been visualized as × in Figure 8.5. Note that one sample from the WEDAR dataset has been classified as an outlier and, thus, removed from the clustering task. This is due to the fact that k-means clustering is susceptible to outliers, thus can lead to misleading clustering results.

We adopted the logistic regression method to predict where learners belong to clusters at the beginning of a new page. Logistic regression has been carefully compared with other methods, such as K-Nearest Neighbors (kNN), and chosen for its computational efficiency, which is essential for our real-time approach. Also, because of its low variance, the method causes fewer overfitting issues, especially for cases with limited data that apply to our scenario. Please note that we also considered directly calculating the distance between the new data points and centroids of each cluster to assign new data points; however, we used the prediction based on the logistics regression for the future framework extension and better applicability of our work with scalability. If we add more features in determining the clusters with increased dimensions, calculating the distance becomes exponentially intensive in terms of computational capacity, which is unsuitable for real-time deployment of our framework.

Thresholds design and updates for different learner clusters

The thresholds (T) of triggering the screen blur have been designed differently for each attention regulation behavior. It combines the mean (μ) and standard deviation (σ) of the duration of each attention regulation behavior, which was derived from [292], that has also been developed from the WEDAR dataset:

$$T(y) = \lfloor \mu(y) + \sigma(y) \rfloor, \tag{8.8}$$

$$\text{Thresholds}_{[\text{eyebrow,blink,mumble,hand,body}]} = [T(\text{duration}^{\text{eyebrow}}),$$
$$T(\text{duration}^{\text{blink}}), T(\text{duration}^{\text{mumble}}), \qquad (8.9)$$
$$T(\text{duration}^{\text{hand}}), T(\text{duration}^{\text{body}})].$$

It is to find the points that would be considered significantly above average, which means significant attention regulation behavior usage compared to the average instances. To make it easily applicable, we only took $\lfloor x \rfloor$, representing the nearest integer to x data point, which made each attention regulation behavior baseline thresholds of the eyebrow, blink, mumble, hand, and body as 2, 2, 9, 9, and 7, respectively. Figure 8.6 shows how the system interface looks once the blur feedback is applied to the text area.

Figure 8.6: Screen blur is triggered for recognizing attention regulation behaviors for longer than adaptive thresholds.

Based on the clustering result at the beginning of each page, Θ_0, Θ_1, and Θ_2 have been designed for learners assigned in C0, C1, and C2, respectively. Since we do not know the absolute weight values proven to work as the most effective intervention, we set the coefficients proportionally based on their centroids. We updated the threshold a maximum of nine times, not making the threshold always bigger than the average attention regulation behaviors. In any case, it was not to trigger the feedback for learners who show the average range of attention regulation behaviors. In this case, the minimum k is calculated as

$$\begin{bmatrix} 2 & 2 & 9 & 9 & 7 \end{bmatrix} \times k^9 \geq \begin{bmatrix} 2 & 2 & 4 & 4 & 4 \end{bmatrix}, \qquad (8.10)$$

$$k \geq \left(\tfrac{23}{10}\right)^{\frac{1}{9}} \approx 0.915. \qquad (8.11)$$

Once applying min-max normalization to centroids of C0, C1, and C2, the behavioral expressiveness of C0 is 0.205, and the reaction time average is 0.135, which made the scaling factor of C0 as *36.133* (1/0.205*1/0.135). The behavioral expressiveness of C1 is 0.834, and the reaction time average is 0.223, making the scaling factor of C1 *5.377* (1/0.834*1/0.223). The behavioral expressiveness of C2 is 0.468, and the reaction time average is 0.705, making the scaling factor of C2 *3.031* (1/0.468*1/0.705). Please note that we considered C0 as an ideal group and made Θ_0 as 1, which does not change the feedback intervals. Based on the

equation (11), we normalized each scaling factor between 0.915 and 1, and derived Θ_0, Θ_1, and Θ_2 as the following:

$$\Theta_0 = 0.915 + \left(\tfrac{36.133 - 3.031}{36.133 - 3.031} \right) \times (1 - 0.915) = 1 \tag{8.12}$$

$$\Theta_1 = 0.915 + \left(\tfrac{5.377 - 3.031}{36.133 - 3.031} \right) \times (1 - 0.915) \approx 0.924 \tag{8.13}$$

$$\Theta_2 = 0.915 + \left(\tfrac{3.031 - 3.031}{36.133 - 3.031} \right) \times (1 - 0.915) = 0.915. \tag{8.14}$$

8.2.5 Experimental conditions
We had four experimental conditions: 1) control group, 2) baseline group, 3) treatment_-subcondition 1 (treatment_sub1), and 4) treatment_subcondition 2 (treatment_sub2) as below.

Control condition (data from 30 participants)
The control condition is the WEDAR dataset with screen blur feedback at randomized timing. The control condition was compared with our prototype with AI-based behavior recognition and feedback. The control group condition has been further compared with treatment conditions with our prototype on learners' self-reported distractions, knowledge gain, and perceived system experiences.

Baseline condition (data from 30 participants)
Since we aimed to understand not only the objective effects of AI-based feedback on distractions but also learners' perceptions of system experiences with the AI-based feedback, we included the baseline group as one experimental condition. The WEDAR dataset does not contain data regarding learners' perceptions of their system experiences; thus, the baseline condition was collected by asking participants about their perceptions of computer screen-based e-readers before the experiment.

Treatment_sub1 (data from 15 participants)
Participants in the Treatment_sub1 condition was based on the prototype that we developed with behavior recognition and AI-based feedback. Learners were tested without being notified about the feedback mechanisms behind the AI-generated screen blurs.

Treatment_sub2 (data from 15 participants)
Participants in the Treatment_sub2 condition was based on the prototype we developed with behavior recognition and AI-based feedback, which is the same as the Treatment_-sub1 condition, but with an explanation of the feedback mechanisms before they were tested. While experiment instructions were given, the definition and examples of attention regulation behaviors and their correlation to self-reported distractions were explained.

8.2.6 Measures
Attention regulation behaviors: behavior labels
In the previous study [156], attention regulation behaviors have been defined as five observable behavioral categories: movements found from eyebrow (e.g., raise or bring together), blinks (e.g., blink flurries, prolonged blink), mumble (e.g., mumble reading), hand

(e.g., touching face, body), and body (e.g., adjusting head, torso, arm). In this work, we followed the annotation criteria and experiment design from [156], so the effect of the behavior recognition-based AI feedback can be directly compared with the WEDAR. After we annotate the newly collected video samples, we further investigate the occurrence of attention regulation behavior in relation to self-reported distraction in our work, in comparison with the WEDAR.

Knowledge gain: multiple-choice & summarization
In this work, learners' knowledge gain is understood in two folds: assessing learners' factual knowledge based on the *multiple-choice questions* [167], while *summarization task* reflects learners' deeper comprehension, knowledge synthesis, and critical thinking skills [170]. We designed the pre-test and post-test questionnaires with the same content, with 10 multiple-choice questions related to the reading material. We extracted the pre-test result from the post-test result to gauge the knowledge gained through the reading. The summarization task was asked for one main topic and the 10 subtopics after reading. To avoid the possible bias from human evaluation, we implemented the Bidirectional Encoder Representations from Transformers (BERT) model, a transformer-based Natural Language Processing (NLP) model, which has also been adopted for the summarization evaluation [309]. Since the model provides scores that average each answer's precision and recall with contextual embeddings, semantic similarity to the original text, and coherence of summarization, we adopted it as an effective automatic summarization evaluation method.

Perceived interaction experience: AttrakDiff questionnaire
AttrakDiff questionnaire, a tool developed for understanding users' perceptions towards interactive products [77], has been used to understand learners' perceived interaction experience with AI-based feedback. It is a tool that has been previously adopted for the Human-Computer Interaction (HCI) [310] and Human-Robot Interaction (HRI) [158] evaluations. The questionnaire consists of 28 questions, and we compared learners' perceptions regarding *computer screen-based e-readers (baseline)* and *our e-reader with AI-based automatic feedback (treatment)* with the pre-post comparison. The questionnaire evaluates four dimensions: 1) Pragmatic quality, which concerns perceptions about the usability of the system, and 2) hedonic identity (hedonic-I), which figures out users' perceptions that identify the system. It includes 3) hedonic stimulation (hedoinc-S), which is about advancing values that the system conveys, and 4) attractiveness, which delivers perceptions regarding the likeability of the system.

8.2.7 Procedure

30 higher education learners (male: 17, female: 13) have been invited for an e-reading task. They were tested with a system with AI-based screen blur feedback based on real-time attention regulation behavior recognition. Participants were students who use the English language for their daily education with frequent computer-based reading (*frequently* to *usually*: 5.8 out of 7-scales Likert). They participated in the study voluntarily via the TU Delft study recruitment portal and campus.

The participants were individually invited to the experiment room with a laptop [2] and a

---

[2]Dell XPS 15 9570, Intel(R) Core(TM) i7-8750H (CPU), NVIDIA GeForce GTX 1050 Ti (GPU)

mouse, which makes the experiment environment similar to their daily computer-based reading environments. Participants were requested to work on the pre-test questionnaire to measure their 1) prior knowledge regarding the reading content (e.g., Disneyland in Paris) with 10 multiple-choice questions. Also, an AttrakDiff questionnaire was given to measure learners' preconceptions about computer screen-based e-readers and their perceptions of them (baseline).

Once the pre-test was finished, instructions were given differently based on the different sub treatment conditions (treatment_sub1, treatment_sub2), though all 30 participants were given the same system. It was to evaluate how the explainability of the feedback would affect the perceptions about their learning experiences with AI-based feedback in e-reading and the subsequent learning outcomes.

1) For 15 participants in the treatment_sub1 condition, the detailed mechanism behind the feedback trigger has not been described. It has only been instructed that feedback occurs based on their behavioral cues to keep them attentive. 2) For another 15 participants in the treatment_sub2 condition, the definition of attention regulation behaviors, examples of behaviors, and their correlation to self-reported distractions have been elaborated. Furthermore, participants in the treatment_sub2 condition were informed that feedback thresholds update on each page, corresponding to one of three groups derived from k-means clustering based on their learning behaviors. During e-reading, all participants were asked to report their distractions via the big button on the interface for the self-reported distraction analysis, regardless of the treatment condition.

After reading, all participants received the same questionnaire: 1) First, the post-test questionnaire had 10 multiple-choice questions that were the same as the pre-test questions. However, unlike the pre-test questionnaire, there were additional questions for participants to summarize the main topic and 10 subtopics in short sentences. 2) There has been a following AttrakDiff questionnaire to ask about their perceived experiences with the system. Please note that all questionnaires have been given paper-based to prevent participants' potential confusion in perceiving the system evaluation process as part of their learning.

8.3 Results

In this section, we investigate the various effects of our system implementation. First, we focus on learners' attention regulation behaviors to understand learners' arousal and perceived distractions on group and individual levels. Second, with statistical and correlation analyses, we study the effects of AI-based feedback on learners' knowledge gain, a fundamental objective of the e-reading activities. Through the feature analysis from the decision tree and coefficient analysis from the logistic regression, we strive to find what works as critical components for predicting high knowledge gain. Lastly, we understand the perceived interaction experiences of learners and see the effects of AI-based feedback and the roles of explainability of feedback.

8.3.1 Effects on Learners' Behaviors: self-reported distractions and attention regulation behaviors

Statistical analysis on self-reported distractions with ANOVA

In this subsection, we investigate self-reported distractions of learners in control, treatment_sub1, and treatment_sub2 conditions. The analysis is closely linked to our implemen-

tation goal, aiming at 1) fewer distractions and 2) extending the overall attention span with decreased distraction intervals. 2-condition and 3-condition comparisons were conducted to see if the explainability of AI-based feedback affected learners' perceived distractions. Furthermore, the one-way Analysis of Variance (ANOVA) was conducted to evaluate if the distinctions between the groups were statistically significant. As represented in Table 8.3, in both 2-condition and 3-condition comparisons, the treatment group reported fewer distractions than the control group. In the 2-condition comparison, the control group reported an average of 11.2 times distractions, while the participants in the treatment condition reported 6.5 times of distractions. The difference between the control and treatment groups' self-reported distinctions in the 2-condition comparison was more evident (p=0.024) than in the 3-condition comparison (p=0.068).

In the 3-condition comparison, treatment_sub1 with AI-based screen blur feedback *without* feedback rules explained, the distraction has been reported on average 6.87 times, while in the treatment_sub2 condition with AI-based screen blur feedback *with* rules explained had self-reported 6.13 times of distractions on average. The result indicates that our AI-based feedback contributed significantly to decreased distractions. The effect was evident even without the explainability of the feedback, which shows that the implementation of the feedback affected the improvement in objective attention management.

Table 8.3: The number of self-reported distractions and their significance in 2-condition comparison and 3-condition comparison.

Number of distraction reports (times)					
Conditions		M(SD)	One-way ANOVA		
		M(SD)	F	df1	df2	p
2-condition comparison	Control	11.2 (9.39)	5.47	1	47.3	**0.024**
	Treatment (sub 1 & 2)	6.50 (5.59)				
3-condition comparison	Control	11.2 (9.37)	2.91	2	34.7	0.068
	Treatment_sub1	6.87 (6.45)				
	Treatment_sub2	6.13 (4.79)				

The interval of the self-reported distractions has statistically significantly (p=0.025) increased in the treatment condition (142 seconds) compared to the control condition (88.6 seconds) in the 2-condition comparison. When looking into the 3-condition comparison, the distraction interval of treatment_sub1 has been reported as 107 seconds, while it from treatment_sub2 has been 178 seconds. It suggests that AI-based feedback has contributed to increased distraction report interval, meaning it has contributed to increased learners' attention spans. The result indicates that having explainability of the feedback contributed to the increased distraction report interval but without statistical significance. In line with the previous result regarding the number of distractions, adopting AI-based feedback positively and significantly influenced distraction intervals and learners' attention spans, an objective contribution of our AI-based screen blur feedback in e-reading, regardless of the feedback explainability. Please note that the self-reported distraction has been analyzed only in available instances. For instance, learners who did not report distractions were not considered for this analysis (each of the two instances in the WEDAR dataset and our work). Also, the distraction interval analysis was not conducted for participants with only one self-reported distraction instance.

Table 8.4: The distraction intervals and their significance in 2-condition comparison and 3-condition comparison.

Intervals of distraction reports (s)					
Conditions		M(SD)	One-way ANOVA		
		M(SD)	F	df1	df2	p
2-condition comparison	Control	88.6 (47.6)	5.49	1	34.1	**0.025**
	Treatment (sub 1 & 2)	142 (108)				
3-condition comparison	Control	88.6 (47.6)	2.88	2	21.1	0.078
	Treatment_sub1	107 (55.6)				
	Treatment_sub2	178 (135)				

Statistical analysis on attention regulation behaviors for the future feedback thresholds design

We analyzed attention regulation behavior durations to see any behavior differences coming from different conditions. We followed the analysis method of the BFLAe framework [292], which we used in our work for deciding the feedback thresholds. This analysis helps understand how long learners typically show attention regulation behaviors in their e-reading. Thus, it gives the average and exceptional ranges of behaviors that we can use for setting up feedback trigger thresholds. As described in the subsection of "Feedback implementation with personalization", the rationale of our current feedback implementation has been based on the attention regulation behavior recognized for extensive periods (i.e., the integer part of (Mean+SD of each attention regulation behavior)), which concluded baseline thresholds of the WEDAR as 2, 2, 9, 9, and 7 seconds for movements of eyebrow, blink, mumble, hand, and body, respectively. Our work represents new baseline thresholds for different sub-conditions: 1, 2, 6, 9, and 7 seconds for AI-based screen blur without explainability while having 1, 2, 11, 9, and 7 seconds for AI-based screen blur with explainability.

Our analysis has shown a consensus about the baseline statistics of each attention regulation behavior regardless of experimental conditions. It helps the general interpretations and utilization of attention regulation behaviors as essential indicators for the AI-based feedback loop design. For example, the median of neutral, eyebrow, blink, mumble, and hand has been 5, 1, 1, 2, and 2 seconds, regardless of the experimental conditions. Similar patterns were observed in the quartiles of various experimental conditions: the first quartiles of neutral, eyebrow, blink, mumble, and hand have been reported as 2, 1, 1, 1, and 1 second(s) in all conditions. The second quartiles of neutral, eyebrow, blink, mumble, and hand behaviors are commonly reported as 5, 1, 1, 2, and 2 second(s). The third quartiles, likely to be related to feedback triggers since the third quartiles hold more prolonged attention regulation behaviors, eyebrow, blink, and hand, were the same as 1, 1, and 4 second(s) regardless of conditions. The result can be used as a foundation for behavior-based attention analysis and feedback design for e-reading.

8.3.2 Effects on Learners' Cognition: Knowledge gain

Statistical analysis with ANOVA

The average knowledge gain, represented by the multiple-choice questions and summarization results, recorded the highest in the treatment_sub2 condition without significance in the 3-conidtion comparision (see Table 8.6). As can be seen from Figure 8.7, the distribution of the knowledge gain evaluated through multiple-choice questions shows statistically significant skewness to the left (skewness: -0.925, std error skewness: 0.309, Z=-2.99),

Table 8.5: One-way ANOVA conducted on three classes (one control, two treatment groups with different feedback instructions) on knowledge gain and perceived interaction experiences.

Attention regulation behavior durations (s)								
Conditions	Labels	Min	Max	M (SD)	Median	Q1	Q2	Q3
Control	Neutral	1.0	124.0	9.44 (12.73)	5.0	2.0	5.0	12.0
	Eyebrow	1.0	5.0	1.20 (0.52)	1.0	1.0	1.0	1.0
	Blink	1.0	5.0	1.14 (0.45)	1.0	1.0	1.0	1.0
	Mumble	1.0	5.0	3.15 (5.11)	2.0	1.0	2.0	3.0
	Hand	1.0	35.0	3.58 (4.88)	2.0	1.0	2.0	4.0
	Body	1.0	62.0	3.20 (2.85)	2.0	1.0	2.0	4.0
Treatment _sub1	Neutral	1.0	117.0	11.61 (15.22)	5.0	2.0	5.0	16.0
	Eyebrow	1.0	2.0	1.07 (0.25)	1.0	1.0	1.0	1.0
	Blink	1.0	5.0	1.12 (0.54)	1.0	1.0	1.0	1.0
	Mumble	1.0	19.0	2.83 (3.30)	2.0	1.0	2.0	3.0
	Hand	1.0	56.0	3.63 (5.40)	2.0	1.0	2.0	4.0
	Body	1.0	18.0	3.33 (2.42)	3.0	2.0	3.0	4.0
Treatment _sub2	Neutral	1.0	227.0	9.65 (13.98)	5.0	2.0	5.0	12.0
	Eyebrow	1.0	4.0	1.05 (0.32)	1.0	1.0	1.0	1.0
	Blink	1.0	3.0	1.14 (0.41)	1.0	1.0	1.0	1.0
	Mumble	1.0	32.0	4.78 (5.98)	2.0	1.0	2.0	6.0
	Hand	1.0	76.0	3.63 (5.7)	2.0	1.0	2.0	4.0
	Body	1.0	42.0	3.42 (3.62)	2.0	1.0	2.0	4.0

Table 8.6: One-way ANOVA conducted in 3-condition comparison with control, Treatment_sub1, and Treatment_sub2 groups on knowledge gain and perceived interaction experience.

Measurement		Control	Treatment_sub1	Treatment_sub2	One-way ANOVA			
		M (SD)			F	df1	df2	p
Knowledge gain	Multiple-choice	7.57 (2.14)	7.00 (2.85)	**8.13 (1.36)**	1.192	2	30.5	0.317
	Summarization	0.820 (0.006)	0.821 (0.006)	**0.825 (0.004)**	2.692	2	24.6	0.88

while the knowledge gain measured through the summarization task shows the skewness to the right without statistical significance (skewness: 0.435, Std Error Skewness: 0.309, Z=1.41). It indicates that the multiple-choice questions were generally perceived as more manageable than the summarization task, with generally higher scores gained than from the summarization task.

We found distinct patterns in the data distributions among experimental conditions in comparing multiple-choice question results. The control group exhibited a broad distribution, with its median centered around 8.0, culminating in a peak of 8.5. On the contrary, the distribution of treatment_sub1, with a median similarly positioned as the control group, starts from a higher point. This distribution remained relatively steady, with a subtle increase extending from the median to a score of 9. treatment_sub1 generally yielded higher scores with a more uniform distribution than the control group. On the other hand, the distribution of treatment_sub2 commenced above 5.0, presenting a relatively elevated median of approximately 8.5 and demonstrating an exponential distribution therein. It indicates that participants in the treatment_sub2 condition generally achieved higher scores with narrower distributions than the control group and treatment_sub1.

In comparing conditions based on the summarization result, there has been a tendency for the control group to have very thick mid-range scorers around the median of 0.820. However, the treatment_sub1 condition started with higher scores than the control group, resulting in generally even distributions with a slight increase at around 8.5 scores. The

distributions of the treatment_sub2 began from a relatively higher point than the control and the treatment_sub1 condition, with an exponential growth at approximately 8.5 to 9 scores.

Overall, implementing the AI-based feedback increased the knowledge gain shown through multiple-choice, with an exponential increase in participants with a high score band of 7.5 to 10. In the summarization task, adopting AI-based feedback with explainability brought students with higher score bands, generally higher than the control group and the treatment condition without feedback explainability without statistical significance.
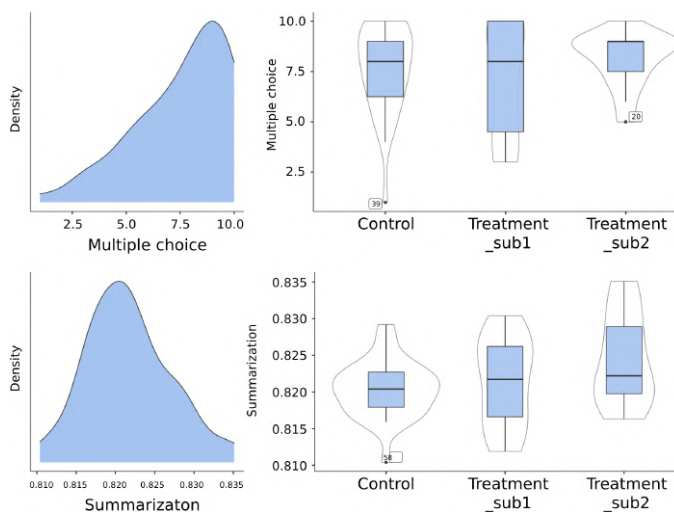
Figure 8.7: Knowledge gain distributions for multiple choice and summarization tasks in 3-condition comparison with control, Treatment_sub1, and Treatment_sub2 groups.

Correlation analysis between results from multiple-choice and summarization task

In this subsection, we investigate a potential correlation between the evaluation results from multiple-choice and summarization. We hypothesized that there would be a positive correlation between them, assuming that high scorers in the multiple-choice question would achieve highly in the summarization task and vice versa.

We first calculated three quartiles (i.e., Q1, Q2, Q3), providing ranges of low (0-25%), mid (25-75%), and high (75-100%) scores gained in each task in three experimental conditions, respectively. We assigned each learner 1, 2, and 3 based on the quartiles they belong to and conducted Pearson's correlation analysis. The result showed no correlation between evaluation results from multiple-choice and summarization (r=-0.125, p=.343), indicating that achieving well in one measure does not guarantee a high score in another in all experimental conditions.

Comparing high-mid-low achievers in retrospect: relation to attention regulation behaviors and perceived interaction experience

In this section, we strived for critical behavioral and affective components that might be related to the high knowledge gain, using correlation and tree-based analysis. We first hypothesized that higher achievers would have better perceptions of the system during their learning and would show more specific behavioral patterns in the learning process. However, unlike our assumption, correlation with perceived interaction experiences was not found in any cases among higher achievers, lower achievers, and combined achievement results.

We further utilized logistic regression and tree-based methods to deduce potential components to predict the high-mid-low achievers in knowledge gain. For the machine-based prediction, the following post-hoc features have been used to train the logistic regression and decision-tree models: "sub-conditions (3-class comparison)", "main conditions (2-class comparison)", "pretest (multiple-choice)", "posttest (multiple-choice)", "Attrakdiff total (AttrakDiff)", "Pragmatic (AttrakDiff)", "Hedonic-I (AttrakDiff)", "Hedonic-S (AttrakDiff)", "Attractiveness (AttrakDiff)", "distraction numbers", and "distraction intervals". Two machine learning models were selected since we can trace the coefficients of each feature or find the feature's importance in retrospect. By understanding the rationale behind machine reasoning, we can further investigate features that might be critical for learners' knowledge gain. We used 60 sample instances with three experimental conditions and adopted the traditional training and testing sample split of 80-20 ratios. As a result, predicting learners' knowledge gain with the multiple-choice questions has achieved 83.33% accuracy through the logistic regression and 91.67% accuracy through the decision tree.

The feature importance analysis from the decision tree (see Figure 8.8) indicated that the most influential feature in predicting the knowledge gain (multiple-choice) has been the "posttest result", regardless of the "pretest result". Learners' perceived "attractiveness" has taken 19.4% of feature importance, while "distraction intervals (9.7%), "Hedonic-I (5.7%)", "Hedonic-S (4.6%)", and "Pragmatic (3.2%)" qualities have been considered importantly for predicting the low-mid-high knowledge gain evaluation with multiple-choice questions.

Though the coefficient does not directly provide the importance of the feature, it is helpful to grasp the directional and proportional relationship between knowledge gain and other indicators. The coefficients from the logistic regression model have shown that the "posttest (coef: 1.409)" result best supports the predictions of the low-mid-high knowledge gain via multiple-choice questions. Also, lower "pretest (coef: -1.149)" scores often result in higher knowledge gain scores, and " Hedonic-S (coef: -0.712)" affects the prediction of low-mid-high achievements in multiple-choice questions. Note that general AttrakDiff evaluation results show negative coefficients with low-mid-high knowledge gain because the lower scores in the AttrakDiff measure indicate a more positive perception of learners. 3-class sub-conditions (0: baseline, 1: treatment_sub1, 2: treatment_sub2) have also had high negative coefficients (coef: -0.547) with the knowledge gain clusters, indicating that the AI feedback contributed to the higher knowledge gain compared to the control group with randomized feedback, and the positive effects of the AI feedback on knowledge gain was stronger with the feedback explained to learners (see the Figure 8.9).

We also strived to predict the low-mid-high level achievers in the summarization task. However, logistic regression and the decision tree did not achieve robust prediction accura-
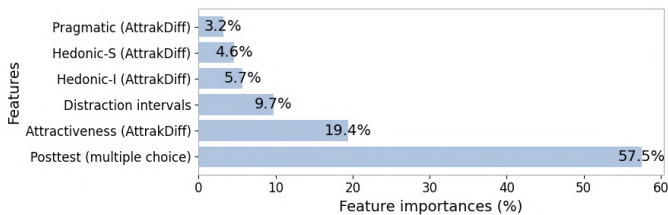
Figure 8.8: Feature importance analysis conducted for predicting knowledge gain clusters (3-level) from multiple-choice, using decision tree.
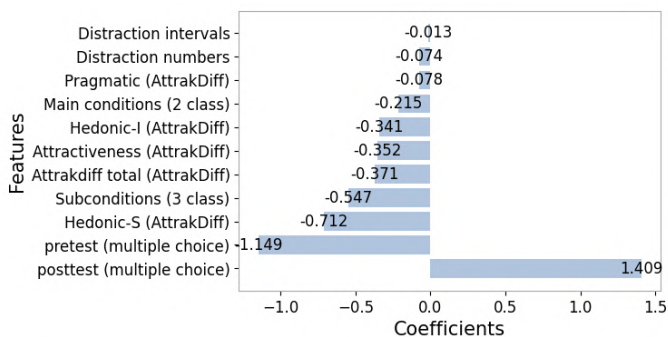

Figure 8.9: Feature importance analysis conducted for predicting knowledge gain clusters (3-level) from multiple-choice, using logistics regression.

cies (33.3% each), showing the same prediction accuracy as random guesses from humans. It might be because the features and the 60 sample instances we used for the prediction were insufficient to decide the summarization scores. Another possibility is that the BERT model's summarization score might not provide a valid evaluation of learners' knowledge gain. Please note that we did not go deeper into such possibilities since conducting the coefficient and feature importance analysis in our work was on finding the potential, influential features that affect the machine prediction on different levels of knowledge gain.

8.3.3 Effects on Learners' Perceptions: Perceived interaction experiences

Table 8.7: One-way ANOVA conducted in 3-condition comparison with control, treatment_sub1, and treatment_sub2 groups on perceived system experience.

Measurement		Baseline	Treatment_sub1	Treatment_sub2	One-way ANOVA			
		M (SD)			F	df1	df2	p
Attrakdiff	Overall	3.73 (0.386)	3.79 (0.991)	**3.29 (0.472)**	4.928	2	24.6	**0.016**
	Pragmatic	**2.84 (0.583)**	3.56 (1.20)	2.99 (0.856)	2.370	2	24.1	0.115
	Hedonic-Identity	3.77 (0.474)	4.00 (0.950)	**3.72 (0.548)**	0.482	2	25.6	0.623
	Hedonic-Simulation	4.77 (0.590)	3.48 (1.14)	**2.99 (0.677)**	39.727	2	25.8	**<.001**
	Attractiveness	3.55 (0.501)	4.14 (1.45)	**3.47 (0.629)**	1.347	2	24.2	0.279

Statistical analysis with ANOVA

As can be seen from Table 8.7, treatment_sub2 has recorded the best evaluation among baseline, treatment_sub1, and treatment_sub2 conditions in the hedonic-I, hedonic-S, and attractiveness measures, while the baseline condition has shown the best evaluation in the pragmatic measure.

According to the ANOVA analysis, statistical significance has been found from the overall AttrakDiff measure, in which we averaged four measures (i.e., pragmatic, hedonic-I, hedonic-S, and attractiveness) and the hedonic-S measures: it indicates that learners' perceptions on overall AttrakDiff measure and the hedonic-S were significantly differently depending on experimental conditions. Through Pearson's correlation analysis, we tried to find the correlation between individual responses from each of the AttrakDiff. The result represents the *overall AttrakDiff* measure to be significantly correlated with all submeasures: *pragmatic* (r=0.638***, p=<.001) *hedonic-I* (r=0.772***, p<.001), *hedonic-S* (r=0.610***, p<.001), and *attractiveness* (r=0.904***, r=p<.001). The notably strong correlation with attractiveness and the fact that the overall AttrakDiff score tends to increase significantly by its attractiveness. *Pragmatic* quality has been significantly correlated with *hedonic-I* (r=0.343**, p=.007) and *attractiveness* (r=0.0.636**, p=.007), indicating that the usability of the system is related to both pleasure derived from task-related use (Hedonic-I) and perceived attractiveness. It did not correlate with *hedonic-S* (r=-0.045, r=0.732), representing that pragmatic quality does not necessarily affect the pleasure derived from the products' features of satisfying learners' psychological needs for novelty and stimulation. *Hedonic-I* has shown a significant correlation with *attractiveness* (r=0.759, p<.001), suggesting that when users derive pleasure by using the system to achieve goals, they also find it more attractive. *Hedonic-S* has shown a correlation with *attractiveness* (r=0.321*, p=0.012), indicating that the aesthetic and appeal of the system are somewhat related to its ability to satisfy a user's need for stimulation and novelty.

AttrakDiff analysis

In this section, we investigate the learners' perceived system experience with descriptive analysis (see Figure 8.10). We compared the learners' responses based on different experimental conditions. Our focus has been: 1) comparing the baseline and the two treatment conditions to understand the perceived effects of an AI-based feedback adoption (baseline vs. treatment (sub 1 & 2)). Second, we strived to understand 2) the role of feedback explainability by comparing responses towards the two treatment conditions (treatment_sub1 vs. treatment_sub2).

Baseline condition with traditional e-readers without feedback has been evaluated as more "practical", "predictable", and "manageable", perceiving it has the most pragmatic quality among 3 experimental conditions. Also, it has been seen as somewhat more "professional" and "presentable" in the hedonic-I measure, indicating the importance of aesthetic considerations in system design, which has not been implemented in our work. Baseline condition has been evaluated as slightly more "undemanding", "pleasant", and "good" in hedonic-S quality and attractiveness, respectively. Other than eight submeasures out of 28, AI-based feedback has been evaluated more positively compared to the baseline: 4 out of 7 from pragmatic quality, 5 out of 7 from hedonic-I, 6 out of 7 from hedonic-S, and 5 out of 7 from attractiveness has been evaluated better than the baseline, showing the improvement made via the AI-based feedback.
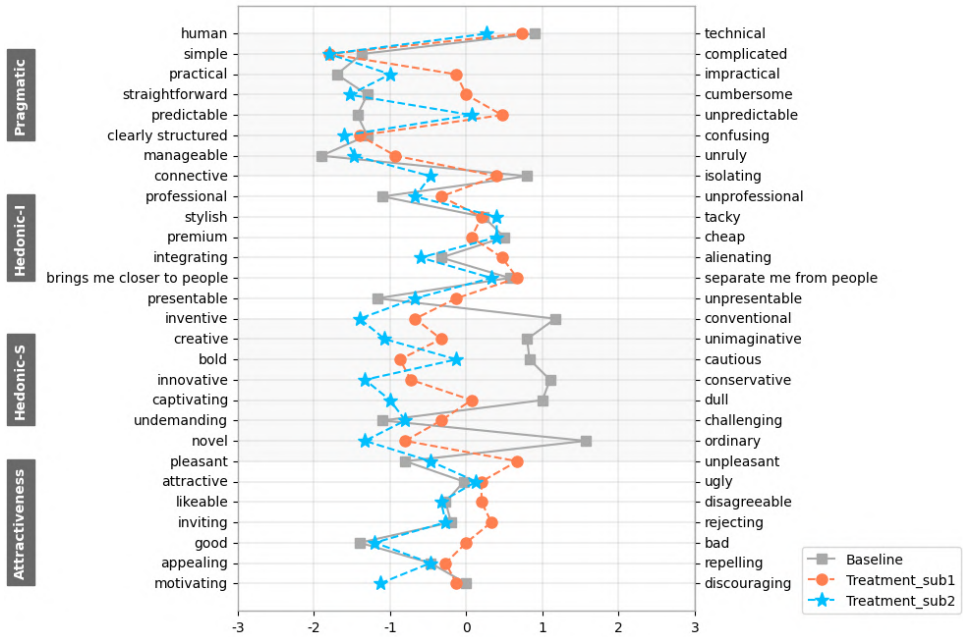
Figure 8.10: AttrakDiff analysis conducted on three experimental conditions: baseline, Treatment_sub1, and Treatment_sub2 conditions.

As mentioned in the subsection of "experimental conditions", the only difference between the treatment_sub1 and treatment_sub2 has been the instructions given to participants regarding the feedback rules though the exact same system was tested. As a result, treatment_sub2 with feedback explainability has been seen as more "human", "simple", "practical", "straightforward", "predictable", "clearly structured", and "manageable" in the *pragmatic* quality measure, while it has been evaluated as "connective", "professional", "integrating", "brings me closer to people", and "presentable" in *hedodnic-I* measure. At the same time, treatment_sub2 has also been perceived as more "inventive", "creative", "innovative", "captivating", "undemanding", and "novel" through *hedonic-S* measures while being assessed as "pleasant", "attractive", "likable", "inviting", "good", "appealing", and "motivating" in terms of attractiveness.

In summary, the results underscore the importance of explainability in AI-generated feedback and its impact on learners' perceptions. The result represents that when feedback rules are clearly communicated and understood, the system tends to be more favored and trusted, which might have impacted enhanced knowledge gain. It emphasizes that agreeable machine reasoning fosters greater human trust and is an essential insight for AI-based feedback implementation.

8.4 Discussion

The importance of feedback explainability in AI-based feedback: As revealed in the analysis of perceived system experience, the overall AttrakDiff and hedonic-S measures have shown statistically significant favor of system users for AI-based automatic feedback with the explainability, compared to feedback without the explainability. The results emphasize the importance of transparency and explainability in the AI-based feedback provision, which extends the discussion of the conventional explainable AI approaches into the realm of feedback design.

Importance of the hedonic attractiveness of the system for learners' perceptions: The correlations between hedonic and pragmatic qualities in the result suggest that these aspects of user experience are not mutually exclusive and can influence each other: a functional product (pragmatic) also tends to be pleasurable to use (hedonic), and vice versa. Attractiveness plays a pivotal role and is significantly correlated with overall submeasures. It underscores the importance of aesthetics and physical appeal in user perception and system experience, which has not been proactively underpinned in our prototype only with the GUI wireframe and, thus, should be studied further with the next round of implementations.

Feedback loop combining with other feedback components: In line with the previous discussion point, the feedback can be understood in line with other feedback components, such as feedback messages and multimodal feedback agents. Our work focused on implementing non-intrusive and timely feedback. However, the feedback messages supported by generative AI and feedback delivered through other multimodal feedback interfaces, such as conversational agents with a physical presence and speech-based interaction, might change the implications of the AI-based feedback.

Exploration of feedback adaptation with generative AI: Since our study aimed to find the potential effects of the AI-based feedback system with the feedback adaptation strategy, we adopted the rule-based thresholds, which are computationally inexpensive for real-time application and with interpretability. However, generative AI might further support our scenario, especially when the feedback is adopted via other conversational agents (e.g., chatbot, human-robot interaction).

8.5 Conclusion

In this work, we developed a feedback loop based on hybrid models to recognize learners' attention regulations in e-reading using skeleton-based neural networks. Theoretical and technological considerations have been aligned to decide the model training, feedback threshold design, feedback adaptation, and system evaluation. We evaluated the effects of implementing AI-based feedback, focusing on learners' behaviors, cognition, and perception by investigating learners' self-reported distractions, knowledge gain, and perceived interaction experiences.

Four experimental conditions have been compared: a control condition with randomized feedback based on the public WEDAR dataset, a baseline condition to understand learners' perceptions about conventional computer-based readers without feedback, and two treatment conditions with and without explainability about the feedback mechanisms.

The result indicates that our AI-based automatic feedback based on attention regulation behavior recognition contributed to fewer distractions and longer attention spans, regard-

less of the feedback's explainability. It shows our system objectively contributed to the attention management of learners engaged in e-reading. Also, implementing feedback has contributed to knowledge gain, helping learners achieve knowledge gain evaluated through multiple-choice and summarization tasks. Regarding the perceived interaction experiences, the evaluation of AI-based feedback with and without the explainability varied: learners highly valued the system's inventiveness, creativity, boldness, innovation, activation, and novelty, which stimulated learners when the instructions about the feedback mechanisms were introduced. It indicates that learners appreciate the system feedback more once they understand the rationale behind the specific feedback.

All in all, our work offers an empirical and comprehensive understanding of closing the AI-based real-time feedback loop in e-reading, with critical insights into designing and implementing the adaptive feedback loop.

9

Summary & Conclusion

9.1 Summary

With the widespread use of digital devices and platforms, e-reading, referred to as digital reading on computers in this dissertation, is becoming more commonplace in formal and informal education. Especially in higher education, the impact of e-reading is more significant since learners are required to register, comprehend, remember, reconstruct, and apply the knowledge based on their independent reading as a part of regular education. E-reading also directly affects learning outcomes, learners' self-efficiency, and learning success. Traditionally, for e-learning, log data for the post-hoc analysis of intervention has been used, while Technology-Enhanced Learning (TEL) approaches, leveraged by sensing and machine learning technologies, developed sophisticated real-time learning analytics and learning support. However, despite the importance of e-reading in higher education and technological advancements that enable seamless TEL applications, real-time intervention design and implementation for e-reading seem scarce. In this regard, this dissertation explores the possibilities of adopting a real-time feedback loop in e-reading, leveraging multimodal data reasoning and learning analytics based on AI technologies.

In this dissertation, **Part I** has been designed to answer the research question "What are the state-of-the-art advancements and challenges in multimodal data aggregation, feedback design, and implementations in the context of TEL?''. Previously developed multimodal learning systems with multimodal data inputs and feedback outputs for various learning objectives have been investigated to explore the ecosystem of TEL through a scoping review, introduced in *Chapter 2*.

A research question, "What theoretical and technical approaches can be taken to recognize learners' attention regulation in e-reading for higher education?'', has been aimed to be tackled in **Part II**. *Chapter 3* explored various observable behavioral indicators from cognitive science, education, and affective computing that show learners' attention loss, which this dissertation defines as "attention regulation behaviors''. Such behaviors have been utilized for training neural networks based on image-based and video-based samples to accurately predict the moments of learners' attention loss *(Chapter 4)* and their cognitive processes, such as learners' usage of higher-level and lower-level thinking skills, during reading practices *(Chapter 5)*.

Part III answered the research question "How can automatic AI-based feedback in e-

reading assist attention management for higher education learners and further affect their learning outcomes, perceptions, and interactions?''. It consists of three studies to design the feedback components to assist learners' recognized distractions. *Chapter 6* investigated various feedback interfaces (e.g., GUI, Speech-based robot), feedback messages, and traits (e.g., meta-cognitive, emphatic). Also, the effects of the feedback implementations on learners' perceptions and learning outcomes (e.g., knowledge gain, perceived learning experience, perceived interaction with the interface) were investigated. *Chapter 7* studied ideal feedback timing by defining thresholds for the feedback trigger for the recognized attention regulation behaviors. It investigated the normal and abnormal ranges of behavior representations of learners and suggested them as a rationale for deciding the feedback trigger thresholds. Lastly, *Chapter 8* combined the hybrid skeleton-based computer vision neural networks with the adaptive feedback strategies for developing a feedback system for e-reading, combining insights from previous explorations.

All in all, the main research question that extends across the dissertation was "How can a multimodal feedback loop, informed by automatic attention recognition, enhance e-reading experiences for higher education learners?''. This chapter summarizes the main findings from each study as conclusions. The limitations of current work and future suggestions are presented in the following general discussion, closing the dissertation.

9.2 Main findings about the state-of-the-art multimodal learning systems (Part I)

The scoping review on multimodal learning systems has been conducted in *Chapter 2*, based on 30 papers published between 2010 and 2023. In terms of multimodal data collection and aggregation, it was found that most systems rely on the sensor-based approach by combining visual, auditory, and tactile inputs aligned with text-based log data, aiming at real-time, on-site analytics and interventions. Those were primarily for individual learners in K-12 and higher education. Computer-vision approaches allowed more stealth learning analytics, enabling automatic recognition and classification of learning states. Log-based data, a traditional resource for learning analytics, has been observed as a standard layer of multimodal data input, while learners' text-based inputs in learning are becoming a more critical resource to understand learning processes and needs with the current advancement of Large Language Models (LLMs). The questionnaire-based and observation-based data stream was a meaningful layer of the data for both a qualitative understanding of learners and as a ground truth of the machine learning model training, where human expertise is critical. For the feedback design, feedback modalities (e.g., visual, auditory, tactile), characteristics (e.g., spatial, temporal), timing (e.g., real-time, post-hoc), functions (e.g., semantic, intuitive), and types (e.g., graphics, dashboards, text, sound effects, music, video, physical movements, vibrations) have been suggested as critical components. Those feedback components were known to interconnect in design and application processes. Specifically, various intuitive and semantic features have been emphasized in graphics, texts, and dashboards. Auditory feedback has also been favored for its semantic and pheonic features for making sound effects, voice, and music. Physical movements and vibrations provide real-time corrections to learners by stimulating the different sensory channels in learning other than primarily dominant visuals, making the feedback more intuitive. Different data streams and feedback designs were found when systems are geared toward cognitive,

affective, and psychomotor domains in learning, which decides where and how multimodal learning systems are implemented. In conceptual learning, language learning, and medical education, multimodal feedback assisted knowledge delivery and comprehension, while multimodal feedback provided real-time evaluation for clear communication. Also, various combinations of visual, auditory, and tactile feedback, often found in sports and musical education, were used to correct learner behaviors in real time or assist further analytics for reflection.

9.3 Main findings about real-time attention recognition (Part II)

Through the literature study, learners who engage in e-reading were found to experience constant fluctuations in attention, and the current attention management has been highly dependent on learners' self-regulation. In *Chapter 3*, various observable indicators of learners' self-regulation from cognitive science, education, and affective computing were investigated. Learners have shown self-regulatory behaviors to sustain their attention level to overcome their perceived distractions, and specific behaviors were shown as efforts to regain learners' cognitive arousal. According to the literature study, eyebrow movements were found to be a sign of cognitive arousal and re-engagement in the task, while blink flurries are known to be spontaneous efforts to sustain attention and increase wakefulness. Learners were known to use mumble reading as auditory stimulation to decode the complicated text better. Hand movements were found to be efforts to re-engage and refocus on the learning task, while body movements show affective states and cognitive arousals. In this regard, such behaviors were defined as "attention regulation behaviors" in this dissertation, indicators of self-aware distractions and voluntary/semi-voluntary actions to regain focus. Pearson's correlation analysis was conducted using the WEDAR dataset in e-reading, collected for the dissertation, which has the video dataset with second-to-second human annotation of attention regulation behaviors, reaction time to stimuli, and self-reported distractions. As a result, significant correlations were found between attention regulation behaviors and self-reported distractions. Also, internal correlations were found among most attention regulation behaviors, which supported the rationale behind the model training for automatically recognizing attention loss based on behavioral cues. Furthermore, second-to-second image-based recognition based on ResNet with various configurations and video-based recognition with CNN-RNN were compared to find the best-performing model. As a result, the video-based method has shown the best performance compared to the image-based methods (75.70% vs. 69.73% in subject-dependent settings and 68.43% vs. 25.90% in subject-independent settings), presumably due to the temporal layers that it contains, which provides information about movements. Also, we conducted attention regulation behavior recognition using various machine learning models (e.g., kNN, SVN, AdaBoost, MLP), using multiple spans of the behavior inputs (2, 4, 8, 16 seconds). The result represented the best accuracy with MLP with a span of 8 seconds, with 89.41±6.91% accuracy. Also, it achieved the most distinctive classification between attentive and distractive states when applying the t-SNE, which supports the choice of the behavior input spans of 8 seconds. This chapter successfully defined behavioral indicators of attention loss and developed models that assist in the recognition of learners' attentional states.

In *Chapter4*, various behavioral indicators were investigated to find the critical component

in predicting learners' usage of Higher-Order Thinking Skills (HOTs) and Lower-Order Thinking Skills (LOTs) in e-reading, leveraging the explainable AI approach. HOTs are the ability to analyze, synthesize, and evaluate knowledge, while LOTs are the ability to remember, understand, and apply knowledge. Evaluating such qualities was found to use different measures such as writing summarization, essays, and conducting oral presentations to evaluate HOTs to see learners' knowledge synthesis and its applications. On the other hand, LOTS were often evaluated through multiple-choice, true/false, and fill-in-the-blank questions to test the knowledge acquisition and its understanding. In this work, a multiple-choice question score has been used to represent learners' LOTs, while learner summarization via Bidirectional Encoder Representations from Transformers (BERT) has been used to gauge learnes' HOTs. Since HOTs and LOTs do not necessarily correlate, we made a matrix with HOTs and LOTs performances as 3 clusters derived from the k-means clustering and 9 clusters derived from quartile analysis on two axes (i.e., HOTs, LOTs). Various behavioral features based on attention regulation behavior occurrences, behavioral expressiveness, reaction time, and reading speed were used to predict learners' HOTs and LOTs combinations. As a result, The prediction results for thinking skill clusters and each three-level HOTs and LOTs demonstrated robust accuracies ranging from 65.33% to 78.66% across different features and their combinations. Attention regulation behavior has been found to be a consistently strong predictor for all levels of HOTs and LOTs. Individual reading speed was relevant only in predicting thinking skill clusters derived from k-means clustering. Expressiveness played an essential role in predicting thinking skill clusters and LOTs, while individual reaction time was influential in predicting HOTs. This chapter uncovered how learners' behaviors work as robust predictors of learners' usage of thinking skills in e-reading. Instructional designers and educational practitioners often miss key characteristics of learners that represent their intervention needs. *Chapter 5* leveraged unsupervised learning techniques to cluster and derive multi-dimensional characteristics of learners as "personas". Personas were made recognizable through further machine learning model training so the work can be used for learning analytics and feedback provision. Based on the WEDAR dataset, three factual learning outcomes (i.e., knowledge gain, perceived interaction experience, and perceived social presence of the conversational agent as feedback interface) have been chosen as low-dimensional features, while the dataset also contains mid and high-dimensional features. Based on the comparison between automatic and manual feature selection methods, manual feature selection based on the low-dimensional features has recorded the best silhouette score, indicating the best distinctions among clusters. The k-means clustering method has been chosen to segment learners into various groups representing personas. Various unsupervised methods, such as hierarchical, DBSCAN, and spectral clustering, were used to cross-validate the clusters. The Chi-squared test has confirmed that using different unsupervised methods has significant homogeneity in categories of contingency, validating the clustering result. Quartile analysis has been applied to six clusters derived from k-means, making six learner persona types with multi-dimensional characteristics. On top of it, recognition of each persona has been strived based on attention regulation behaviors. In the process, various sampling methods (e.g., instant vs. cumulative, learning-phased vs. time-based) and various classical machine learning methods in its persona prediction were examined. Aside from performing the 6-class classification to predict six persona types,

four persona predictions with the most feedback needs have also been conducted. The SVM classifier has shown relatively stable and robust performances in the time duration-based method, proven as the most appropriate classifier for real-time feedback loop development, having stable 65%-70% accuracies on both instant and cumulative samples for the 4-class classification task. All in all, this chapter provided the feedback adaptation strategy based on the multi-dimensional learner features that represent various learner types effectively and efficiently.

9.4 Main findings about the real-time feedback design (Part III)

In *Chapter 6*, the impact of the conversational agent and its empathic and metacognitive feedback were investigated compared to the GUI-based system. Two e-reading systems, based on the Human-Robot Interaction (HRI) and GUI, were developed respectively. Learners' knowledge gain, perceived interaction experience, and perceived social presence from the HRI-based and GUI-based interfaces were compared. HRI-based feedback was implemented via a Furhat robot with a human face, facial expressions, and speech-based input and output, while GUI-based feedback had a pop-up with yes or no buttons for user inputs. During the e-reading, quiz questions were given on the screen at the end of each subtopic. At the same time, only the HRI-based system provided emphatic and metacognitive feedback through the robot speech regarding learners' answers and correctness. The result indicates that comparatively higher knowledge gain and perceived knowledge gain were achieved with the HRI feedback. Also, the HRI feedback was emotionally favored for being inventive, creative, innovative, captivating, and challenging, which is related to hedonic-stimulation quality. In the meanwhile, the overall social presence has been highly evaluated with the HRI-based system compared to the GUI-based system, especially having significance in the average social presence measure and co-presence measure, indicating that learners appreciated the physical presence of the robot interface and perceived the HRI as "social beings" in the interaction process. The accuracy of recognizing the attention regulation behaviors based on various deep learning models, such as CNN-RNN, CNN-LSTM, and CNN-Transformer, have been compared, resulting in the CNN-RNN as the best-performing model with 72.97% accuracy. Furthermore, attention regulation behaviors were further used to predict learners' knowledge gain, perceived interaction experience, and perceived social presence. In predicting the 3-level knowledge gain with high-mid-low distinctions, the SVM classifier has performed 74.29%. For predicting the perceived interaction experience, the random forest has shown the highest performance in predicting the overall experience, pragmatic quality, raw hedonic-identity, raw hedonic-stimulation, and raw and normalized attractiveness scores, ranging across 70% to 92.5% accuracies. This chapter has proven that attention regulation behavior is a robust predictor of learners' attentional, cognitive, and psychological states in e-reading and interaction with feedback. *Chapter 7* aligned four stages of system architecture in developing the real-time feedback loop for attentive e-reading. The framework is an iterative loop of four stages: 1) capturing learners' behaviors in e-reading via webcam, 2) reasoning and recognizing the learners' distractions based on the neural networks leveraged by computer vision, 3) intervention (screen blur) based on the pre-designed feedback rules, and 4) triggering the positive cognitive and behavioral changes of learners, which we fundamentally aimed as extended attention span and fewer distractions. With iterations, feedback triggers behavioral changes

in learners and simultaneously works as the next round of input for attention recognition. This chapter also defines the reasoning behind having the screen-blur feedback. The fundamental function of the blur feedback was to trigger the learner's arousal while not bothering the primary e-reading task performances while introducing a non-intrusive but alarming intervention commonly referred to as a challenge in real-time intervention design. At the same time, the reaction time to the blurred stimuli can simultaneously work as data points for future feedback loops because it provides meaningful insights into learners' arousal. Lastly, the chapter also investigates the normal and abnormal ranges of attention regulation behaviors, which can be further utilized for the feedback timing design. Based on the quartile analysis of the duration of each attention regulation behavior occurrence in the WEDAR dataset, low-mid-high ranges of attention regulation behavior thresholds have been derived, which defines the normal and abnormal behavior ranges of learning behaviors. This chapter has contributed to the system architecture design, the type of feedback, and the rationale behind the feedback timing design.

Previously found insights have been comprehensively combined in *Chapter 8* for developing a prototype. Implementing a real-time feedback loop that helps the attention span with fewer distractions has been the fundamental focus of the e-reading system with AI-based feedback. From the literature study, it has been found that skeleton-based behavior recognition has advantages in lower computational requirements and broader applicability without environmental constraints compared to the video-based method. Multi-dimensional matrices of facial and body landmarks have been derived based on the MediaPipe framework to leverage the advantage of having temporal layers in data, an advantage of the video-based method. Using the WEDAR dataset, binary and multi-class models have been developed to recognize attention regulation behaviors effectively. At the same time, a model fusion of binary and multi-class classification has been applied to minimize the false-positive recognition of distractions and subsequent feedback triggers, which is critical for learners' perceived trustworthiness toward the feedback and the system. The primary feedback rule has been that once the attention regulation behaviors are recognized for more than the time defined in the predefined thresholds, the blur was applied to the screen, and learners were gently reminded of their distractions. At the same time, clicking on the button to deactivate the blur screen was meant to cut the tendency for distraction. Feedback design took an adaptive approach in that the threshold was adapted on every page based on learners' behaviors on the previous page. Based on the behavioral expressiveness and reaction speed to the blur, which indicated the distractions and arousal in behaviors, three learner clusters were derived from k-means clustering. The blur was designed to be given more frequently for learners with more intervention needs. Several aspects of the prototype were evaluated on various experimental conditions: distractions, learning performances, and learners' perceptions were evaluated on 1) a control group with randomized feedback timing, 2) a treatment group with the AI-based screen blur feedback without explanation about the feedback mechanism, and 3) a treatment group with AI-based screen blur feedback with explanation about the feedback mechanism. Also, 4) a baseline group representing learners' preliminary perceptions about e-reading experiences before implementing measures have taken place. Both treatment conditions, with or without explanation about the feedback mechanism, showed fewer reports of distractions with a longer attention span than the control group, with statistical significance derived from ANOVA analysis. In knowledge gain,

treatment conditions with feedback explainability performed better in both multiple-choice and summarization tasks. Perceived interaction experience, measured by the AttrakDiff questionnaire, treatment conditions with feedback explainability has shown the best evaluation on the overall AttrakDiff score, Hedonic-Identity, Hedonic-Stimulation, and Attractiveness. At the same time, pragmatic quality was evaluated better in the baseline group, representing the direction of future improvement. This chapter validated the effectiveness of AI-based automatic feedback in decreasing learners' distractions and assisting more attentive e-reading, with insights into the importance of feedback explainability and aesthetic attractiveness in learners' perceptions and experiences.

9.5 General Discussion

As articulated in the scoping review on multimodal learning systems in *Chapter 2*, generalization and personalization of learning analytics and feedback design are significant challenges in closing the feedback loop. That is because systems should be general enough to accommodate the learning needs of general learners by adequately recognizing and responding to the targeted actions. At the same time, specific personalization needs should be effectively tackled without harming generalization. In this thesis, data-driven learner segmentation approaches have been applied to address such issues so that feedback can adapt to specific persona types that represent particular learners *(Chapter 5)*. In the process, combining the human experts' insights and machine learning techniques was found critical for complementing the limitations of machine reasoning. For instance, structuring the learning data for the first round feature vector design and validating the clusters has been essential for data-driven persona development, where human expertise and domain knowledge can benefit the quality of the result. Also, the semantical explainability of models *(Chapter 4)* and the design of AI-based feedback *(Chapter 8)* are still highly dependent on human expertise regardless of the advancement of AI, emphasizing the complementary relationship between human expertise and AI technologies.

Another point has been that general multimodal learning systems lack overarching frameworks with higher-level learning objectives. Due to the domain-specific and context-specific nature of the learning analytics and intervention design, most existing frameworks need further support from the expanded models or frameworks. This dissertation defined attention regulation behaviors as target behaviors that provide clues for feedback generation. It also specified behaviors in *Chapter 3* that could successfully correlate with various learning outcomes. Further developed machine learning models were combined with diverse feedback interface with empathic and meta-cognitive prompts *(Chapter 6)* with the refinement of feedback timing *(Chapter 7)*. However, a more theoretical framework for behavior analysis and content-level instructional design in e-reading would strengthen a holistic understanding of feedback on learners' behavioral, cognitive, and psychological changes.

Lastly, closing the feedback loops in the context of Multimodal Learning Analytics (MMLA) creates various challenges and opportunities. Though this dissertation successfully closed the feedback loop based on attention regulation behavior recognition with the skeleton-based framework and adaptive feedback strategies (Chapter 8), there is still room for improvement. For instance, though an automatic AI-feedback system has brought longer attention spans of learners and less frequent distractions regardless of the explainability of

the feedback, the evaluation of learners' perceptions varied significantly depending on the explainability. It represents "how" the feedback is explained, which critically affects the perception of AI-based automatic feedback and their learning experiences. As revealed, different feedback interfaces bring different perceptions and dynamics between learners and the system even with the same content (Chapter 6). Thus, composing a feedback ecosystem with other types of real-time feedback (e.g., VR/AR-based interface) could change the whole learning dynamics and interaction between learners and the system. Also, as represented in *Chapter 8*, the system's aesthetical quality affects the hedonic identity and attractiveness, which highly affects the general perceptions of the system. Though only the primary wireframes of the Graphic User Interface (GUI) have been suggested in this dissertation, there is much to be explored and improved based on further enhancements with considerations of various User Interface (UI) components.

Samenvatting
(Summary in Dutch)

9.6 Samenvatting

Met het wijdverbreide gebruik van digitale apparaten en platforms wordt e-lezen, in dit proefschrift digitaal lezen op een computerscherm genoemd, steeds gewoner in het formele en informele onderwijs. Vooral in het hoger onderwijs is de impact van e-lezen groter, omdat van leerlingen wordt verwacht dat ze de kennis op basis van hun zelfstandig lezen registreren, begrijpen, onthouden, reconstrueren en toepassen als onderdeel van het reguliere onderwijs. Het heeft een rechtstreekse invloed op de leerresultaten, de zelfredzaamheid van de leerlingen en hun leerprestaties. Traditioneel werd voor e-learning gestreefd naar loggegevens van MOOCS voor post-hoc analyse en interventie, terwijl Technology-Enhanced Learning (TEL) benaderingen, gebruikmakend van sensing en machine learning technologieën, plaatsvonden voor real-time leeranalyse en leerondersteuning. Echter, ondanks het belang van e-lezen in het hoger onderwijs en technologische vooruitgang die naadloze TEL-toepassing mogelijk maakt, lijken real-time interventieontwerp en -implementatie voor e-lezen schaars. In dit verband verkent dit proefschrift de mogelijkheden van een real-time feedbacklus bij e-lezen, gebruikmakend van multimodale dataredenering en leeranalyse op basis van AI-technologieën. Al met al was de belangrijkste onderzoeksvraag van het proefschrift "Hoe kan een multimodale feedbacklus, geïnformeerd door automatische aandachtsherkenning, de e-leeservaring van studenten in het hoger onderwijs verbeteren?".

In dit proefschrift, **Part I** is ontworpen om de onderzoeksvraag te beantwoorden "Wat zijn de state-of-the-art ontwikkelingen en uitdagingen in multimodale data aggregatie, feedback ontwerp en implementaties in de context van TEL". Eerder ontwikkelde multimodale leersystemen met multimodale data inputs en feedback outputs voor verschillende leerdoelen zijn onderzocht om het ecosysteem van TEL te verkennen door middel van een scoping review geïntroduceerd in *Chapter 2*.

De onderzoeksvraag, "Welke theoretische en technische benaderingen kunnen worden gebruikt om de aandachtsregulatie van lerenden bij e-lezen voor het hoger onderwijs te herkennen?", is behandeld in **Part II**. *Chapter 3* onderzocht verschillende waarneembare gedragsindicatoren uit de cognitiewetenschappen, het onderwijs en affectieve informatica die het aandachtsverlies van leerlingen aantonen, die in dit proefschrift worden gedefinieerd als "aandachtsregulatiegedragingen". Dergelijke gedragingen zijn gebruikt voor het trainen van neurale netwerken op basis van afbeelding en video voorbeelden om nauwkeurig de momenten van aandachtsverlies (*Chapter 4*) en cognitieve processen, zoals het gebruik van hogere en lagere denkvaardigheden, tijdens het lezen te voorspellen bij leerlingen (*Chapter 5*).

9

9.7 Algemene Discussie

Zoals verwoord in de scoping review over multimodale leersystemen in *Chapter 2*, zijn generalisatie en personalisatie van learning analytics en feedbackontwerp belangrijke uitdagingen bij het sluiten van de feedbacklus. Systemen moeten namelijk algemeen genoeg zijn om tegemoet te komen aan de leerbehoeften van algemene leerlingen door de gerichte acties adequaat te herkennen en erop te reageren. Tegelijkertijd moeten specifieke personalisatiebehoeften effectief worden aangepakt zonder de generalisatie te schaden. In dit proefschrift zijn gegevensgestuurde leerlingsegmentatiebenaderingen toegepast om dergelijke problemen aan te pakken, zodat de feedback loop zich kan aanpassen aan specifieke personatypes die bepaalde leerlingen vertegenwoordigen (*Chapter 5*). Tijdens het proces bleek dat het combineren van de inzichten van menselijke experts en machine-learningtechnieken cruciaal is voor het aanvullen van de beperkingen van machine-reasoning. Het structureren van de leerdata voor de eerste ronde van het featurevectorontwerp en het valideren van de clusters was bijvoorbeeld essentieel voor datagestuurde persona-ontwikkeling, waarbij menselijke expertise en domeinkennis de kwaliteit van het resultaat ten goede kunnen komen. Ook de semantische verklaarbaarheid van modellen (*Chapter 4*) en het ontwerp van AI-gebaseerde feedback (*Chapter 8*) zijn nog steeds sterk afhankelijk van menselijke expertise, ongeacht de vooruitgang van AI, wat de complementaire relatie tussen menselijke expertise en AI-technologieën benadrukt.

Een ander punt is dat algemene multimodale leersystemen geen overkoepelende kaders hebben met leerdoelen op een hoger niveau. Vanwege de domeinspecifieke en contextspecifieke aard van de leeranalyse en het interventieontwerp, hebben de meeste bestaande frameworks verdere ondersteuning nodig van de uitgebreide modellen of frameworks. Dit proefschrift definieerde aandacht regulerend gedrag als doelgedrag dat aanwijzingen geeft voor het genereren van feedback. Het specificeerde ook gedragingen in *Chapter 3* die succesvol konden correleren met verschillende leerresultaten. Verder ontwikkelde machine-learning modellen werden gecombineerd met diverse feedback interfaces met empathischnadrukkelijke en meta-cognitieve prompts (*Chapter 6*) met de verfijning van feedback timing (*Chapter 7*). Een meer theoretisch kader voor gedragsanalyse en instructieontwerp op inhoudsniveau bij e-lezen zou echter een meer holistisch begrip van feedback op gedrags-, cognitieve en psychologische veranderingen bij leerlingen versterken.

Tot slot creëert het sluiten van de feedbacklussen in de context van Multimodal Learning Analytics (MMLA) verschillende uitdagingen en kansen. Hoewel dit proefschrift met succes de feedbacklus heeft gesloten op basis van aandachtsregulatie gedragsherkenning met het skelet-gebaseerde framework en adaptieve feedbackstrategieën (Chapter 8), is er nog ruimte voor verbetering. Hoewel een automatisch AI-feedbacksysteem bijvoorbeeld heeft geleid tot een langere aandachtsspanne van lerenden en minder frequente afleiding ongeacht de uitlegbaarheid van de feedback, varieerde de evaluatie van de perceptie van lerenden aanzienlijk afhankelijk van de uitlegbaarheid. Het is "hoe" de feedback wordt uitgelegd, wat de perceptie van AI-gebaseerde automatische feedback en de leerervaringen kritisch beïnvloedt. Zoals aangetoond, zorgen verschillende feedbackinterfaces voor verschillende percepties en dynamieken tussen lerenden en het systeem, zelfs bij dezelfde inhoud (*Chapter 6*). Door een feedbackecosysteem samen te stellen met andere soorten realtime feedback (bijvoorbeeld VR/AR-gebaseerde interfaces) kan de hele leerdynamiek en interactie tussen lerenden en het systeem dus veranderen. Zoals weergegeven in *Chapter 8*, beïnvloedt

de esthetische kwaliteit van het systeem de hedonische identiteit en aantrekkelijkheid, wat de algemene perceptie van het systeem sterk beïnvloedt. Hoewel alleen de primaire grafische gebruikersinterface (GUI) kaders zijn gebruikt in dit proefschrift, kan er nog veel worden onderzocht en verbeterd op basis van verdere overwegingen van verschillende gebruikersinterface (UI) componenten.

9

학위논문 국문 초록
(Summary in Korean)

9.8 요약

디지털 기기 및 플랫폼 사용 저변의 확대에 따라 스크린 기반 읽기 (e-reading)가 공식 및 비공식 교육에서 더욱 보편화되고 있다. 특히 e-reading의 영향력은 학습자의 독립적 읽기를 기반으로 한 지식 습득, 이해, 기억, 재구성 및 활용을 정규 교육과정의 일부로 포괄하는 고등 교육에서 더욱 두드러진다. 때문에 e-reading은 학습 결과, 학습자의 자기 효용감 및 학습 성공에 매우 중대한 영향을 끼친다. 전통적인 이러닝 학습 지원 (e-learning support)은 1) MOOC의 로그 데이터를 활용한 사후 분석 및 피드백, 그리고 2) 센싱 및 머신러닝을 활용한 기술 기반 학습 (Technology-Enhanced Learning; TEL)의 도입 및 실시간 학습 분석 (Learning Analytics; LA)이라는 두 가지 큰 흐름으로 요약된다. 하지만 고등 교육에서의 e-reading의 중요성과 이를 둘러싼 실시간 기술 기반 학습 인프라의 비약적 발전에도 불구하고 e-reading을 중심으로 한 실시간 피드백 시스템의 디자인 및 기술적 구현은 이루어지지 않았다. 이러한 맥락에서 본 학위논문에서는 AI를 기반으로 한 멀티모달 추론 (Multimodal Reasoning) 및 학습 분석을 중심으로 e-reading을 위한 실시간 피드백 프레임워크를 제안한다. 본 학위논문을 관통하는 주요 연구 질문은 'AI 기반의 멀티모달 인터렉션 루프 (loop)가 고등 교육 학습자의 e-reading 내 주의력에 어떻게 도움을 줄 수 있는가?' 이다.

Part I에서는 '기술 기반 학습 환경에 적용 가능한 멀티모달 데이터 활용, 피드백 디자인, 그리고 시스템 개발을 위한 선행 과제는 무엇인가?' 라는 연구 질문에 대해 탐구하였다. 이를 위해 Chapter 2에서는 주제범위 문헌고찰 (Scoping Review) 방법을 통해 '기술 기반 학습에서 멀티모달 데이터 수집, 피드백 도입이 다양한 학습 목표를 성취하기 위해 어떻게 적용되었는가?' 라는 연구 질문을 가지고 멀티모달 데이터 수집 및 융합, 피드백 디자인, 그리고 교육 환경 내 적용이라는 세가지 하위 주제에 접근하였다.

Part II에서는 '고등교육 내 e-reading에서 학습자의 주의력을 시스템 기반으로 인식하기 위한 이론적, 기술적 접근은 무엇인가?'라는 연구 질문에 답하고자 하였다. 이에 따라 다양한 인지 과학 (Cognitive Science), 교육 (Education), 감성 컴퓨팅 (Affective Computing) 지표에 대한 탐구를 기반으로 학습자의 주의력 상실을 예측하기 위한 주요 지표를 도출하고, 이를 학습자들의 e-reading 시 나타나는 '주의력 조절 행동 (Attention Regulation Behaviors)' 으로 정의하였다 (Chapter 3). 또한 해당 행동 지표들을 기반으로 이미지 및 비디오 기반 신경망 (Neural Network)을 학습하여 읽기 과제 수행 중 학습자의 주의력 상실 (Chapter 4) 및 고수준/저수준 사고 기술의 활용 (Higher-Order and Lower-Order Thinking Skills; HOTs and LOTs)과 같은 e-reading 내 학습자의 인지 수준(Chapter 5)을 예측하고자 하였다.

Part III에서는 'e-reading에서 인공지능 기반의 실시간 피드백이 학습자의 주의력 관리 및 학습 결과, 사용자 경험 및 시스템과의 인터렉션에 어떤 영향을 끼치는가?' 라는 연구 질문에 관해 고찰하였다. 이를 위해 해당 연구에서는 Part II에서 다루어진 학습자들의 주의력 조절 행동과 다양한 피드백 구성 요소 (Feedback Components)들 간 상관관계에 집중하였다. Chapter 6에서는 다양한 피드백 인터페이스(예: GUI, 음성 대화 로봇(Speech-based Robot)), 메시지 (Content), 특성 (예: 메타인지적 (Meta-cognitive), 공감하는 (Empathic)), 타이밍에 대해 고찰하였다. 이를 기반으로 다양한 피드백 구성 요소가 학습자의 인식과 학습 결과(예: 지식 습득 (Knowledge Gain), 학습 경험 (Perceived Learning Experience), 인터페이스와의 상호 작용 (Perceived Interaction Experience))에 어떠한 영향을 끼치는지 탐구하였다. Chapter 7에서는 학습자가 e-reading에서 보이는 주의력 상실 지표 표현의 정상 범위와 비정상 범위를 정의하여 피드백 작동 (trigger)의 임계값을 제시하는 피드백 타이밍 프레임워크를 제안하였다. 마지막으로 Chapter 8에서는 스켈레톤 기반의 (Skeleton-based) 컴퓨터 비전 (Computer Vision) 신경망 모델 (Neural Networks)과 GUI 기반의 적응형

(Adaptive) 피드백 디자인을 개발하여 AI 피드백이 e-reading 학습자의 행동, 인지, 지각, 인터렉션에 끼치는 영향력에 대해 고찰하였다

9.9 총론

Chapter 2의 멀티모달 학습 시스템을 위한 주제범위 문헌고찰을 통해 분석된 바와 같이, 학습 분석 및 피드백 설계 시의 알고리듬 일반화 (Generalization) 및 개인화 (Personalization)는 피드백 루프를 완결하는 데 필수적으로 고려 되어야 하는 요소이다. 학습 분석 및 피드백 시스템의 도입 목적은 시스템이 목표로 하는 사용자의 행동을 정확히 감지하고, 이를 통해 보편적인 학습자의 학습 요구 (Learning Needs)에 대응할 수 있을 만큼 충분히 일반적이어야 하기 때문이다. 한편으로, 알고리듬은 일반화 요소를 저해하지 않는 수준에서 특정 학습자들의 개인적인 니즈 또한 효과적으로 해결 될 수 있도록 디자인 되어야 한다. 본 최종 학위논문에서는 이러한 문제를 해결하기 위해 비지도 학습 (Unsupervised Learning) 기법을 통한 데이터 기반 학습자 세분화 (Segmentation) 접근법을 적용하여, 손쉽게 특정 학습자군을 대표하는 '페르소나' 유형을 도출하고, 이에 맞게 피드백을 제공할 수 있도록 했다 (Chapter 5). 이 과정에서 기계 추론 (Machine Reasoning)의 한계를 보완하기 위해 인간 전문가의 인사이트와 머신러닝 기술의 결합이 강조되었다. 예를 들어, 비지도 학습의 1차적 특징 벡터 (Feature Vector) 설계를 위해 학습 데이터를 구조화하고, 추후 도출 된 클러스터를 검증하는 작업은 데이터 기반 페르소나 개발 (Data-driven Persona Development)에 필수적인 과정으로, 이 과정에서 인간의 전문성 (Expertise)과 도메인 지식 (Domain Knowledge)은 클러스터링 품질에 큰 영향을 끼친다. 또한 기계학습 모델의 의미론적 (Semantic) 설명 가능성 (Interpretability) (Chapter 4)과 AI 기반 피드백 도입에 (Chapter 8) 필수적인 인간의 지능 (Human Intelligence)은 인간과 AI 기술의 상호보완적 관계를 기반으로 한 융합 지능 (Hybrid Intelligence)의 중요성을 강조한다.

또한 해당 학위논문에서는 일반적인 멀티모달 학습 시스템 개발을 위한 고수준 (High-level) 학습 목표 (Learning Objectives)가 포함된 포괄적 (Overarching) 프레임워크가 부족하다는 점에 주목하였다. 도메인 별 (Domain-specific), 상황 별 (Context-specific) 특성으로 인해 대부분의 학습 분석 및 피드백 설계에서 기존 프레임워크는 모델 확장 또는 타 프레임워크의 추가적 연계를 필요로 한다. 본 학위논문에서는 주의력 조절 행동을 피드백 생성의 단서가 되는 목표 행동으로 정의하였으며 (Chapter 3), 또한 Chapter 4에서는 성공적 학습 결과와 긴밀하게 연관되는 학습 행동을 '설명가능한 인공지능 (Explainable AI)' 프레임워크를 활용해 연구하였다. 이를 기반으로 추후 개발된 머신러닝 모델은 피드백 타이밍의 세분화 (Chapter 8)와 함께 로봇 인터페이스를 기반으로 한 공감 및 메타인지 프롬프트 (Chapter 6)의 영향력을 주제로 한다. 넓은 연구 범위를 고려할 때, 행동 기반 분석 및 피드백 컨텐츠를 위한 총체적 프레임워크의 개발은 학습자의 행동적, 인지적, 심리학적 변화에 대한 보다 유기적이고 심층적 이해를 가능케 할 것으로 보인다.

마지막으로 해당 학위논문에서는 멀티모달 학습 분석 (Multimodal Learning Analytics; MMLA)의 맥락에서 피드백 루프를 완결 짓기 위해 해결되어야 할 어려움 및 향후 연구 방향성을 공유한다. 본 연구에서는 스켈레톤 기반 행동 감지 프레임워크와 적응형 피드백 전략 (Chapter 8)을 결합하는 방식으로 피드백 루프를 매듭지었지만, 여전히 개선점은 존재한다. 예를 들어, 도입된 인공지능 기반 (AI-based)의 자동 (Automatic) 피드백 시스템은 피드백의 설명 가능성 (Explainability)에 관계없이 학습자의 주의 집중 시간 (Attention Span)을 길어지게 하는 한편 주의력 상실 (Distraction)의 빈도를 감소케 했지만, 학습자의 시스템에 대한 인지적 (Perceived) 평가는 피드백의 설명 가능성에 따라 크게 달라졌다. 이는 피드백이 학습자 주의 집중에 객관적 도움을 줌에도 피드백이 '어떻게' 설명 되는지에 따라 AI 기반 자동 피드백에 대한 학습자의 인식 및 학습 경험이 크게 달라짐을 의미한다. 앞서 살펴본 바와 같이, 피드백 인터페이스에 따라 같은 콘텐츠를 제공함에도 학습자의 시스템에 대한 인식 및 인터페이스와의 역학관계 (Dynamics)는 크게 달라졌다 (Chapter 6). 따라서 다양한 유형의 실시간 피드백 (예: VR/AR 기반 인터페이스)의 도입을 통한 시스템 생태계 (Ecosystem)의 다양화는 e-reading의 양상 뿐 아니라 학습자와 인터페이스 간 역학관계 및 상호 작용에 큰 변화를 불러올 것이다. 또한, Chapter 8에서 살펴본 바와 같이 시스템의 미적 품질 (Aesthetic Quality)은 시스템의 정체성과 관련된 쾌락적 특질 (Hedonic-Identity Quality)과 매력도 (Attractiveness)에 영향을 미치며, 이는 시스템에 대한 일반적인 인식에도 큰 영향을 선사한다는 점이 확인되었다. 이 학위논문에서는 그래픽

사용자 인터페이스 (GUI)를 구성하는 기본적이고 기능적인 (Functional) 와이어 프레임 (wireframe)만이 제시되었지만, 다양한 사용자 인터페이스 (User Interface; UI) 및 구성 요소를 고려한 시스템의 추가적 미적 보완은 시스템에 대한 전반적 인식에 한층 더 긍정적인 변화를 가져올 것으로 전망한다.

9

Bibliography

References

[1] Fakher Jaoua, Hussein M Almurad, Ibrahim A Elshaer, and Elsayed S Mohamed. E-learning success model in the context of covid-19 pandemic in higher educational institutions. *International Journal of Environmental Research and Public Health*, 19(5):2865, 2022.

[2] Donatella Persico, Stefania Manca, and Francesca Pozzi. Adapting the technology acceptance model to evaluate the innovative potential of e-learning systems. *Computers in Human Behavior*, 30:614–622, 2014.

[3] Huiyong Li, Rwitajit Majumdar, Mei-Rong Alice Chen, Yuanyuan Yang, and Hiroaki Ogata. Analysis of self-directed learning ability, reading outcomes, and personalized planning behavior for self-directed extensive reading. *Interactive Learning Environments*, pages 1–20, 2021.

[4] Ángel Garralda Ortega, Abel Hon Man Cheung, and Michelle Yuen Shan Fong. Developing e-reading pedagogies informed by research. *Journal of World Languages*, (0), 2022.

[5] Jane Oakhill, Kate Cain, and Carsten Elbro. Reading comprehension and reading comprehension difficulties. *Reading development and difficulties: Bridging the gap between research and practice*, pages 83–115, 2019.

[6] M Horton-Ramos. Reading in the digitized era: Analyzing esl graduate students'e-reading habit. *Asian EFL*, 27(1):67–85, 2020.

[7] Jan D Vermunt. Metacognitive, cognitive and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher education*, 31(1):25–50, 1996.

[8] Jordan M Barkley. Reading education: is self-efficacy important? *Reading Improvement*, 43(4):194–211, 2006.

[9] Mart Van Dinther, Filip Dochy, and Mien Segers. Factors affecting students' self-efficacy in higher education. *Educational research review*, 6(2):95–108, 2011.

[10] Jaclyn Broadbent and Walter L Poon. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The internet and higher education*, 27:1–13, 2015.

[11] Yoon Lee. Flower: Feedback loop for group work supporter. *The International Learning Analytics and Knowledge Conference (LAK demo session)*, 2020.

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

[13] Dimitrios Pnevmatikos, Panagiota Christodoulou, and Nikolaos Fachantidis. 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*, pages 153–165. Springer, 2018.

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

[15] Michae Xuelin Huang, Jiajia Li, Grace Ngai, Hong Va Leong, and Andreas Bulling. Moment-to-moment detection of internal thought from eye vergence behaviour. *arXiv e-prints*, pages arXiv–1901, 2019.

[16] Gregory JH Colflesh and Andrew RA Conway. Individual differences in working memory capacity and divided attention in dichotic listening. *Psychonomic bulletin & review*, 14(4):699–703, 2007.

[17] Selene Mota and Rosalind W Picard. Automated posture analysis for detecting learner's interest level. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, volume 5, pages 49–49. IEEE, 2003.

[18] Sonja Walcher, Christof Körner, and Mathias Benedek. Looking for ideas: Eye behavior during goal-directed internally focused cognition. *Consciousness and cognition*, 53:165–175, 2017.

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

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

[21] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 305–312, 2011.

[22] Haoyu Chen, Esther Tan, Yoon Lee, Sambit Praharaj, Marcus Specht, and Guoying Zhao. Developing ai into explanatory supporting models: An explanation-visualized deep learning prototype. In *The International Conference of Learning Science (ICLS)*, 2020.

[23] Sidney D'Mello, Tanner Jackson, Scotty Craig, Brent Morgan, P Chipman, Holly White, Natalie Person, Barry Kort, R El Kaliouby, Rosalind Picard, et al. Autotutor

detects and responds to learners affective and cognitive states. In *Workshop on emotional and cognitive issues at the international conference on intelligent tutoring systems*, pages 306–308, 2008.

[24] Sidney D'Mello and Art Graesser. Mining bodily patterns of affective experience during learning. In *Educational data mining 2010*, 2010.

[25] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 12(2):505–523, 2018.

[26] Yoon Lee, Gosia Migut, and Marcus Specht. Role of multimodal systems in technology enhanced learning (TEL): A scoping review (in press). In *18th European Conference on Technology Enhanced Learning (EC-TEL23)*, 2023.

[27] Yoon Lee, Bibeg Limbu, Zoltan Rusak, and Marcus Specht. Role of multimodal learning systems in technology-enhanced learning (tel): A scoping review. In *Responsive and Sustainable Educational Futures*, 2023.

[28] Temitayo Deborah Oyedotun. Sudden change of pedagogy in education driven by covid-19: Perspectives and evaluation from a developing country. *Research in Globalization*, 2:100029, 2020.

[29] Sanna Järvelä, Jonna Malmberg, Eetu Haataja, Marta Sobocinski, and Paul A Kirschner. What multimodal data can tell us about the students' regulation of their learning process? *Learning and instruction*, 72:101203, 2021.

[30] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[31] Yoon Lee and Marcus Specht. 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)*. ACM, 2023.

[32] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachsler. From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, 34(4):338–349, 2018.

[33] Elena V Frolova, Olga V Rogach, and Tatyana M Ryabova. Digitalization of education in modern scientific discourse: New trends and risks analysis. *European journal of contemporary education*, 9(2):313–336, 2020.

[34] Cristobal Romero and Sebastian Ventura. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1355, 2020.

[35] Abelardo Pardo, Jelena Jovanovic, Shane Dawson, Dragan Gašević, and Negin Mirri-ahi. Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, 50(1):128–138, 2019.

[36] Hilary Arksey and Lisa O'Malley. Scoping studies: towards a methodological frame-work. *International journal of social research methodology*, 8(1):19–32, 2005.

[37] B. Cope, M. Kalantzis, and New London Group. *Multiliteracies: Literacy Learning and the Design of Social Futures*. Literacies (Routledge). Routledge, 2000.

[38] Yoon Lee, Haoyu Chen, Guoying Zhao, and Marcus Specht. 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)*, pages 319–328. ACM, 2022.

[39] Gualtiero Volpe, Ksenia Kolykhalova, Erica Volta, Simone Ghisio, George Waddell, Paolo Alborno, Stefano Piana, Corrado Canepa, and Rafael Ramirez-Melendez. A multimodal corpus for technology-enhanced learning of violin playing. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, pages 1–5, 2017.

[40] Ana Cristina Pires, Ewelina Bakala, Fernando González-Perilli, Gustavo Sansone, Bruno Fleischer, Sebastián Marichal, and Tiago Guerreiro. Learning maths with a tangible user interface: Lessons learned through participatory design with children with visual impairments and their educators. *International Journal of Child-Computer Interaction*, 32:100382, 2022.

[41] Mathieu Chollet, Pranav Ghate, Catherine Neubauer, and Stefan Scherer. Influence of individual differences when training public speaking with virtual audiences. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 1–7, 2018.

[42] Federico Domínguez and Katherine Chiluiza. Towards a distributed framework to analyze multimodal data. In *Proc. of Workshop Cross-LAK–held at LAK '16*, pages 52–57, 2016.

[43] Xavier Ochoa, Federico Domínguez, Bruno Guamán, Ricardo Maya, Gabriel Falcones, and Jaime Castells. The rap system: Automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 360–364, 2018.

[44] Raffaello Brondi, Massimo Satler, Carlo Alberto Avizzano, and Paolo Tripicchio. A multimodal learning system for handwriting movements. In *2014 International Conference on Intelligent Environments*, pages 256–259. IEEE, 2014.

[45] Grega Jakus, Kristina Stojmenova, Sašo Tomažič, and Jaka Sodnik. A system for efficient motor learning using multimodal augmented feedback. *Multimedia Tools and Applications*, 76(20):20409–20421, 2017.

[46] Renlong Ai, Marcela Charfuelan, Walter Kasper, Tina Klüwer, Hans Uszkoreit, Feiyu Xu, Sandra Gasber, and Philip Gienandt. Sprinter: Language technologies for interactive and multimedia language learning. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2733–2738, 2014.

[47] Nordine Sebkhi, Dhyey Desai, Mohammad Islam, Jun Lu, Kimberly Wilson, and Maysam Ghovanloo. Multimodal speech capture system for speech rehabilitation and learning. *IEEE Transactions on Biomedical Engineering*, 64(11):2639–2649, 2017.

[48] Akira Maezawa and Kazuhiko Yamamoto. Muens: A multimodal human-machine music ensemble for live concert performance. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4290–4301, 2017.

[49] Jules Françoise, Sarah Fdili Alaoui, Thecla Schiphorst, and Frédéric Bevilacqua. Vocalizing dance movement for interactive sonification of laban effort factors. In *Proceedings of the 2014 conference on Designing interactive systems*, pages 1079–1082, 2014.

[50] Maria João Silva. Children using electronic sensors to create and use knowledge on environmental health. *First Monday*, 2020.

[51] Bibeg Hang Limbu, Halszka Jarodzka, Roland Klemke, and Marcus Specht. Can you ink while you blink? assessing mental effort in a sensor-based calligraphy trainer. *Sensors*, 19(14):3244, 2019.

[52] Mutlu Cukurova, Michail Giannakos, and Roberto Martinez-Maldonado. The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology*, 51(5):1441–1449, 2020.

[53] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, 27:1245–1249, 2020.

[54] Ali Asadipour, Kurt Debattista, and Alan Chalmers. Visuohaptic augmented feedback for enhancing motor skills acquisition. *The Visual Computer*, 33(4):401–411, 2017.

[55] Ali Asadipour, Kurt Debattista, Vinod Patel, and Alan Chalmers. A technology-aided multi-modal training approach to assist abdominal palpation training and its assessment in medical education. *International Journal of Human-Computer Studies*, 137:102394, 2020.

[56] Cristian Bernareggi, Dragan Ahmetovic, and Sergio Mascetti. μgraph: Haptic exploration and editing of 3d chemical diagrams. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 312–317, 2019.

[57] Bosede Iyiade Edwards, Kevin S Bielawski, Rui Prada, and Adrian David Cheok. Haptic virtual reality and immersive learning for enhanced organic chemistry instruction. *Virtual Reality*, 23(4):363–373, 2019.

[58] Min Fan, Alissa N Antle, and Emily S Cramer. Design rationale: Opportunities and recommendations for tangible reading systems for children. In *Proceedings of the The 15th International Conference on Interaction Design and Children*, pages 101–112, 2016.

[59] Bri Hightower, Silvia Lovato, Jordan Davison, Ellen Wartella, and Anne Marie Piper. Haptic explorers: Supporting science journaling through mobile haptic feedback displays. *International Journal of Human-Computer Studies*, 122:103–112, 2019.

[60] Jun Lee, WonJong Kim, Anna Seo, JiSun Jun, SeungYeon Lee, Jee-In Kim, KiDong Eom, Muwook Pyeon, and Hanku Lee. An intravenous injection simulator using augmented reality for veterinary education and its evaluation. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 31–34, 2012.

[61] Korman Maria, Alessandro Filippeschi, Emanuele Ruffaldi, Yifat Shorr, and Daniel Gopher. Evaluation of multimodal feedback effects on the time-course of motor learning in multimodal vr platform for rowing training. In *2015 International Conference on Virtual Rehabilitation (ICVR)*, pages 158–159. IEEE, 2015.

[62] Tatiana Ortegon, David Acosta, Sebastian Salgado, Wilhelm Mino, Joss Moo-Young, Danny Luk, Connor Smiley, Tom Tsiliopoulos, Jacky Yang, Oscar I Caldas, et al. Prototyping interactive multimodal vr epidural administration. In *2019 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–3. IEEE, 2019.

[63] Beryl Plimmer, Peter Reid, Rachel Blagojevic, Andrew Crossan, and Stephen Brewster. Signing on the tactile line: A multimodal system for teaching handwriting to blind children. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 18(3):1–29, 2011.

[64] Emanuele Ruffaldi, Alessandro Filippeschi, Carlo Alberto Avizzano, Benoît Bardy, Daniel Gopher, and Massimo Bergamasco. Feedback, affordances, and accelerators for training sports in virtual environments. *Presence: Teleoperators and Virtual Environments*, 20(1):33–46, 2011.

[65] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 539–546, 2015.

[66] Saurabh Shukla, Ashutosh Shivakumar, Miteshkumar Vasoya, Yong Pei, and Anna F Lyon. ileap: A human-ai teaming based mobile language learning solution for dual language learners in early and special educations. *International Association for Development of the Information Society*, 2019.

[67] John A Sokolowski, Hector M Garcia, William Richards, and Catherine M Banks. Developing a low-cost multi-modal simulator for ultrasonography training. In *Proceedings of the Conference on Summer Computer Simulation*, pages 1–5, 2015.

[68] Peter Van Rosmalen, Dirk Börner, Jan Schneider, Olga Petukhova, and Joy Van Helvert. Feedback design in multimodal dialogue systems. In *CSEDU (2)*, pages 209–217, 2015.

[69] Soonja Yeom, Derek L Choi-Lundberg, Andrew Edward Fluck, and Arthur Sale. Factors influencing undergraduate students' acceptance of a haptic interface for learning gross anatomy. *Interactive Technology and Smart Education*, 2017.

[70] Nikoleta Yiannoutsou, Rose Johnson, and Sara Price. Exploring how children interact with 3d shapes using haptic technologies. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*, pages 533–538, 2018.

[71] Amitha Mathew, P Amudha, and S Sivakumari. Deep learning techniques: an overview. *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*, pages 599–608, 2021.

[72] Muhammad Arifur Rahman, David J Brown, Nicholas Shopland, Andrew Burton, and Mufti Mahmud. Explainable multimodal machine learning for engagement analysis by continuous performance test. In *Universal Access in Human-Computer Interaction. User and Context Diversity: 16th International Conference, UAHCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part II*, pages 386–399. Springer, 2022.

[73] Jiyou Jia, Yunfan He, and Huixiao Le. A multimodal human-computer interaction system and its application in smart learning environments. In *Blended Learning. Education in a Smart Learning Environment: 13th International Conference, ICBL 2020, Bangkok, Thailand, August 24–27, 2020, Proceedings 13*, pages 3–14. Springer, 2020.

[74] Aysu Ezen-Can, Kristy Elizabeth Boyer, Shaun Kellogg, and Sherry Booth. Unsupervised modeling for understanding mooc discussion forums: a learning analytics approach. In *Proceedings of the fifth international conference on learning analytics and knowledge*, pages 146–150, 2015.

[75] Shuangyan Liu and Mathieu d'Aquin. Unsupervised learning for understanding student achievement in a distance learning setting. In *2017 IEEE Global Engineering Education Conference (EDUCON)*, pages 1373–1377. IEEE, 2017.

[76] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

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

[78] Beatrice Alenljung, Jessica Lindblom, Rebecca Andreasson, and Tom Ziemke. User experience in social human-robot interaction. In *Rapid automation: concepts, methodologies, tools, and applications*, pages 1468–1490. IGI Global, 2019.

[79] David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[80] K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363, 1993.

[81] Chih-Ming Chen and Sheng-Hui Huang. Web-based reading annotation system with an attention-based self-regulated learning mechanism for promoting reading performance. *British Journal of Educational Technology*, 45(5):959–980, 2014.

[82] Chih-Ming Chen. Personalized e-learning system with self-regulated learning assisted mechanisms for promoting learning performance. *Expert Systems with Applications*, 36(5):8816–8829, 2009.

[83] Ian Roffe. E-learning: engagement, enhancement and execution. *Quality assurance in education*, 2002.

[84] RK Rahul, S Shanthakumar, P Vykunth, and K Sairamnath. Real-time attention span tracking in online education. In *2020 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–4. IEEE, 2020.

[85] Hasnan Baber. Modelling the acceptance of e-learning during the pandemic of covid-19-a study of south korea. *The International Journal of Management Education*, 19(2):100503, 2021.

[86] Maryam Shafiei Sarvestani, Mehdi Mohammadi, Jalil Afshin, and Laleh Raeisy. Students' experiences of e-learning challenges; a phenomenological study. *Interdisciplinary Journal of Virtual Learning in Medical Sciences*, 10(3):1–10, 2019.

[87] Shan Zhang, Zihan Yan, Shardul Sapkota, Shengdong Zhao, and Wei Tsang Ooi. Moment-to-moment continuous attention fluctuation monitoring through consumer-grade eeg device. *Sensors*, 21(10):3419, 2021.

[88] Michael Esterman and David Rothlein. Models of sustained attention. *Current opinion in psychology*, 29:174–180, 2019.

[89] Pablo Oyarzo, David D Preiss, and Diego Cosmelli. Attentional and meta-cognitive processes underlying mind wandering episodes during continuous naturalistic reading are associated with specific changes in eye behavior. *Psychophysiology*, page e13994, 2022.

[90] Douglas Derryberry. Attention and voluntary self-control. *Self and identity*, 1(2):105–111, 2002.

[91] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10631–10642, 2021.

[92] Haoyu Chen, Xin Liu, Xiaobai Li, Henglin Shi, and Guoying Zhao. Analyze sponta-neous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–8, 2019.

[93] Sharifah Noor Masidayu Sayed Is, Nor Azlina Ab Aziz, and Siti Zainab Ibrahim. A comparison of emotion recognition system using electrocardiogram (ecg) and pho-toplethysmogram (ppg). *Journal of King Saud University-Computer and Information Sciences*, 2022.

[94] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

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

[96] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98, 1969.

[97] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*, volume 10. Ishk, 2003.

[98] Connie De Vos, Els Van der Kooij, and Onno Crasborn. Mixed signals: Combining linguistic and affective functions of eyebrows in questions in sign language of the netherlands. *Language and speech*, 52(2-3):315–339, 2009.

[99] Maria L Flecha-Garcia. Eyebrow raising, discourse structure, and utterance function in face-to-face dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28, 2006.

[100] Mayada R Eesa. Facial expressions a study of eyebrow movement during conversation. *Ahl Al-Bait Jurnal*, 1(10), 2010.

[101] Keith Rayner. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506, 2009.

[102] David Hoppe, Stefan Helfmann, and Constantin A Rothkopf. Humans quickly learn to blink strategically in response to environmental task demands. *Proceedings of the National Academy of Sciences*, 115(9):2246–2251, 2018.

[103] Christina A Chu, Mark Rosenfield, and Joan K Portello. Blink patterns: reading from a computer screen versus hard copy. *Optometry and Vision Science*, 91(3):297–302, 2014.

[104] Robert Schleicher, Niels Galley, Susanne Briest, and Lars Galley. Blinks and sac-cades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51(7):982–1010, 2008.

[105] Giuseppe Barbato, Vittoria De Padova, Antonella Raffaella Paolillo, Laura Arpaia, Eleonora Russo, and Gianluca Ficca. Increased spontaneous eye blink rate following prolonged wakefulness. *Physiology & behavior*, 90(1):151–154, 2007.

[106] Youngjun Cho. Rethinking eye-blink: Assessing task difficulty through physiological representation of spontaneous blinking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[107] Charles W McMonnies. The clinical and experimental significance of blinking behavior. *Journal of Optometry*, 13(2):74–80, 2020.

[108] Daniel Todorovic. Choosing what to read out loud while studying: The role of agency in production. 2020.

[109] Suzanne M Prior, Kimberley D Fenwick, Katie S Saunders, Rachel Ouellette, Chantell O'Quinn, and Shannon Harvey. Comprehension after oral and silent reading: Does grade level matter? *Literacy Research and Instruction*, 50(3):183–194, 2011.

[110] Young-Suk Grace Kim, Yaacov Petscher, and Christian Vorstius. Unpacking eye movements during oral and silent reading and their relations to reading proficiency in beginning readers. *Contemporary Educational Psychology*, 58:102–120, 2019.

[111] Rebecca Boehme and Håkan Olausson. Differentiating self-touch from social touch. *Current Opinion in Behavioral Sciences*, 43:27–33, 2022.

[112] Shimpei Ishiyama, Lena V Kaufmann, and Michael Brecht. Behavioral and cortical correlates of self-suppression, anticipation, and ambivalence in rat tickling. *Current Biology*, 29(19):3153–3164, 2019.

[113] Stephanie Margarete Mueller, Sven Martin, and Martin Grunwald. Self-touch: contact durations and point of touch of spontaneous facial self-touches differ depending on cognitive and emotional load. *PloS one*, 14(3):e0213677, 2019.

[114] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective Computing and Intelligent Interaction*, pages 71–82. Springer, 2007.

[115] Abhishek Revadekar, Shreya Oak, Aumkar Gadekar, and Pramod Bide. Gauging attention of students in an e-learning environment. In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, pages 1–6. IEEE, 2020.

[116] Andrea Veronese, Mattia Racca, Roel Stephan Pieters, and Ville Kyrki. Probabilistic mapping of human visual attention from head pose estimation. *Frontiers in Robotics and AI*, 4:53, 2017.

[117] Justine Cassell, Y Nakano, T Bickmore, C Sidner, and Charles Rich. Annotating and generating posture from discourse structure in embodied conversational agents. In *Workshop on representing, annotating, and evaluating non-verbal and verbal communicative acts to achieve contextual embodied agents*, 2001.

[118] Myrthe Faber, Robert Bixler, and Sidney K D'Mello. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*, 50(1):134–150, 2018.

[119] Timothy Bell. Extensive reading: Speed and comprehension. *The reading matrix*, 1(1), 2001.

[120] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[121] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[122] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[123] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[124] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999.

[125] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[126] Damla Arifoglu and Abdelhamid Bouchachia. Activity recognition and abnormal behaviour detection with recurrent neural networks. *Procedia Computer Science*, 110:86–93, 2017.

[127] Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54(3):2259–2322, 2021.

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

[129] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Applied Sciences*, 11(2):796, 2021.

[130] Haoyu Chen, Zitong Yu, Xin Liu, Wei Peng, Yoon Lee, and Guoying Zhao. 2nd place scheme on action recognition track of eccv 2020 vipriors challenges: an efficient optical flow stream guided framework. *arXiv preprint arXiv:2008.03996*, 2020.

[131] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989.

[132] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[133] Robert E Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.

[134] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[135] Kalpathy Ramaiyer Subramanian. Myth and mystery of shrinking attention span. *International Journal of Trend in Research and Development*, 5(1), 2018.

[136] Donald Eric Broadbent. A mechanical model for human attention and immediate memory. *Psychological review*, 64(3):205, 1957.

[137] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[138] Sajjad Hussain, Maksal Minaz, N Ahmad, and Muhammad Idris. The effect of e-reading and printed document reading on student's comprehension and retention power. In *International Conference on Computational and Social Sciences*, volume 25, page 08, 2015.

[139] Michael F Shaughnessy. An interview with miriam schcolnik: Reading, e-reading and writing and their assessment. 2020.

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

[141] Yoon Lee and Marcus Specht. Multimodal WEDAR dataset for attention regulation behaviors, self-reported distractions, reaction time, and knowledge gain in e-reading , 2023.

[142] Robert Bixler and Sidney D'Mello. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*, 26:33–68, 2016.

[143] Joseph Tao-yi Wang. Pupil dilation and eye tracking. *A handbook of process tracing methods for decision research: A critical review and user's guide*, pages 185–204, 2011.

[144] Dario D Salvucci and Joseph H Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78, 2000.

[145] Songpeng Yan, Michael Hahn, and Frank Keller. Modeling fixation behavior in reading with character-level neural attention. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

[146] Ziming Liu. Digital reading. *Chinese Journal of Library and Information Science (English edition)*, page 85, 2012.

[147] Livia Popa, Ovidiu Selejan, Allan Scott, Dafin F Mureşanu, Maria Balea, and Alexandru Rafila. Reading beyond the glance: eye tracking in neurosciences. *Neurological Sciences*, 36(5):683–688, 2015.

[148] Kristy Roschke and Ralph Radach. Perception, reading, and digital media. In *The cognitive development of reading and reading comprehension*, pages 33–52. Routledge, 2016.

[149] Lih-Juan ChanLin. Reading strategy and the need of e-book features. *The Electronic Library*, 31(3):329–344, 2013.

[150] Yoon Lee, Bibeg Limbu, Zoltan Rusak, and Marcus Specht. Role of multimodal learning systems in technology-enhanced learning (tel): A scoping review. In *European Conference on Technology Enhanced Learning*, pages 164–182. Springer, 2023.

[151] Xuan Liu, Jiachen Ma, and Qiang Wang. A social robot as your reading companion: exploring the relationships between gaze patterns and knowledge gains. *Journal on Multimodal User Interfaces*, pages 1–21, 2023.

[152] Yoon Lee, Gosia Migut, and Marcus Specht. What attention regulation behaviors tell us about learners in e-reading?: Adaptive data-driven persona development and application based on unsupervised learning. *IEEE Access*, 2023.

[153] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. Explainable ai: current status and future directions. *arXiv preprint arXiv:2107.07045*, 2021.

[154] José M Alonso and Gabriella Casalino. Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In *Higher Education Learning Methodologies and Technologies Online: First International Workshop, HELMeTO 2019, Novedrate, CO, Italy, June 6-7, 2019, Revised Selected Papers 1*, pages 125–138. Springer, 2019.

[155] Rania Qasrawi and Abdullah BeniAbdelrahman. The higher and lower-order thinking skills (hots and lots) in unlock english textbooks (1st and 2nd editions) based on bloom's taxonomy: An analysis study. *International Online Journal of Education and Teaching*, 7(3):744–758, 2020.

[156] Yoon Lee, Haoyu Chen, Guoying Zhao, and Marcus Specht. Wedar: Webcam-based attention analysis via attention regulator behavior recognition with a novel e-reading dataset. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 319–328, 2022.

[157] Inge Molenaar. Towards hybrid human-ai learning technologies. *European Journal of Education*, 57(4):632–645, 2022.

[158] Yoon Lee and Marcus Specht. 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*. Association for Computing Machinery (ACM), 2023.

[159] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, 2022.

[160] Mutlu Cukurova, Qi Zhou, Daniel Spikol, and Lorenzo Landolfi. Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough? In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 270–275, 2020.

[161] Liudmila A Dikaya, Garen Avanesian, Igor S Dikiy, Vladimir A Kirik, and Valeria A Egorova. How personality traits are related to the attitudes toward forced remote learning during covid-19: Predictive analysis using generalized additive modeling. In *Frontiers in Education*, volume 6, page 629213. Frontiers Media SA, 2021.

[162] Larry W Howard, Thomas Li-Ping Tang, and M Jill Austin. Teaching critical thinking skills: Ability, motivation, intervention, and the pygmalion effect. *Journal of Business Ethics*, 128:133–147, 2015.

[163] Benidiktus Tanujaya, Jeinne Mumu, and Gaguk Margono. The relationship between higher order thinking skills and academic performance of student in mathematics instruction. 2017.

[164] Yee Mei Heong, Widad Binti Othman, Jailani Bin Md Yunos, Tee Tze Kiong, Razali Bin Hassan, and Mimi Mohaffyza Binti Mohamad. The level of marzano higher order thinking skills among technical education students. *International Journal of Social Science and Humanity*, 1(2):121, 2011.

[165] Sowmya Narayanan, Muhammad Adithan, et al. Analysis of question papers in engineering courses with respect to hots (higher order thinking skills). *American Journal of Engineering Education (AJEE)*, 6(1):1–10, 2015.

[166] Elena Tikhonova and Nataliya Kudinova. Sophisticated thinking: Lower order thinking skills. In *Proceedings of the 2nd International Multidisciplinary Scientific Conferences on Social Sciences and Arts*, volume 2, pages 352–360, 2015.

[167] Yousef Abosalem. Assessment techniques and students' higher-order thinking skills. *International Journal of Secondary Education*, 4(1):1–11, 2016.

[168] Jamie L Jensen, Mark A McDaniel, Steven M Woodard, and Tyler A Kummer. Teaching to the test... or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, 26:307–329, 2014.

[169] Gulistan Mohammed Saido, Saedah Siraj, Abu Bakar Bin Nordin, and Omed Saadallah Al_Amedy. Higher order thinking skills among secondary school students in science learning. *MOJES: Malaysian Online Journal of Educational Sciences*, 3(3):13–20, 2018.

[170] Brigitte A McKown and Cynthia L Barnett. Improving reading comprehension through higher-order thinking skills. *Online Submission*, 2007.

[171] Bhawani Prasad Mainali. Higher order thinking in education. *Academic Voices: A Multidisciplinary Journal*, 2:5–10, 2012.

[172] Arnita Cahya Saputri, Yudi Rinanto, Nanik Murti Prasetyanti, et al. Improving students' critical thinking skills in cell-metabolism learning using stimulating higher order thinking skills model. *International Journal of Instruction*, 12(1):327–342, 2019.

[173] Marvin Zuckerman. *Behavioral expressions and biosocial bases of sensation seeking*. Cambridge university press, 1994.

[174] Andrea Zunino, Jacopo Cavazza, and Vittorio Murino. Revisiting human action recognition: Personalization vs. generalization. In *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part I 19*, pages 469–480. Springer, 2017.

[175] Jochem van Kempen, Gerard M Loughnane, Daniel P Newman, Simon P Kelly, Alexander Thiele, Redmond G O'Connell, and Mark A Bellgrove. Behavioural and neural signatures of perceptual decision-making are modulated by pupil-linked arousal. *Elife*, 8:e42541, 2019.

[176] Herbert Bless and Klaus Fiedler. Affective states and the influence of activated general knowledge. *Personality and Social Psychology Bulletin*, 21(7):766–778, 1995.

[177] Marc Brysbaert. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of memory and language*, 109:104047, 2019.

[178] Joachim Wirth, Ferdinand Stebner, Melanie Trypke, Corinna Schuster, and Detlev Leutner. An interactive layers model of self-regulated learning and cognitive load. *Educational Psychology Review*, 32(4):1127–1149, 2020.

[179] Mary C Dyson and Mark Haselgrove. The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, 54(4):585–612, 2001.

[180] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[181] Sandeep Trivedi and Nikhil Patel. Clustering students based on virtual learning engagement, digital skills, and e-learning infrastructure: Applications of k-means, dbscan, hierarchical, and affinity propagation clustering. *Sage Science Review of Educational Technology*, 3(1):1–13, 2020.

[182] Murali Shanker, Michael Y Hu, and Ming S Hung. Effect of data standardization on neural network training. *Omega*, 24(4):385–397, 1996.

[183] Richard D Goffin, R Blake Jelley, Deborah M Powell, and Norman G Johnston. 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):251–268, 2009.

[184] Yoon Lee, Gosia Migut, and Marcus Specht. What attention regulation behaviors tell us about learners in e-reading?: Adaptive data-driven persona development and application based on unsupervised learning. *IEEE Access*, pages 1–17, 2023.

[185] Plinio Thomaz Aquino Junior and Lucia Vilela Leite Filgueiras. User modeling with personas. In *Proceedings of the 2005 Latin American conference on Human-computer interaction*, pages 277–282, 2005.

[186] Mikio Aoyama. Persona-scenario-goal methodology for user-centered requirements engineering. In *15th IEEE International Requirements Engineering Conference (RE 2007)*, pages 185–194. IEEE, 2007.

[187] Abdelsalam M Maatuk, Ebitisam K Elberkawi, Shadi Aljawarneh, Hasan Rashaideh, and Hadeel Alharbi. The covid-19 pandemic and e-learning: challenges and opportunities from the perspective of students and instructors. *Journal of computing in higher education*, 34(1):21–38, 2022.

[188] Jauwairia Nasir, Utku Norman, Wafa Johal, Jennifer K Olsen, Sina Shahmoradi, and Pierre Dillenbourg. 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)*, pages 1–7. IEEE, 2019.

[189] Yoon Lee, Haoyu Chen, Esther Tan, Sambit Praharaj, and Marcus Specht. Flower: Feedback loop for group work supporter. In *The International Learning Analytics and Knowledge Conference (LAK demo session)*, 2020.

[190] S Ghazal Ghalebandi and Abdullah Noorhidawati. Engaging children with pleasure reading: The e-reading experience. *Journal of Educational Computing Research*, 56(8):1213–1237, 2019.

[191] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, and Bernard J Jansen. A survey of 15 years of data-driven persona development. *International Journal of Human–Computer Interaction*, 37(18):1685–1708, 2021.

[192] Alan Cooper. The inmates are running the asylum. In *Software-Ergonomie'99*, pages 17–17. Springer, 1999.

[193] Dimitra Chasanidou, Andrea Alessandro Gasparini, and Eunji Lee. Design thinking methods and tools for innovation. In *Design, User Experience, and Usability: Design Discourse: 4th International Conference, DUXU 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings, Part I*, pages 12–23. Springer, 2015.

[194] Joni Salminen, Kathleen Guan, Soon-gyo Jung, Shammur A Chowdhury, and Bernard J Jansen. A literature review of quantitative persona creation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[195] Jon Brickey, Steven Walczak, and Tony Burgess. A comparative analysis of persona clustering methods. In *Proceedings of the Sixteenth Americas Conference on Information Systems*, 2010.

[196] Xiang Zhang, Hans-Frederick Brown, and Anil Shankar. Data-driven personas: Constructing archetypal users with clickstreams and user telemetry. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5350–5359, 2016.

[197] Mostafa Mesgari, Chitu Okoli, and Ana Ortiz de Guinea. Affordance-based user personas: A mixed-method approach to persona development. In *Twenty-first Americas Conference on Information Systems*. Citeseer, 2015.

[198] James E Nieters, Subbarao Ivaturi, and Iftikhar Ahmed. Making personas memorable. In *CHI'07 extended abstracts on Human factors in computing systems*, pages 1817–1824, 2007.

[199] John Pruitt and Jonathan Grudin. Personas: practice and theory. In *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15, 2003.

[200] Jennifer McGinn and Nalini Kotamraju. Data-driven persona development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1521–1524, 2008.

[201] Yoon Lee and Marcus Specht. Multimodal SKEP dataset for attention regulation behaviors, knowledge gain, perceived learning experience, and perceived social presence in e-learning with a conversational agent. `https://data.4tu.nl/datasets/4c9de645-ca88-4b45-8fc7-2fc325f191dc/1`, 2023.

[202] Bharatwaja Namatherdhala, Noman Mazher, and Gopal Krishna Sriram. A comprehensive overview of artificial intelligence tends in education. *International Research Journal of Modernization in Engineering Technology and Science*, 4(7), 2022.

[203] Ningyu Zhang, Gautam Biswas, and Yi Dong. Characterizing students' learning behaviors using unsupervised learning methods. In *International conference on artificial intelligence in education*, pages 430–441. Springer, 2017.

[204] Karina Huang, Tonya Bryant, and Bertrand Schneider. Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. *International Educational Data Mining Society*, 2019.

[205] Onofrio Rosario Battaglia, Benedetto Di Paola, and Claudio Fazio. A new approach to investigate students' behavior by using cluster analysis as an unsupervised methodology in the field of education. *Applied Mathematics*, 7(15):1649–1673, 2016.

[206] Abdallah Moubayed, Mohammadnoor Injadat, Abdallah Shami, and Hanan Lutfiyya. Student engagement level in an e-learning environment: Clustering using k-means. *American Journal of Distance Education*, 34(2):137–156, 2020.

[207] Udoinyang G Inyang, Uduak Augustine Umoh, Ifeoma C Nnaemeka, and Samuel A Robinson. Unsupervised characterization and visualization of students' academic performance features. *Comput. Inf. Sci.*, 12(2):103–116, 2019.

[208] Steve Mulder and Ziv Yaar. *The user is always right: A practical guide to creating and using personas for the web*. New Riders, 2006.

[209] Tomasz Miaskiewicz, Tamara Sumner, and Kenneth A Kozar. A latent semantic analysis methodology for the identification and creation of personas. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1501–1510, 2008.

[210] Bart Rienties, Avinash Boroowa, Simon Cross, Chris Kubiak, Kevin Mayles, and Sam Murphy. 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.

[211] Elisa Prati, Margherita Peruzzini, Marcello Pellicciari, and Roberto Raffaeli. How to include user experience in the design of human-robot interaction. *Robotics and Computer-Integrated Manufacturing*, 68:102072, 2021.

[212] Nargiza Gaybullaevna Dilova. Formative assessment of students' knowledge–as a means of improving the quality of education. *Scientific reports of Bukhara State University*, 5(3):144–155, 2021.

[213] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 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*, pages 229–238, 2021.

[214] James V Hoffman and Cherry L Kugle. A study of theoretical orientation to reading and its relationship to teacher verbal feedback during reading instruction. *The Journal of Classroom Interaction*, pages 2–7, 1982.

[215] Dale H Schunk and Jo Mary Rice. Learning goals and progress feedback during reading comprehension instruction. *Journal of Reading Behavior*, 23(3):351–364, 1991.

[216] Elaheh Shahmir Shourmasti, Ricardo Colomo-Palacios, Harald Holone, and Selina Demi. User experience in social robots. *Sensors*, 21(15):5052, 2021.

[217] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[218] Justyna Gerłowska, Urszula Skrobas, Katarzyna Grabowska-Aleksandrowicz, Agnieszka Korchut, Sebastian Szklener, Dorota Szczęśniak-Stańczyk, Dimitrios Tzovaras, and Konrad Rejdak. Assessment of perceived attractiveness, usability, and societal impact of a multimodal robotic assistant for aging patients with memory impairments. *Frontiers in neurology*, 9:392, 2018.

[219] Matthieu Destephe, Martim Brandao, Tatsuhiro Kishi, Massimiliano Zecca, Kenji Hashimoto, and Atsuo Takanishi. Walking in the uncanny valley: importance of the attractiveness on the acceptance of a robot as a working partner. *Frontiers in psychology*, 6:204, 2015.

[220] Smita Singh, Eric D Olson, and Chin-Hsun Ken Tsai. Use of service robots in an event setting: Understanding the role of social presence, eeriness, and identity threat. *Journal of Hospitality and Tourism Management*, 49:528–537, 2021.

[221] Nan Liang and Goldie Nejat. A meta-analysis on remote hri and in-person hri: What is a socially assistive robot to do? *Sensors*, 22(19):7155, 2022.

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

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

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

[225] Jacqueline M Kory-Westlund and Cynthia Breazeal. 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*, 6:54, 2019.

[226] Matthew Lombard, Theresa B Ditton, Daliza Crane, Bill Davis, Gisela Gil-Egui, Karl Horvath, Jessica Rossman, and S Park. 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*, volume 240, pages 2–4, 2000.

[227] James J Cummings and Erin E Wertz. Capturing social presence: concept explication through an empirical analysis of social presence measures. *Journal of Computer-Mediated Communication*, 28(1):zmac027, 2023.

[228] Iolanda Leite, Carlos Martinho, Andre Pereira, and Ana Paiva. 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*, pages 669–674. IEEE, 2009.

[229] Stephanie A Andel, Triparna de Vreede, Paul E Spector, Balaji Padmanabhan, Vivek K Singh, and Gert-Jan De Vreede. Do social features help in video-centric online learning platforms? a social presence perspective. *Computers in Human Behavior*, 113:106505, 2020.

[230] Charlotte A Jones-Roberts. Increasing social presence online: Five strategies for instructors. *FDLA Journal*, 3(1):8, 2018.

[231] Mete Akcaoglu and Eunbae Lee. Increasing social presence in online learning through small group discussions. *International Review of Research in Open and Distributed Learning*, 17(3):1–17, 2016.

[232] Julie E Kendall and Kenneth E Kendall. Enhancing online executive education using storytelling: an approach to strengthening online social presence. *Decision Sciences Journal of Innovative Education*, 15(1):62–81, 2017.

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

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

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

[236] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*, pages 114–130. Springer, 2012.

[237] Salem Alelyani, Jiliang Tang, and Huan Liu. Feature selection for clustering: A review. *Data Clustering*, pages 29–60, 2018.

[238] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[239] Maha Alkhayrat, Mohamad Aljnidi, and Kadan Aljoumaa. A comparative dimensionality reduction study in telecom customer segmentation using deep learning and pca. *Journal of Big Data*, 7:1–23, 2020.

[240] Dehui Luo, Xiang Wan, Jiming Liu, and Tiejun Tong. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. *Statistical methods in medical research*, 27(6):1785–1805, 2018.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[256] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. In *International Journal of Computer Vision*, 2023.

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

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

[259] Neil A Bradbury. Attention span during lectures: 8 seconds, 10 minutes, or more?, 2016.

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

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

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

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

[264] Pieter Desmet. Designing emotions, 2002.

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

[266] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

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

[268] Yoon Lee, Gosia Migut, and Marcus Specht. Behavior-based feedback loop for attentive e-reading (bflae): A real-time computer vision approach. In *nternational IJCAI Workshop on Micro-gesture Analysis for Hidden Emotion Under-standing*, 2023.

[269] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. Window size impact in human activity recognition. *Sensors*, 14(4):6474–6499, 2014.

[270] Edna Holland Mory. Feedback research revisited. In *Handbook of research on educational communications and technology*, pages 738–776. Routledge, 2013.

[271] Umair Ahmed, Abdussalaam Iyanda Ismail, Meryem Fati, and Mohammed Ali Akour. E-learning during covid-19: understanding the nexus between instructional innovation, e-psychological capital, and online behavioural engagement. *Management in Education*, page 08920206211053101, 2021.

[272] Pushkar Dubey, Resham Lal Pradhan, and Kailash Kumar Sahu. Underlying factors of student engagement to e-learning. *Journal of Research in Innovative Teaching & Learning*, 16(1):17–36, 2023.

[273] Leili Yekefallah, Peyman Namdar, Rahman Panahi, and Leila Dehghankar. Factors related to students' satisfaction with holding e-learning during the covid-19 pandemic based on the dimensions of e-learning. *Heliyon*, 7(7), 2021.

[274] Xin-Yu Wang, Guang Li, Summaira Malik, and Ahsan Anwar. Impact of covid-19 on achieving the goal of sustainable development: E-learning and educational productivity. *Economic Research-Ekonomska Istraživanja*, 35(1):1950–1966, 2022.

[275] Zethembe Mseleku. A literature review of e-learning and e-teaching in the era of covid-19 pandemic, 2020.

[276] Sherif Welsen, Matthew Pike, and James Walker. Engineering student attitudes to e-reading in remote teaching environments. In *2020 IFEES World Engineering Education Forum-Global Engineering Deans Council (WEEF-GEDC)*, pages 1–6. IEEE, 2020.

[277] Will Drover, Matthew S Wood, and Andrew C Corbett. Toward a cognitive view of signalling theory: Individual attention and signal set interpretation. *Journal of management studies*, 55(2):209–231, 2018.

[278] Katrien Verhoeven, Geert Crombez, Christopher Eccleston, Dimitri ML Van Ryck-eghem, Stephen Morley, and Stefaan Van Damme. The role of motivation in distracting attention away from pain: an experimental study. *PAIN®*, 149(2):229–234, 2010.

[279] Sandy Kane, Miriam Lear, and Cecilia Maxine Dube. Reflections on the role of metacognition in student reading and learning at higher education level. *Africa Education Review*, 11(4):512–525, 2014.

[280] Catherine A Spann, James Schaeffer, and George Siemens. Expanding the scope of learning analytics data: Preliminary findings on attention and self-regulation using wearable technology. In *Proceedings of the seventh international learning analytics & knowledge conference*, pages 203–207, 2017.

[281] Krishna Regmi and Linda Jones. A systematic review of the factors–enablers and barriers–affecting e-learning in health sciences education. *BMC medical education*, 20(1):1–18, 2020.

[282] Michael Yi-chao Jiang, Morris Siu-yung Jong, Wilfred Wing-fat Lau, Yan-li Meng, Ching-sing Chai, and Mengyuan Chen. Validating the general extended technology acceptance model for e-learning: Evidence from an online english as a foreign language course amid covid-19. *Frontiers in Psychology*, 12:671615, 2021.

[283] Christopher Brooks, Craig Thompson, and Stephanie Teasley. A time series interaction analysis method for building predictive models of learners using log data. In *Proceedings of the fifth international conference on learning analytics and knowledge*, pages 126–135, 2015.

[284] Ioana Jivet, Jacqueline Wong, Maren Scheffel, Manuel Valle Torre, Marcus Specht, and Hendrik Drachsler. Quantum of choice: How learners' feedback monitoring decisions, goals and self-regulated learning skills are related. In *LAK21: 11th international learning analytics and knowledge conference*, pages 416–427, 2021.

[285] Robert Bodily, Judy Kay, Vincent Aleven, Ioana Jivet, Dan Davis, Franceska Xhakaj, and Katrien Verbert. Open learner models and learning analytics dashboards: a systematic review. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 41–50, 2018.

[286] Michail Giannakos, Mutlu Cukurova, and Sofia Papavlasopoulou. Sensor-based analytics in education: Lessons learned from research in multimodal learning analytics. In *The Multimodal Learning Analytics Handbook*, pages 329–358. Springer, 2022.

[287] Imène Jraidi and Claude Frasson. Student's uncertainty modeling through a multimodal sensor-based approach. *Journal of Educational Technology & Society*, 16(1):219–230, 2013.

[288] Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D Lytras, Farhat Abbas, and Jalal S Alowibdi. Predicting student performance using advanced learning analytics. In *Proceedings of the 26th international conference on world wide web companion*, pages 415–421, 2017.

[289] Yelin Kim, Tolga Soyata, and Reza Feyzi Behnagh. Towards emotionally aware ai smart classroom: Current issues and directions for engineering and education. *IEEE Access*, 6:5308–5331, 2018.

[290] Kshitij Sharma and Michail Giannakos. Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, 51(5):1450–1484, 2020.

[291] Gerges H Samaan, Abanoub R Wadie, Abanoub K Attia, Abanoub M Asaad, Andrew E Kamel, Salwa O Slim, Mohamed S Abdallah, and Young-Im Cho. Mediapipe's landmarks with rnn for dynamic sign language recognition. *Electronics*, 11(19):3228, 2022.

[292] Yoon Lee, Gosia Migut, and Marcus Specht. Behavior-based feedback loop for attentive e-reading (bflae): A real-time computer vision approach. In *Micro-gesture Analysis for Hidden Emotion Understanding 2023*, page 12, 2023.

[293] Barry J Zimmerman and Adam R Moylan. Self-regulation: Where metacognition and motivation intersect. In *Handbook of metacognition in education*, pages 299–315. Routledge, 2009.

[294] Chiu-Lin Lai and Gwo-Jen Hwang. Strategies for enhancing self-regulation in e-learning: a review of selected journal publications from 2010 to 2020. *Interactive learning environments*, pages 1–23, 2021.

[295] John Hattie and Helen Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[296] Logan Fiorella, Jennifer J Vogel-Walcutt, and Sae Schatz. Applying the modality principle to real-time feedback and the acquisition of higher-order cognitive skills. *Educational Technology Research and Development*, 60:223–238, 2012.

[297] Richard E Mayer and Roxana Moreno. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52, 2003.

[298] Eduardo Salas and Janis A Cannon-Bowers. The science of training: A decade of progress. *Annual review of psychology*, 52(1):471–499, 2001.

[299] Paul Ayres and John Sweller. The split-attention principle in multimedia learning. *The Cambridge handbook of multimedia learning*, 2:135–146, 2005.

[300] Kiavash Bahreini, Rob Nadolski, and Wim Westera. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments*, 24(3):590–605, 2016.

[301] S Athi Narayanan, M Prasanth, P Mohan, MR Kaimal, and Kamal Bijlani. Attention analysis in e-learning environment using a simple web camera. In *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*, pages 1–4. IEEE, 2012.

[302] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019, 2019.

[303] Arkendu Sen and Shiang Harn Liew. Augmented reality and its use in education. *Encyclopedia of Education and Information Technologies*, pages 202–211, 2020.

[304] Uwe Maier and Christian Klotz. Personalized feedback in digital learning environments: Classification framework and literature review. *Computers and Education: Artificial Intelligence*, 3:100080, 2022.

[305] Abelardo Pardo, Kathryn Bartimote, Simon Buckingham Shum, Shane Dawson, Jing Gao, Dragan Gašević, Steven Leichtweis, Danny Liu, Roberto Martínez-Maldonado, Negin Mirriahi, et al. Ontask: Delivering data-informed, personalized learning support actions. 2018.

[306] Hexiang Huang, Xupeng Guo, Wei Peng, and Zhaoqiang Xia. Micro-gesture classification based on ensemble hypergraph-convolution transformer. In *Micro-gesture Analysis for Hidden Emotion Understanding 2023*, page 9, 2023.

[307] Stéphan Vincent-Lancrin and Reyer Van der Vlies. Trustworthy artificial intelligence (ai) in education: Promises and challenges. 2020.

[308] Daniele Di Mitri, Jan Schneider, and Hendrik Drachsler. Keep me in the loop: Real-time feedback with multimodal data. *International Journal of Artificial Intelligence in Education*, pages 1–26, 2021.

[309] Tinghuai Ma, Qian Pan, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. T-bertsum: Topic-aware text summarization based on bert. *IEEE Transactions on Computational Social Systems*, 9(3):879–890, 2021.

[310] Ingo Siegert, Matthias Busch, and Julia Krüger. Does users' system evaluation influence speech behavior in hci?–first insights from the engineering and psychological perspective. In *Konferenz Elektronische Sprachsignalverarbeitung*, pages 241–248. TUDpress, Dresden, 2020.

List of Publications

Published Work

1. **Lee, Y.**, *Migut, G., & Specht, M.*: (2023). What Attention Regulation Behaviors Tell Us About Learners in E-reading?: Adaptive Data-driven Persona Development and Application based on Unsupervised Learning. IEEE Access. (SCIE, IF=3.9).

2. **Lee, Y.**, *Chen, H., Zhao, G., & Specht, M.* (2022). WEDAR: Webcam-based Attention Analysis via Attention Regulator Behavior Recognition with a Novel E-reading Dataset. In Proceedings of the 2022 International Conference on Multimodal Interaction (pp. 319-328).

3. **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 LAK23: 13th International Learning Analytics and Knowledge Conference (pp. 520-530).

4. **Lee, Y.**, *Limbu, B., Rusak, Z., & Specht, M.* (2023). Role of Multimodal Learning Systems in Technology-Enhanced Learning (TEL): A Scoping Review. In European Conference on Technology Enhanced Learning (pp. 164-182). Cham: Springer Nature Switzerland. *Best Paper Award Nominee.*

5. **Lee, Y.**, *Migut, G., & Specht, M.*: (2023). Behavior-based Feedback Loop for Attentive E-reading (BFLAe): A Real-Time Computer Vision Approach. In Proceedings of IJCAI-2023 (32nd International Joint Conference on Artificial Intelligence) Workshop&Challenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA 2023)

6. **Lee, Y.**, *Chen, H., Tan, E., Praharaj, S., & Specht, M.*: (2020). FLOWer: Feedback Loop for Group Work Supporter. In The International Learning Analytics and Knowledge Conference (LAK demo session).

7. *Chen, H., Tan, E.,* **Lee, Y.**, *Praharaj, S., Specht, M., & Zhao, G.*: (2020). Developing AI into explanatory supporting models: An explanation-visualized deep learning prototype. In 14th International Conference of the Learning Sciences: The Interdisciplinarity of the Learning Sciences, ICLS 2020. International Society of the Learning Sciences.

Work in Submission

8. [Submitted to a peer-reviewd journal] **Lee, Y.**, *Migut, G., & Specht, M.*: (2023). Unveiling Behavior-based Explainable AI: Predicting Learners' Higher-Order and Lower-Order Thinking Skills through Attention Regulation Behaviors in E-reading.

9. [Submitted to a peer-reviewd journal] **Lee, Y.**, *Migut, G., & Specht, M.*: (2023). An AI-based Feedback Loop for Attention Management in E-reading: Adaptation Strategies for Real-time Distraction Recognition and Feedback Implementation.

Acknowledgments

Special thanks to my supervisors, Marcus Specht and Gosia Migut, and my former supervisor, Zoltan Rusak, for their uncompromising support, guidance, and the intellectual freedom afforded me throughout this journey. Your unconditional trust and patience have been the pillars of my academic growth. I cannot thank you enough for all the opportunities that you have offered me.

I extend my gratitude to the esteemed members of my graduation committee: Prof. Dr. M.A Neerincx, Prof. Dr. H. Drachsler, Prof. Dr. M. Cukurova, Prof. Dr. I. Molenaar, Prof. Dr. H. Jarodzka, and Prof. Dr. Ir. G.J.P.M. Houben. Your insightful reviews and mentorship have been invaluable in refining and finalizing my thesis.

My heartfelt appreciation goes to my family and friends, whose tireless support and encouragement have been my stronghold. I'm especially grateful to Haoyu, Sunjoo, Ruben, and Danielle for their companionship throughout my journey in Delft. I sincerely thank my fellow PhDs and Postdocs at LDE-CEL. Your fellowship and support have greatly enriched my PhD experience.

Sincerely, Yoon
Delft, February 2024

Yoon Lee
PhD Candidate, TU Delft

Interactive Intelligence: Multimodal AI for Real-Time Interaction Loop towards Attentive E-Reading

Contact

📞 (+31) 616 825 610

✉️ y.lee@tudelft.nl

Profile

DATE OF BIRTH
16-06-1990

NATIONALITY
South Korean

GENDER
Female

Language

NATIVE KOREAN

WORKING PROFICIENCY
ENGLISH (C1)

ELEMENTARY CHINESE

Skills

PROGRAMMING
- Higher Intermediate Python
- Intermediate Matlab
- Intermediate HTML/CSS
- Beginner Kotlin
- Beginner Arduino

DESIGN
- Advanced Adobe Photoshop, Illustrator, Indesign
- Advanced Rhinoceros
- Advanced AutoCAD
- Higher Intermediate Adobe Premiere, After Effects
- Higher Intermediate 3ds MAX
- Intermediate Solidworks, Catia, Alias
- Intermediate FIGMA

About me

Hi! I'm Yoon, a PhD Candidate in Computer Science at TU Delft, trying to use various machine learning methods to tackle cognitive challenges in higher education. With a background spanning design to software engineering, my drive is to develop solutions to solve real-world human attention and create feedback loops in e-reading, my current PhD research topic. As for me, I am a very self-motivated, responsible, and persistent person. I excel in translating multidisciplinary insights into tangible advancements. With a toolbox from software engineering to design thinking, I'm on a mission to make things that matter.

Education

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE, TU DELFT, DELFT, NETHERLANDS
PhD Candidate (Computer Science) | Aug 2019 - Feb 2024

- Graduation Project - Interactive Intelligence: Multimodal AI for Real-Time Interaction Loop towards Attentive E-Reading
- Affiliated with the Leiden-Delft-Erasmus Center for Education and Learning (LDE-CEL)
- Supervised by Prof. Dr. Marcus Specht

FACULTY OF INDUSTRIAL DESIGN ENGINEERING, DELFT, TU DELFT, NETHERLANDS
Master of Science (Industrial Design Engineering) | Aug

- Specialized in Medisign (Design Innovation for Healthcare)
- Graduation Project - Noise fatigue in the ICU: platform for sound data collection and visualization
- Affiliated with Critical Alarms Lab (CAL) and Erasmus MC
- Supervised by Dr. Elif Ozcan Vieira

FACULTY OF INDUSTRIAL DESIGN, HONGIK UNIVERSITY, SEOUL, SOUTH KOREA
Bachelor of Arts (Industrial Design) | Mar 2010 - Feb 2015

- Specialized in Integrated Product Design
- Minored in Business Administration
- Winning merit-based scholarships for 8 times out of 8 semesters for academic excellence: full-funded scholarship for 1 year for entrance (100%), Jaju scholarship 2 times (80% of tuition), Changjo scholarship 3 times (60% of tuition), Hyupdong scholarship 1 time (40% of tuition)
- Graduation Project-Appcessory Design for Preventing Drowsy Driving (IoT design, mock-up, exhibited in Hongik Graduation Show, SK Telecom Smart Appcessory Design Competition)

References

PROF. DR. MARCUS SPECHT

Full professor, Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft, Netherlands

M.M.Specht@tudelft.nl

DR. GOSIA MIGUT

Assistant professor, Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft, Netherlands

M.A.Migut@tudelft.nl

DR. ZOLTAN RUSAK

IT Business Consultant for Manufacturing, Henkel, Netherlands

Former Assistant professor, Faculty of Industrial Design Engineering, TU Delft, Netherlands

zoltan.rusak@henkel.com

Work Experience

R & D SUPPORT DEPARTMENT, HYUNDAI MOTOR GROUP, SEOUL, SOUTH KOREA
Research Intern (1st prize, intern evaluation) | Jun 2014 - Aug 2014

Unit Design Team
- Hyundai WIA Design Guideline
- CNC milling Machine Handle Design

Research Engineer | Jan 2015 - May 2017

Hyundai WIA H/W Design Lead
- New Generation Concept Design for Hyundai WIA Mother Machines
- Controller Design for New Line-up; iTROL+ Bar Type, Folder Type (Government R & D Project)
- Design Patents: iTROL+ Bar Type (3009075930000), iTROL+ Folder Type (3009043320000)
- Design Awards: PINUP DESIGN AWARD iTROL+ Bar Type (Best 100), iTROL+ Folder Type (Finalist)

Hyundai WIA S/W Design Lead
- GUI Guidelines for iTROL+
- Mother machine U/X Strategy for Government R&D Development

Project Experience & Collaboration

Interactive Intelligence: Multimodal AI for Real-Time Interaction Loops Towards Attentive E-Reading
PhD Project | Aug 2019 - Feb 2024

Research Questions
- What theoretical and technical approaches can be taken to recognize learners' attention regulation in e-reading for higher education?
- How can automatic AI-based real-time feedback in e-reading assist attention management for higher education learners on their knowledge gain, perceptions, and interaction qualities with the system?
- Collaborations: INSYGHTLab (TU Delft, Netherlands), Nanyang Technological University (Singapore), Stanford University (United States), Oulu University (Finland), and Harbin University (China)

OER PROJECT (NTU): UNCOVERING THE PROCESS AND OUTCOME OF COMPUTER-SUPPORTED-COLLABORATIVE-LEARNING (CSCL) USING MULTIMODAL LEARNING ANALYTICS, PI:CHEW LEE
Technological Consultant | Aug 2019 - Sep 2021

Duties
- Provision of technical consultancy to the design and development of Real-time feedback loop design and visualization for MMLA for collaborative learning,
- Deliver post-hoc dashboard design and visualization for instructors and students
- Write and publish papers on top-conferences (ICLS 2020, LAK 2020)
- Link to Demo video: https://www.youtube.com/watch?v=pfblSWSznB0&ab_channel=YoonLee

FORD X CHILDREN ENTERTAINING EDUCATIONAL AR EXPERIENCE FOR CHILDREN IN AN AUTONOMOUS CAR

Student R & D Project | Aug 2017 - Jan 2018

- Link to presentation video: https://www.youtube.com/
  watch?v=7UXVPqMBzJU&ab_channel=YoonLee

Teaching Merits

Coach/mentor | Aug 2019 - Ongoing

Weekly teaching, assignment design, feedback, evaluation

- Bachelor Seminar (19/20 Q3), Computer Science, TU Delft: systematic
  literature review on multimodal learning systems
- Research Project (20/21 Q4), Computer Science, TU Delft: using
  multimodal sensors to research to analyze learners' sustained attention
  in e-reading for higher education
- Software Project (20/21 Q4), Computer Science, TU Delft: exploring
  various indicators and train machine learning models for attention
  management in e-reading
- Living Education Lab (23 Q1), Minor, Leiden University: design thinking
  and technologies for educational applications

Publications

Lee, Y., Migut, G., & Specht, M. (2023). What Attention Regulation
Behaviors Tell Us About Learners in E-reading?: Adaptive Data-driven
Persona Development and Application based on Unsupervised Learning.
IEEE Access (SCIE, IF=3.9).

Lee, Y., & Specht, M. (2023, March). 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,
CORE A=excellent).

Lee, Y., Chen, H., Zhao, G., & Specht, M. (2022, November). WEDAR:
Webcam-based Attention Analysis via Attention Regulator Behavior
Recognition with a Novel E-reading Dataset. In Proceedings of the
2022 International Conference on Multimodal Interaction (ICMI, CORE
B=good to very good).

Lee, Y., Limbu, B., Rusak, Z., & Specht, M. (2023, August). Role of
Multimodal Learning Systems in Technology-Enhanced Learning (TEL):
A Scoping Review. In European Conference on Technology Enhanced
Learning (ECTEL, CORE B=good to very good). **Best paper nominee**

Lee, Y., Migut, G., & Specht, M. (2023). Behavior-based Feedback Loop
for Attentive E-reading (BFLAe): A Real-Time Computer Vision Approach
(IJCAI, CORE A*=exceptional).

Chen, H., Tan, E., **Lee, Y.**, Praharaj, S., Specht, M., & Zhao, G. (2020, June).
Developing AI into explanatory supporting models: An explanation-
visualized deep learning prototype. In 14th International Conference of
the Learning Sciences: The Interdisciplinarity of the Learning Sciences,
ICLS 2020. International Society of the Learning Sciences.

Chen, H., Yu, Z., Liu, X., Peng, W., **Lee, Y.**, & Zhao, G. (2020). 2nd place
scheme on action recognition track of eccv 2020 vipriors challenges:
an efficient optical flow stream guided framework. arXiv preprint
arXiv:2008.03996.

Lee, Y., Chen, H., Tan, E., Praharaj, S., & Specht, M. (2020). FLOWer: Feedback Loop for Group Work Supporter. In The International Learning Analytics and Knowledge Conference (LAK demo session).

Datasets

Lee, Y., Specht, M. (2023): Multimodal WEDAR dataset for attention regulation behaviors, self-reported distractions, reaction time, and knowledge gain in e-reading. Version 1. 4TU.ResearchData. dataset. **https://doi.org/10.4121/8f730aa3-ad04-4419-8a5b-325415d2294b.v1**

Lee, Y., Specht, M. (2023): Multimodal SKEP dataset for attention regulation behaviors, knowledge gain, perceived learning experience, and perceived social presence in e-learning with a conversational agent. Version 1. 4TU.ResearchData. dataset. **https://doi.org/10.4121/4c9de645-ca88-4b45-8fc7-2fc325f191dc.v1**

Patents & Awards

PATENTS

- KR Patent: Hyundai WIA mother machine controller bar type (3009075930000)
- KR Patent: Hyundai WIA mother machine controller folder type (3009043320000)

AWARDS

- 2023 European Conference on Technology Enhanced Learning Best Paper Award Nominee (Role of Multimodal Learning Systems in Technology-Enhanced Learning (TEL): A Scoping Review)
- 2016 PINUP Design Awards Best 100 (iTROL+ Bar Type)
- 2016 PINUP Design Awards Finalist (iTROL+ Folder Type)
- 2014 SK Telecom Smart Appcessory Design Competition Finalist (Grabeat)

Auditory

Multimodal Reasoning

Visual

Tactile

TUDelft