EYE-TEACH

Deliverable 3.1

System Requirements and Literature Review of Potential AI Solutions

Author(s):	Milos Kravcik, Nghia Duong-Trung, Michael Dietrich, Alexander Zimmermann (DFKI); Konstantinos Tsiakas, Deniz Iren (OUNL); Francesca Zermiani, Ladislao Salmerón (UVEG); Ludovica Marinuci (CNR); Michael Raschke (Blickshift); Gustaf Öqvist Seimyr (Lexplore)
Editor(s):	Milos Kravcik (DFKI)
Responsible Organisation:	DFKI
Version-Status:	V1
Submission date:	27/06/2025
Dissemination level:	PU – Public

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Table of Contents

1. Executive Summary

The combination of advanced ET technology and AI offers great potential for improving educational methodologies. ET data is inherently temporal and sequential, typically captured as time-series data. This temporal nature requires ML methods capable of processing and learning from sequential patterns. Depending on the analysis goal, different ML paradigms can be employed: classification, regression and clustering. These algorithms have the advantage of being interpretable and computationally efficient, making them suitable for real-time applications in educational contexts. However, they often require manual feature engineering and struggle to capture complex temporal dependencies inherent in raw gaze sequences.

Classical ML algorithms such as SVM, DT, RF, Naive Bayes, and KNN have been used to analyze ET data. These models are particularly effective when gaze features are aggregated over time windows into summary statistics. To better capture the temporal dynamics of ET data, more advanced models like RNN and their variants such as LSTM have been employed. These algorithms can process raw, unsegmented gaze sequences and learn temporal patterns associated with attention, cognitive load, or emotional arousal. However, RNNs and LSTMs often require larger datasets to train effectively and are less interpretable than traditional models.

At the frontier of ML applications in ET are Transformer-based models, which have recently gained importance for their ability to model long-range dependencies in sequential data without relying on recurrent structures. Transformers use self-attention mechanisms to dynamically weigh the relevance of each gaze point in the sequence, making them highly effective in capturing both short- and long-term interactions between eye movements and cognitive states. This is crucial for converting the granularity of ET data into a format that can effectively inform teaching strategies.

In this deliverable, we summarised the requirements for AI solutions that can translate eye movements into comprehensible outputs, supporting teachers in diagnosing reading difficulties and giving them greater control in the classroom and during one-to-one tutorials. We identified the most relevant eye movement metrics, including fixation duration, saccade length, reading speed, the word-frequency effect, skipping rate, first-pass fixation duration, gaze duration and regression rate, as well as suitable public datasets such as CopCo, GECO, MECO, MQA-RC, OneStop, PoTeC, SB-SAT and WebQAmGaze. We also considered both low-end and high-end solutions for our development, enabling normal usage with webcams in classroom settings, as well as evaluation with high-resolution ET systems. Finally, we specified the human-centred, AI-based requirements alongside the ethical and legal requirements, in which transparency, explainability and human autonomy play an important role.

2. Abbreviations and definitions

2.1. Abbreviations

AI	Artificial Intelligence
DL	Deep Learning
DT	Decision Trees
ET	Eye Tracking
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
ML	Machine Learning
RF	Random Forest
RNN	Recurrent Neural Network
SVM	Support Vector Machines

2.2. Literature review

A literature review is a comprehensive survey and critical analysis of academic sources (such as books and journal articles) on a specific topic. It provides an overview of existing knowledge, identifies key theories and methods, and highlights research gaps. It also situates a new research project within the broader scholarly conversation. In essence, it demonstrates an understanding of a subject area and its relevant research landscape.

2.3. Requirements analysis

Requirements analysis is the systematic process of defining, documenting and managing the needs and expectations of stakeholders regarding a new or modified product or service. It is a crucial phase in any project, particularly in software development, ensuring that the final product meets user needs and business objectives. The process involves understanding what the product should do and how it should behave within certain constraints.

3. Introduction

EYE-TEACH is dedicated to enhancing educational methodologies by using AI and eye-tracking technologies to support teachers in assessing and improving student's reading comprehension skills. By integrating these innovative tools, the project aims to bridge the gap between technology and pedagogy, ultimately encouraging a more inclusive and effective learning environment. As part of WP3 – *AI Tools and Technological Solutions*, the project focuses on the design and development of a pilot system that can transform ET data into comprehensible outputs.

EYE-TEACH aims to assess and enhance the potential and impact of new technological developments in eye-tracking and AI to support teachers' pedagogical skills in assessing and supporting reading comprehension skills in students. The development and deployment of new technologies by EYE-TEACH will directly contribute to this goal, however the intended scale of the project goes far beyond the roll-out of new solutions. The ultimate goal of strengthening social and economic resilience and inclusive growth through enhancing complementarity between technology and human skills impacts society as a whole and requires a societal shift in attitudes and behaviours that will be facilitated through the widespread uptake of the EYE-TEACH results.

WP3 aims to design and develop a pilot AI system for transforming ET data to a comprehensible output.

The general objectives of WP3 are:

- Identifying models that are useful for transforming ET metrics into comprehensible outputs
- Designing suitable UI solutions for our target group
- Developing an AI-based pilot system to support reading comprehension
- Preparing guidelines on how to use our AI-based system to support reading comprehension

According to the EYE-TEACH Grant Agreement:

The Exploration phase (Month 1-15) will allow for scoping and assessing the current and existing state-of-the-art and technology readiness of AI-assisted-ET-analytics in different educational contexts. After establishing consortium communication channels and completing the initial project startup tasks, phase 1 will kick-off with a workshop organised by WP5 in order to explore good practices on AI-assisted ET-analytics systems. The workshop will involve all the partners of the EYE-TEACH project. The insights gained during the workshop will be used to develop vignettes in co-creation with stakeholders (the interaction ecosystem) that will be used to map education professionals' needs, acceptance of and readiness to adopt AI-assisted ET-analytics tools (WP1). In parallel, a set of ET metrics suitable to detect differences in learner processes and available datasets for training purposes of AI algorithms will be identified (WP2). For transforming ET metrics to comprehensible output, the project will explore system requirements and define a data framework (WP3). Ethical and legal issues in the field as well as related to the work of all three aforementioned WPs will be investigated in order to set up an ethical framework for building the pilot system (WP4). Synergies between work packages will be coordinated and progress, quality and risks will be monitored throughout the project timeline (WP6).

This work plan is reflected in the structure of this deliverable. It is informed by the outcomes of the teacher workshop (WP1) as well as relevant eye-movement metrics and public datasets identification (WP2), complemented by existing legal regulations (WP4).

The aim of this deliverable is to clarify various aspects of the research and development in our project. This includes the target groups, application scenarios, human-centered design aspects, relevant metrics and datasets, AI models, output forms and ethical issues. The purpose of this is to support the main objective of improving reading comprehension at various levels of the education system.

4. Literature Review

The combination of advanced ET technology and AI offers great potential for improving educational methodologies. AI's ability to manage and interpret complex data sets can bridge the gap between the wealth of data that ET yields and the actionable insights that educators require. Although ET can capture nuanced parameters reflecting cognitive processes during text comprehension, the raw data can be overwhelming and uninformative for teachers without proper analysis.

ET is a valuable tool for assessing student engagement and performance. Significant differences in saccade latency were found between high and low performers (Dass et al., 2023). Furthermore, ET can measure levels of engagement through indicators such as fixation duration and pupil dilation during experiential learning tasks (Jaiswal et al., 2023).

ET data is inherently temporal and sequential, typically captured as time-series data consisting of gaze coordinates, fixation durations, saccades, and pupil dilations over time. This temporal nature requires ML methods capable of processing and learning from sequential patterns. Depending on the analysis goal, different ML paradigms can be employed: classification (e.g., distinguishing engaged from disengaged students), regression (e.g., predicting cognitive load or comprehension score based on gaze metrics), and clustering (e.g., grouping learners by similar reading behaviors or attention patterns). These algorithms have the advantage of being interpretable and computationally efficient, making them suitable for real-time applications in educational contexts. However, they often require manual feature engineering and struggle to capture complex temporal dependencies inherent in raw gaze sequences.

Classical ML algorithms such as SVM, DT, RF, Naive Bayes, and KNN have been used to analyze ET data. These models are particularly effective when gaze features are aggregated over time windows into summary statistics, such as mean fixation duration, number of saccades, or blink rate. For instance, classification tasks such as distinguishing between experts and novices based on gaze behavior in problem-solving scenarios can be effectively performed using SVM or RF. Similarly, simple linear regression models can estimate task difficulty or student performance from basic gaze metrics.

To better capture the temporal dynamics of ET data, more advanced models like RNN and their variants such as LSTM have been employed. These algorithms can process raw, unsegmented gaze sequences and learn temporal patterns associated with attention, cognitive load, or emotional arousal. Sequence models are particularly valuable when the timing and order of gaze events are crucial for interpretation, such as detecting mind wandering or predicting when a learner is about to disengage. However, RNNs and LSTMs often require larger datasets to train effectively and are less interpretable than traditional models.

At the frontier of ML applications in ET are Transformer-based models (Hollenstein et al., 2021), which have recently gained importance for their ability to model long-range dependencies in sequential data without relying on recurrent structures. Transformers use self-attention mechanisms to dynamically weigh the relevance of each gaze point in the sequence, making them highly effective in capturing both short- and long-term interactions between eye movements and cognitive states. This is crucial for converting the granularity of ET data into a format that can effectively inform teaching strategies. A review by Skaramagkas et al. (2021) emphasised the significance of ET metrics, such as pupil diameter and blink rate, in identifying emotional and cognitive states with the aid of AI. While most algorithms can recognise and classify levels or states of each process with relatively high accuracy, no studies have focused on discriminating between emotional and cognitive processes. This provides a current research gap in the literature and a prospective future research direction. The combination of ET data and AI offers valuable insights into human behaviour and emotions (Kędras & Sobiecki, 2023), and also has the potential to transform educational practices by providing real-time, gaze-based interventions to improve learning outcomes. However, while transformer models offer state-of-the-art performance, they come with increased computational demands and require extensive labeled data for training. Nonetheless, their capacity to model nuanced human behaviors from ET data makes them a powerful tool in the development of next-generation educational technologies.

Research suggests that AI-powered ET systems could provide educators with valuable insights into students' cognitive processes. Šola et al. (2024) emphasise the importance of analysing video content to assess its suitability for students (e.g. whether it is overwhelming, dull, or interesting). This assessment can be informed by several factors, including content complexity, information density, and audience engagement metrics.

This literature review aims to collect and analyse information on the ET technology, human-centred design aspects, ML techniques available, the solutions provided by our project partners, Blickshift and Lexplore, as well as public datasets.

4.1. ET technology

ET is frequently employed in cognitive and educational research to analyse reading behaviour (Soh, 2016; Özer & Özdemir, 2021), among other things. However, many modern ET solutions are expensive, primarily due to the high-end processing hardware required to analyse infrared camera images. They also typically necessitate a controlled laboratory environment, thereby limiting their scalability in real educational settings.

ET hardware can differ in various ways, such as the human interface, the tracking area, and the specifications (e.g. spatial resolution, sample rate, and accuracy). Several types of ET devices are available:

- *Head-stabilised ET* enables high-fidelity vision experiments, but limits participant comfort and natural interaction.
- *Mobile ET* devices (head-mounted) are ideal for real-world experiments.
- *Remote ET* does not require contact with the participant, enabling natural interaction in vision experiments; however, excessive head movement and sunlight are limitations of this type of ET. The most common remote ET methods are *webcam-based* systems and systems that use specialised infrared cameras.

4.1.1. Remote ET systems

Remote ET systems use cameras and software to monitor a user's gaze direction and eye movements without any physical contact. Unlike traditional eye trackers, they can capture these movements in more natural settings. Such systems are commonly used in research and user experience testing, as well as in gaming, to gain insight into how people interact with interfaces and environments.

The system can estimate the direction of the user's gaze and their point of focus by tracking the relative position of the pupil and glints. Sophisticated algorithms and software are employed to process image data, calibrate the system and determine gaze positions. Overall, remote eye-tracking systems are a valuable tool for understanding human behaviour and improving user experience in a variety of applications. Examples of Remote ET Systems include:

- *Tobii:* A well-known provider of both desktop and VR/AR ET systems, with a focus on usability and integration with various software and games (https://gaming.tobii.com/product/eye-tracker-5/).
- *SR Research (EyeLink):* Known for high-precision eye trackers, particularly the EyeLink 1000 Plus (https://www.sr-research.com/eyelink-1000-plus/).
- *RemoteEye:* An open-source, high-speed remote eye tracker developed at University of Tübingen (Hosp et al, 2020).

4.1.2. Webcam based ET

Recent advances in computer vision have paved the way for scalable ET in education using only a web camera system, which is available to wide masses of users. When combined with ML or DL, this provides an unobtrusive method of predicting reading comprehension levels by considering gaze-based features and textual characteristics. Furthermore, this brings ET technology out of controlled laboratory environments and into more natural settings. It can facilitate valuable insights into cognitive load, engagement levels and information processing strategies by tracking gaze patterns.

Webcam-based ET can be supported by libraries such as *WebGazer* (https://github.com/brownhci/WebGazer), which infer the eye gaze locations of website visitors on a page in real time. This technology can be integrated into any website that wishes to gain a better understanding of its visitors and improve their user experience. Furthermore, it runs entirely in the client browser.

Gaze.io is software that aims to enable desktop navigation and application control using the user's eye gaze as an input, with only the onboard webcam and in variable lighting conditions (https://github.com/asheeshr/gaze-io-system).

A review of webcam ET in education (Dostálová & Plch, 2023) showed that only a few studies aimed to experimentally explore cognitive processes. Working with ET data is a rather extensive process, ranging from basic on- and off-screen gaze tracking to analysing detailed metrics of saccades and fixations, also at the level of selected areas of interest. The choice of ET metrics depends on the main objectives of the research in question.

Guan et al. (2022) employed webcam ET alongside ML methods to investigate the correlation between reading behaviour and performance. The selected ET features were related to various fixation parameters and the frequency of regressions on pages. They found that the longer participants focused on an area, the poorer their reading performance was, a finding supported by several studies. However, technical constraints and the small number of participants limited the accuracy of this experiment.

Hutt et al. (2023) provided evidence for a scalable solution for predicting attention and reading comprehension using only stock web cameras, delivering sufficiently precise and accurate gaze measurements. They used Webgazer and converted the total gaze raw data into global gaze features (general eye movement data independent of the presented stimuli) and local gaze features (based on both gaze location and the relative screen content).

Research shows that webcam-based gaze features moderately correlate with eye tracker-based features, and that they can model user engagement and comprehension with a similar level of accuracy (Hutt & D'Mello, 2022). Other studies have demonstrated that free, open-source, webcam-based ET can be used to assess differences in reading patterns and online thought patterns (Wong et al., 2023). Nevertheless, the performance of webcam-based ET technology varies greatly depending on external factors such as the type of lighting or the position of the participant (Yüksel, 2023; Hagihara et al, 2024).

4.2. Human-centered design aspects

The development of AI-assisted educational tools, particularly those involving complex data such as ET, requires grounding in human-centered design (HCD) principles and alignment with human-centered AI (HCAI) frameworks. Research in AI in Education (AIEd) and Human-Centered Learning Analytics (HCLA) aims to identify and address the challenges that may arise while deploying AI-based methods in educational technologies and tools (Yang et al. 2021), highlighting the importance of involving end-users, especially teachers and students, throughout the design process to ensure relevance, usability, and interpretability (Topai et al. 2024). However, there is still a gap between the potential of LA and its actual implementation in educational practice. Furthermore, there is the challenge of developing systems that are not only technically robust but also socially and ethically aligned. HCAI frameworks have been proposed to support the development of LA-based tools in line with HCAI principles such as safety, reliability, and trustworthiness (Alfredo et al. 2024). For example, balancing model accuracy with model interpretability can increase trustworthiness and reliability. In the context of our tool, these principles suggest the need for human-centered approaches, including transparent modeling of gaze-based data, ensuring that models do not systematically misrepresent student behavior across different reading strategies or demographics, and human-in-the-loop mechanisms for teachers to control and override AI decisions.

In terms of model functionalities, the objective of the ML models should be in line with the purpose of the LA dashboard and its affordances. Different types of dashboards require different visualizations and underlying ML models (van Leeuwen, 2019). *Mirroring* dashboards focus on displaying indicators to reflect existing data without making judgments or recommendations. *Alerting* dashboards aim to communicate data transformations and employ heuristics or ML models to notify users when certain behaviors are detected, typically to indicate potential issues. *Advising* or *guiding* dashboards aim to provide justifications and actionable recommendations to improve interpretation and learning outcomes. Another taxonomy classifies dashboards into *explanatory* – focusing on addressing teachers' specific questions, and *exploratory* – focusing on sensemaking (Chen et al., 2019). Based on their purpose, these types of dashboards include diverse types of LA analytics, considering their level of analysis, namely: *descriptive*, *diagnostic*, *predictive* and *prescriptive*. Depending on the purpose of a dashboard and the type of LA, several modeling approaches can be utilized, including predictive and clustering methods, recommendation systems, and control algorithms. For example, a mirroring dashboard could utilize descriptive LA to visualize the progress of student groups using an unsupervised ML method for clustering, while a diagnostic dashboard could employ an explainable predictive model which can provide a justification for a detected event.

Focusing on how to make intelligent tutoring systems and dashboards transparent to teachers, a taxonomy has been developed to classify systems based on the types of the underlying components or models used (Ley et al., 2023). There are three main types of models used in LA dashboards and adaptive learning systems; *learner model*, which describes the student's skills, learning strategies, etc., *domain model*, which characterizes task aspects, including task difficulty and learning goals, and *instructional model,* which can dictate appropriate teaching strategies, such as hinting, feedback and explanations. These models may use different types of data: (a) (*multimodal) learning analytics*, which provide information about the student learning processes, (b) c*ontextual data,* which provide information related to the task (activity) and the environment, and (c) *self-reported data* which are provided by users (teachers and students) either as self-assessment measures (e.g., perceived performance) or user's perception related to their interaction with the system (e.g., perceived trust and usability).

Visualizations of gaze data have been used to support both students and teachers to provide them with insights about how they process text, videos, and graphs. In order to communicate the context of reading activities of students, a set of dynamic and static visualizations has been designed (Spakov et al., 2017). Gaze and Word Replay are dynamic, animated views which visualize the reading activity over time and can help teachers to understand how the individual students read, as well as to compare the skills of a group of students reading the same text. Word Reading durations and Student Summary are static views that summarize the reading activity data and provide an overview of students and their reading activity, e.g., active vocabulary. In the context of remote instruction for programming, visualizations are used to summarize multiple students' learning behavior through clustering of gaze data accompanied with contextual information, e.g., scrolling bar location (Yao et al., 2018). The goal is to examine how gaze visualizations from multiple students impact instruction and understanding of the students' perspective, in order to develop a gaze-based tool as a feedback tool for students, as an informative tool for teachers, as well as a tool which can support instructional design revisions (Boels, 2022). The concept of interactive visualization has been introduced to enhance the usability of HCAI-enabled tools, along with a set of design guidelines for constructing visual representations in support of HCAI principles (Hoque et al., 2024), including *Tackling Human Concerns Directly* – addressing human concerns directly in the design of HCAI tools for fostering trust, transparency, and acceptance, *Encouraging Interaction* – fostering human-data and human-AI interactions to enhance user engagement and understanding, and *Show, Don't Tell* – designing visualizations that lay out the data landscape without steering users towards a specific interpretation, and support user agency and autonomy. Furthermore, designing effective visualization guidance mechanisms for teacher-facing dashboards should consider teachers' *Visual Literacy (VL)* in order to enhance teachers' interpretation by: (i) highlighting key elements with heuristics, (ii) pairing visuals with descriptive text, (iii) removing non-essential details, and (iv) testing different dashboard views (Pozdniakov et al., 2025).

4.3. ML-techniques used in ET systems

In their review of ET metrics used to detect emotional and cognitive processes, Skaramagkas et al. (2021) identified visual attention and cognitive workload as key metrics. Visual attention enables individuals to focus on relevant details while ignoring distractions. Cognitive workload, on the other hand, reflects the mental effort required for a task and is influenced by factors such as task complexity, presentation style and processing effort. The authors also addressed the automatic recognition of these processes through ML techniques. Most studies exploit features derived from pupil diameter and blink characteristics. The SVM classification system is either the one used in most of the studies, or the one that helps researchers achieve the highest accuracy in predicting emotions. This is possibly due to the various advantages of SVM classifiers, particularly with regard to relatively small datasets and clear class separation.

A systematic literature review (Arnold et al., 2025) revealed that ET technologies and ML play a transformative role in our understanding of reading behaviour, attention, and cognitive workload. However, challenges such as data scarcity, limited generalisability and biases in existing methodologies remain. Closing these gaps by developing standardised frameworks, creating diverse datasets, and advancing synthetic data generation could improve the accessibility, accuracy, and real-world applicability of ET solutions for enhancing reading comprehension and focus. Studies using hybrid models (such as CNN–LSTM combinations) have demonstrated consistently superior accuracy compared to traditional methods. These results corroborate earlier research emphasising the importance of model architecture and dataset quality for achieving reliable eye movement classifications. This review identifies opportunities to advance ET applications in reading. Integrating advanced ML models into commercial ET software could improve the real-time analysis of reading behaviours, which would benefit education. Mobile ET systems also present opportunities for improvement. Developing lightweight algorithms and adaptable hardware could improve accuracy in environments with limited resources, such as classrooms, by mitigating environmental challenges such as lighting variations and head movement.

ML algorithms have played a key role in improving the accuracy, calibration and quality of ET data. Classification and regression models have successfully identified and predicted gaze error levels in different conditions (Kar, 2020). Conventional ML models, such as RF, SVM, KNN, gradient boosting and DT, have been used to analyse eye movement data and categorise visual attention patterns. A review examined the use of ML and DL in the development of reliable eye movement classification (Fikri et al., 2021). ML methods have also been used to enhance the accuracy of webcam-based gaze estimation (Jain et al., 2024). The field of DL approaches to ET is growing, with some researchers using the video frames of low-cost web cameras (Zdarsky et al., 2021).

AI models, such as convolutional neural networks (CNNs), can be used effectively to interpret biometric data, including ET, in order to assess cognitive states of learners. Such models can provide valuable insights into individual engagement and cognitive activity, which are critical for improving educational outcomes. In particular, the CNN model has demonstrated superior performance in interpreting complex data patterns (Jamil & Belkacem, 2024).

Previous studies have addressed reading comprehension using various modelling approaches, such as SVM with a Fisher kernel derived from generative models of individual gaze patterns (Makowski et al., 2019), recurrent and convolutional neural networks (Ahn et al., 2020), linear regression models (Mézière et al., 2023) and multimodal language models (Shubi et al., 2024). These studies emphasise the complexity of the task and highlight the need for more advanced and effective modelling techniques.

4.4. Blickshift solutions

Understanding how people read and process written information is a key research area in cognitive psychology, education, linguistics, and human-computer interaction. With **Blickshift Analytics** (www.blickshift.com), researchers have access to a specialized, interactive software solution that brings structure, efficiency, and depth to the analysis of eye movement data in reading studies.

4.4.1. Comprehensive Tool for Gaze Data Analysis

Blickshift Analytics (Fig. 1) offers a timeline-based approach to analyzing ET data, allowing researchers to explore complex gaze datasets across time and experimental conditions. The platform supports data from all major ET systems and provides a suite of tools to preprocess, visualize, and quantify eye movements.

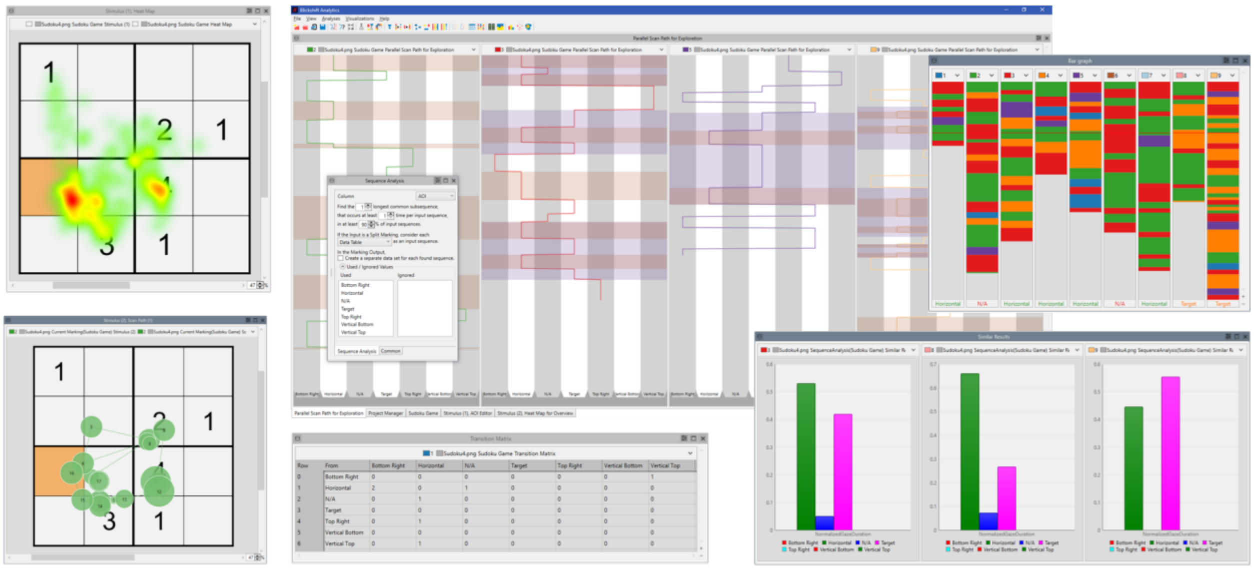

Fig. 1: Blickshift Analytics is a specialized, interactive software solution that brings structure, efficiency, and depth to the analysis of eye movement data.

In reading studies, participants' gaze behavior is typically recorded while they read text presented on a screen. These recordings capture fixations (moments when the eyes remain relatively still and visual information is processed), saccades (rapid eye movements between fixations), regressions (backward saccades), and scanpaths (sequences of eye movements). Blickshift Analytics allows users to analyze all these components in a flexible and highly visual environment.

4.4.2. Core Features Supporting Reading Research

Timeline-Based Exploration

The timeline view in Blickshift Analytics provides an intuitive way to explore individual and aggregated gaze behavior. Researchers can scroll through time-synchronized visualizations of fixations and saccades and correlate them with the underlying text stimuli. This is especially helpful in identifying when and where readers encounter difficulties or pause for longer periods.

Area-of-Interest (AOI) Analysis

In reading studies, specific words, phrases, or text regions are often marked as Areas of Interest (AOIs). Blickshift Analytics supports the fast and flexible creation of AOIs and provides detailed metrics such as:

- Total fixation duration per AOI

- Number of fixations
- First fixation duration
- Time to first fixation
- Dwell time
- Refixation counts

These metrics help to quantify the reader's attention to specific linguistic features (e.g., keywords, difficult vocabulary, syntactic structures), visual elements (e.g., bold or italic text), or design components (e.g., headlines, navigation links).

Regression and Re-reading Analysis

Regressions—backward saccades to previously read text—are a critical indicator of reading difficulty or complex information processing. Blickshift Analytics makes it easy to identify and analyze such regressions. The software provides visual representations and statistical summaries of re-reading behavior, helping researchers investigate:

- Which text segments are re-read
- How frequently regressions occur
- Whether regressions correlate with text difficulty or comprehension

Comparison Across Participants and Conditions

Reading behavior often varies based on the individual (e.g., age, reading ability) or the experimental condition (e.g., type of text, font style, layout). With its powerful filtering and grouping functions, Blickshift Analytics allows researchers to compare gaze data across multiple participants or study conditions. Metrics can be aggregated, averaged, and exported for statistical analysis or used directly within the platform for visual comparisons.

Advanced Data Integration

Blickshift Analytics can be combined with other experimental data, such as:

- Behavioral responses (e.g., comprehension questions)
- EEG or fNIRS signals (for neurocognitive research)
- User interactions (e.g., scrolling behavior in digital reading)

This makes it an excellent choice for multimodal studies where ET is just one part of the research setup.

Flexible Export and Reporting

The software supports detailed data export to CSV format for further analysis in R, Python, or SPSS. Visualizations such as scanpaths, heatmaps, and fixation plots can be exported as images or integrated into automated reports. This ensures that insights can be communicated clearly to stakeholders, collaborators, or broader audiences.

4.4.3. Applications in Reading Studies

Blickshift Analytics can be used to investigate a wide range of questions in reading studies, including:

- How do readers process complex or technical texts?
- Which parts of a sentence draw the most attention?
- How does typography affect reading fluency?
- What is the difference in gaze behavior between skilled and less-skilled readers?
- How do different age groups approach digital reading?

Blickshift Analytics is a powerful solution for researchers who want to analyze eye movements with scientific precision and efficiency. In reading studies, where the nuances of attention, comprehension, and processing speed matter, the software offers the right combination of interactivity, structure, and analytical power.

4.5. Lexplore solutions

Lexplore is a service for systematic reading development that combines ET with AI to evaluate students' reading abilities quickly and objectively (www.lexplore.com). The service helps schools work systematically on reading development by screening all students, providing insights into their reading levels, and offering automated recommendations to improve their reading skills. Additionally, Lexplore promotes equity by enabling data-driven decision-making and systematic follow-up for all students.

The service is based on three components: Assessment, Insights and Reading – which each will be outlined in the next sections, in addition to a summary of beneficial experiences of using Lexplore.

4.5.1. Assessments

Lexplore's assessment was developed at Karolinska Institutet in Sweden, based on over 30 years of scientific research. By combining ET and AI, a student's reading ability can be analyzed in just a few minutes. The service provides teachers with real-time data for interventions that directly correlate with students' reading abilities. The screening is quick, objective, and accurate, while also enabling individualized learning and resource optimization.

Benefits of Assessing with Lexplore:

- Objective and equitable.
- Time-efficient (< 5 min) and as accurate as more time-consuming methods.
- Measures both oral and silent reading.
- Results are delivered immediately with visual graphs that are easy to understand and provide powerful visualizations of reading.
- Enables data-driven decision-making and helps teachers monitor progress over time.
- Students find it fun - which rarely is the case with traditional methods.

4.5.2. Insights

Lexplore provides powerful insights into students' reading development that are easy to understand and use. These insights are delivered in the form of visual graphs, offering a clear picture of reading progress at the organizational, school, group, and individual levels. The insights allow for early interventions to improve results and help teachers plan resources effectively.

Benefits of Insights from Lexplore:

- Ability to track progress over time and tailor interventions and exercises for each student.
- Support for data-driven decision-making and equitable education with unbiased data.
- Insights gathered in one place make it easy for teachers to understand each student's development and adjust teaching.
- Automated recommendations for fluency training and books at the right level for each student.

- Support for engaging students in their own reading development and creating a platform for collegial learning and discussion among teachers.

4.5.3. Reading

To help students develop their reading skills, Lexplore offers several interventions, including fluency training and automatic book recommendations based on the student's reading level. These interventions are designed to support each student at their current level, aiming to help all students reach their full potential.

Support for Reading from Lexplore:

- Fluency training designed to improve students' decoding and reading fluency. Based on the research model "The Simple View of Reading," students practice the connection between sounds and letters in a structured way. The exercises are tailored to the student's reading level and can be done individually or in groups.
- Teaching material aimed at creating balanced reading instruction that increases engagement. With the support of digital training, teachers can focus on decoding, fluency, language comprehension, and reading strategies, leading to improved reading comprehension. The material is designed for grades 1-9 and developed by special educators based on international research.
- Library that makes it easy for students to log their reading, both for physical and digital books. This tool helps include parents in the reading development and allows tracking of students' reading activities in one place. Students gain access to a high-quality library with engaging books tailored to their reading level, and they can earn points, challenge friends, and take quizzes to boost reading comprehension.

4.5.4. Experiences

Over one million assessments have been conducted using the method, and last school year, 200,000 students were tested across nearly 1,000 schools. Sweden, Norway, and the UK are the primary markets, but through partnerships, Lexplore also reaches international schools in countries such as Portugal, Brazil, South Africa, and India.

Experiences of Lexplore:

- Lexplore helps schools prioritize reading and work systematically on reading development. By screening all students and tracking their progress over time, municipalities like Lidingö have seen an increase in strong readers and a decrease in struggling readers. Lexplore provides quick and objective insights that support teachers in working with decoding, reading strategies, and comprehension.
- Lexplore is time-efficient and saves teachers' valuable time. Through quick screening and automatic recommendations, teachers can focus on teaching instead of time-consuming assessments. Screening results are available within seconds, making it easier for teachers to act swiftly on the insights.
- Lexplore motivates both students and teachers. By screening in a fun and engaging way, students get the chance to be involved in their own reading development. Level-appropriate content based on the screening boosts students' confidence and motivation by giving them the right challenges at the right time. Automatic fluency training recommendations also make teaching easier and more motivating for teachers.
- Lexplore promotes equity by giving all students the same opportunities to develop their reading skills. By using objective data, teachers can ensure that each student receives the right support, and the insights help plan early interventions to improve outcomes. An example of this is Kristinehamn municipality, where systematic collaboration between schools led to greater equity for students.
- Working on reading skills in all subjects can have a significant impact on students' grades. Uppsala municipality found that when students' reading skills improved across all subjects, their grades increased. Students went from reading silently in class to reading aloud with their peers, engaging in text discussions, and practicing vocabulary, leading to immediate results.

4.6. Public Datasets

WebQAmGaze (Ribeiro et al., 2023) is a multilingual dataset of 600 participants. The stimuli consist of performing a read-at-your-own-pace reading task, a comprehension quiz, a question-oriented information seeking task, and a respective question-answering prompt. The Eye-Tracker is a webcam. The data includes stimuli, gaze metrics, and questionnaire information. The dataset is accessible and has a working codebase. It is a small dataset with 700 MB in size.

WEyeDS (Evdokimov et al., 2024) is a dataset of 38 participants. The stimuli consist of an abstract prosaccade task (react to color-changes of dots on the screen). The Eye-Tracker is a webcam. The data includes stimuli and face images. The dataset is inaccessible due to the author's non-response, but it has a working codebase. The data size is unknown.

CopCo (Hollenstein et al., 2022) is a dataset of 22 Danish-speaking participants. The stimuli consist of reading Danish text of speeches, followed by a comprehension quiz. The Eye-Tracker is an EyeLink 1000 Plus. The data includes stimuli and common eye metrics (fixations, saccades). The dataset is accessible. The codebase can be installed using manual fixes, but breaks API of used (outdated) libraries - generally requires extensive fiddling to work properly. The total size is 4 GB.

Eye-Tracking Dyslexia Dataset (**ETDD70**) (Sedmidubsky et al., 2024) is a dataset of 70 young (9-10 years) czech-speaking participants with a 50-50 split by dyslexia. The stimuli consists of three reading tasks (syllables only, meaningful text, pseudo text). The Eye-Tracker is unknown. The data includes stimuli and common eye metrics (fixations, saccades). The dataset is accessible, but has no codebase. The total size is 600 MB.

Potsdam Textbook Corpus (**PoTeC**) (Jakobi et al., 2025) is a dataset of 75 German-speaking participants (university students) grouped by reading comprehension. The stimuli consist of 12 demanding (university level) reading comprehension tasks followed by a multiple-choice quiz. The Eye-Tracker is an EyeLink 1000. The data includes stimuli, common eye metrics (fixations, saccades) and many derived features (word-, text- ,reader-, fixation-, trial-level). The dataset is accessible and has a working codebase. The total size is 3 GB.

Multilingual Eye-Tracking Corpus (**MECO**) (Siegelman et al., 2022) is a multilingual (13) dataset with approximately 50 participants per language. The stimuli consist of a reading task followed by a questionnaire. The Eye-Tracker is a EyeLink Portable Duo 1000/1000+. The data includes stimuli, common eye metrics (fixations, saccades). The dataset is accessible. The codebase is written in R and works using RStudio. The total size is 150 MB.

Ghent Eye-Tracking Corpus (**GECO**) (Cop et al, 2017) is a dataset of 19 Dutch/English bilinguals and 14 English monolinguals. The stimuli consists of reading an entire book (5000 sentences) followed by a quiz. The Eye-Tracker is an EyeLink 1000. The data includes stimuli and common eye metrics (fixations, saccades.). The dataset is accessible; however, no codebase is available. Several analyses were done in R scripts according to the paper. The total dataset's size is 500 MB. (Note: Hollenstein & Beinborn (2021) reported an alignment error in GECO and proposed a solution. We will further investigate this issue and carefully evaluate whether the GECO dataset should be excluded from the EYE-TEACH system training and development.)

ETH-XGaze (Zhang et al., 2020) is a dataset of 110 participants. The stimuli consist of head poses, gaze points, and lighting conditions. The Eye-Tracker was a rig setup of 18 Canon 250D cameras. The dataset is accessible (comes in three different image resolutions, resulting in different sizes: 130 GB, 500 GB, 7 TB). The codebase works. The dataset supports 3D gaze estimation, robustness testing under extreme head poses and lighting, cross-person/domain generalization, and evaluation at multiple image resolutions. Main metric: angular error in predicted gaze direction.

The **Dundee** Corpus (Kennedy, 2003) is a dataset of 10 English-speaking and 10 French-speaking participants. The stimuli consist of reading newspaper articles in their respective language. The Eye-Tracker is a "Dr Bouis Oculometer Eyetracker". The dataset is not public and no codebase exists. The data size is unknown. The original paper seems unavailable.

ZuCo (Hollenstein et al., 2018) is a dataset of 12 participants recording eye-tracking and EEG data. The stimuli consists of 3 reading tasks (sentiment analysis, comprehension, task-specific) followed by quizzes. The Eye-Tracker was EyeLink 1000+, The EEG system was a 128-channel HydroCel Geodesic Sensor Net (HCGSN). The data includes stimuli, EEG data, and common eye metrics (fixations, saccades). The dataset is accessible. The codebase exists (Matlab based). The total dataset's size is 100 GB.

ZuCo 2.0 (Hollenstein et al., 2020) consists of simultaneous eye-tracking and EEG during natural reading annotation. It contains gaze and brain activity data of 739 English sentences, 349 in a normal reading paradigm and 390 in a task-specific paradigm, in which the 18 participants actively search for a semantic relation type in the given sentences as a linguistic annotation task. The complete dataset contains about 70 GB of files.

CELER (Berzak et al., 2022) is a dataset of 365 L1 and L2 English participants. The dataset covers several readings tasks followed by a comprehension quiz. The Eye-Tracker was an EyeLink 1000/1000+. The data includes stimuli, common eye metrics, and derived metrics from two analyses ((i) eye movement measures, and (ii) the effect of frequency, suprisal, and word length on reading times). The dataset requires a paid license, a codebase exists, but has not been tested. The total dataset's size is unknown.

Provo corpus (Luke & Christianson, 2018) aims to correlate ET with predictability norms for cloze scores, the orthographic form, morphosyntactic, and semantic information for each read word i.e. some words are skipped and some are fixated upon. The corpus was gathered in two stages: the survey stage to create the predictability norms and the second stage of reading/eye-tracking. A total of 470 participants were reached. The Eye-Tracker was an EyeLink 1000+. The data includes stimuli, predictability norms and common eye metrics. The dataset is accessible, but has no codebase. The total size is 1 GB.

Valliappan et al. (2020) dataset is a study of 17 participants participating in 10 English reading comprehension tasks and follow-up quizzes on their smartphones. The data was collected using a special-designed app. It includes gaze estimates, no face images, and must be requested. The trained models must also be requested. There is supplementary material that explains implementation details, but no codebase.

Potsdam-Allahabad Hindi Eyetracking Corpus (**PAH**) (Husain et al., 2015) aims to characterize reading difficulty by word-level metrics (such as syllable length). A total of 30 participants perform native reading tasks of Hindi-Urdu texts while being eye-tracked. The Eye-Tracker was a SMI iView X HED. The dataset exists as an R-language package (2 MB) with 33 features.

OneStop (Berzak et al., 2025) is a large-scale English corpus of eye movements in reading with 360 L1 participants, 2.6 million word tokens, and 152 hours of recorded ET data. The dataset was collected using an EyeLink 1000 Plus eyetracker. The dataset release includes Interest Area Reports (features aggregated at the word level), Fixation Reports (features aggregated at the level of fixations/saccades). The raw data is in EDF and ASCII formats, and consists of a detailed participant questionnaire. To facilitate analyses, the dataset's authors further provide precomputed text annotations: word length, frequency, and surprisal, part-of-speech tags, and syntactic dependency trees. The full dataset is 3.65 GB.

MQA-RC (Sood et al., 2020) is a dataset of 23 participants. The stimuli consist of reading movie plots and follow-up questions, allowing researchers to observe changes in reading behavior in three comprehension tasks. They use 32 movie plots, with corresponding question-answer (QA) pairs, from the benchmark MovieQA dataset~\cite{tapaswi2016movieqa}. The Eye-Tracker was Tobii 600Hz. The dataset is available on request. The codebase is available.

SB-SAT (Ahn et al., 2020) decodes a reader's eye movements to predict their level of text comprehension and related states. Eye movements were recorded from 95 people reading 4 published SAT passages, each followed by corresponding SAT questions and self-evaluation questionnaires. A sequence of 21 fixation-location (x,y), fixation-duration, and pupil-size features were extracted from the reading behavior. The full dataset is not available. A version of train, validation, and test sets are available with a codebase. However, the codebase was not well-organized and used the outdated Tensorflow framework, requiring much effort to update and test.

IITB-Hallucination Gaze corpus (**IITB-HGC**) (Maharaj et al., 2023) aims to detect hallucination from gaze signals from human. The authors employed 5 annotators for annotating 500 instances of claim-context pairs, carefully derived from the FactCC dataset. During the annotation process, they captured the fixation durations of annotators on both the claim and context texts, along with their corresponding labels. The authors identified two attention strategies: global attention, which focuses on the most informative sentence, and local attention, which focuses on important words within a sentence. The dataset's size is quite small, e.g., 7 MB, and consists of 2500 rows. The codebase is available but is dedicated to demonstrate hallucination detection.

InteRead (Zermiani et al., 2024) consists of ET reading dataset with interruptions, comprising gaze data collected from 57 participants engaged in an interrupted reading task. There are 5247 tokens spanning across 28 pages of text, six of which include interruptions. The authors used a Tobii Pro Spectrum screen-based eye tracker operating at a sampling frequency of 1200 Hz. All stimuli were presented on the native Tobii Pro Spectrum screen (EIZO FlexScan EV2451) with dimensions of 52.8x29.7cm and a resolution of 1920x1080px. Two Python-based frameworks were used: PsychoPy and PyGaze. However, the codebase is written in R. The dataset's size is 950 MB.

Lee et al. (2024) dataset raises a caution against using different comprehension tests interchangeably in reading research and highlights the importance of test format and construct validity. This study investigates the "jingle fallacy" in reading comprehension tests—the mistaken belief that similarly named tests measure the same construct—by comparing the Nelson Denny Reading Test (NDRT) and the Wechsler Individual Achievement Test (WIAT-II). Eye movements were recorded from the right eye only using an EyeLink 1000 tracker. The dataset includes eye-tracking and comprehension accuracy data from 91 university students reading passages under high/low comprehension demands. It also contains standardized scores from the

NDRT and WIAT-II tests. Key eye metrics include fixations, saccade lengths, and reading times. The authors found that high scorers on the NDRT adjusted their reading strategies more efficiently (e.g., faster reading, fewer fixations) than low scorers, especially as trials progressed. A cleaned version of the data (3 MB) and R scripts are publicly available; however, it is for demonstration purposes.

4.6.1. Advantages

The currently collected ET datasets for reading comprehension offer several advantages. A significant strength is the linguistic diversity represented across the datasets, including multilingual corpora like MECO and WebQAmGaze, as well as monolingual datasets in underrepresented languages such as Czech (ETDD70), Hindi (PAH), and Danish (CopCo). This broad coverage facilitates cross-linguistic comparisons and multilingual model training. Furthermore, the reading tasks within these datasets are varied and well-aligned with comprehension goals, ranging from naturalistic book reading (GECO), comprehension quizzes (PoTeC, Lee et al.), and question-answering (MQA-RC), to more complex task-specific reading involving annotation (ZuCo 2.0). Some datasets, ZuCo and ZuCo 2.0, integrate EEG data, providing multimodal insights into cognitive processes during reading. However, it might be different in our project's settings. In addition, several datasets such as OneStop and Provo include rich annotations at multiple levels (e.g., word, sentence, and fixation), offering valuable resources for linguistic and psycholinguistic analyses. Overall, many datasets are accompanied by working codebases and are openly accessible, supporting reproducible research and practical experimentation.

4.6.2. Disadvantages

Despite these strengths, several limitations hinder the usability and consistency of the available datasets. Accessibility remains a key issue, as some datasets are either only partially available (e.g., SB-SAT), require licensing (e.g., CELER), or lack clear channels for obtaining data (e.g., Valiapan). In terms of technical usability, codebases are not uniformly available or functional; for instance, datasets like GECO and ETDD70 lack codebases altogether, while others such as CopCo and SB-SAT rely on outdated libraries or frameworks that require substantial effort to fix. Another challenge lies in sample size disparity: while some datasets include hundreds of participants (e.g., WebQAmGaze, MECO), others have relatively small cohorts (e.g., ZuCo, CopCo), which may affect the statistical power and generalizability of findings. Additionally, the heterogeneity of ET hardware and sampling rates—from low-resolution webcam-based tracking to high-end systems like EyeLink 1000+—introduces variation in data quality and interpretability. This inconsistency extends to the ET metrics provided, with some datasets lacking standard features or sufficient metadata to enable fine-grained analysis.

4.6.3. Recommended Datasets for Reading Comprehension Tasks

Among the available resources, several datasets stand out for their completeness, usability, and relevance to reading comprehension. The OneStop corpus is a premier example, combining a large participant base, extensive ET data, and detailed linguistic annotations suitable for both NLP and cognitive modeling. PoTeC is also a strong candidate, featuring comprehension quizzes, a structured set of academic reading tasks, and derived features that enable complex analysis. For multilingual research, MECO is particularly valuable due to its broad language coverage and reliable R-based codebase. In contexts requiring cognitive or neural interpretation, ZuCo and ZuCo 2.0 provide a unique combination of ET and EEG data during diverse reading tasks. Finally, the Provo dataset offers a well-annotated corpus focusing on lexical predictability and fixation behavior, making it suitable for psycholinguistic modeling and reader variability studies. These datasets, taken together, provide a representative and versatile foundation for empirical studies in ET-based reading comprehension research.

4.6.4. Ethical and Privacy aspects

WebQAmGaze: Informed consent to participation and use of data for research purposes (Ribeiro et al., 2023); ethical approval by Ethics Commission of the Faculty of Humanities of the University of Copenhagen (no approval number/code reported). **Further information will be requested from the study authors regarding the participants' consent to the reuse of their data for other research purposes compatible with the training of AI systems.** Fully anonymized dataset is publicly available under Creative Commons Attribution 4.0 International Public License (https://github.com/tfnribeiro/WebQAmGaze).

WEyeDS: No information on informed consent and ethical approval of the study is available (Evdokimov et al., 2024). **This information will be requested from the authors of the study if the dataset is chosen for the training phase of the AI system.** Fully anonymized dataset is available “only for research purposes”. In order to receive a download link, you need to provide an institutional email, your name and a brief description of how you intend to use the dataset (https://robbinslab.github.io/weyeds/).

CopCo: Informed consent to participation and reuse of data (Hollenstein et al., 2022); ethical approval by the Research Ethics Committee at the Faculty of Humanities of the University of Copenhagen (no approval number/code reported). Fully anonymized dataset is publicly available under CC-By Attribution 4.0 International (https://osf.io/ud8s5/).

ETDD70: Ethical approval by the Research Ethics Committee of Masaryk University (Sedmidubsky et al., 2024) in Brno, Czech Republic (no approval number/code reported). **Information regarding the informed consent given by study participants will be requested from the authors if the dataset is chosen for the training phase of the AI system.** Fully anonymized dataset is publicly available under CC-By Attribution 4.0 International (https://zenodo.org/records/13332134).

PoTeC: Informed consent to participation (Jakobi et al., 2024) and use of data for research purposes (no approval number/code reported). **Further information is needed regarding participants' consent to the reuse of their data for other research purposes compatible with training AI systems.** Fully anonymized dataset is publicly available under CC-By Attribution 4.0 International (https://github.com/DiLi-Lab/PoTeC).

MECO: Informed consent to participate and use of data (Siegelman et al., 2022); Ethics clearance was obtained by each participating site from the ethics research board of the corresponding institution or country (no approval number/code reported). **Further information is needed especially regarding participants' consent to the reuse of their data for other research purposes**. Fully anonymized dataset is publicly available under CC-By Attribution 4.0 International (https://meco-read.com/).

GECO: No information on informed consent and ethical approval of the study is available (Cop et al., 2017) . **In case the dataset is suitable for training the AI system, full information on ethical and privacy aspects will be required.** Dataset is available: https://expsy.ugent.be/downloads/geco/.

OneStop: All the participants provided written consent to participate in the experiment (Berzak et al., 2025) under IRBs protocols 1605559077 (MIT) and 183157-2022 (Technion). **Further specification is needed regarding participants' consent to**

the reuse of their data for other research purposes. Fully anonymized data is available under CC-By Attribution 4.0 International (https://lacclab.github.io/OneStop-Eye-Movements/).

ETH-XGaze: No information on informed consent and ethical approval of the study is available (Zhang et al. 2020). **In case the dataset is suitable for training the AI system, full information on ethical and privacy aspects will be required.** The ETH-XGaze Dataset is available on request (cf. https://ait.ethz.ch/xgaze).

The Dundee Corpus: No information on informed consent and ethical approval of the study is available (Kennedy et al., 2013). **In case the dataset is suitable for training the AI system, full information on ethical and privacy aspects will be required.** The dataset is not publicly available.

ZuCo: All participants gave written consent for their participation and the re-use of the data prior to the start of the experiments (Hollenstein et al., 2018). The study was approved by the Ethics Commission of the University of Zurich (no approval number reported). Fully anonymized data is available under CC-By Attribution 4.0 International (https://statics.teams.cdn.office.net/evergreen-assets/safelinks/1/atp-safelinks.html).

ZuCo 2.0: All participants gave written consent for their participation and the re-use of the data prior to the start of the experiments (Hollenstein et al., 2020). The study was approved by the Ethics Commission of the University of Zurich (no approval number reported). Fully anonymized data is available under CC-By Attribution 4.0 International (https://osf.io/2urht/).

CELER: All the participants provided written consent to take part in the experiment (Berzak et al., 2022). There is no reference to ethical approval of the study. **Further specification is needed regarding participants' consent to the reuse of their data for other research purposes**. Fully anonymized dataset is available under CC-By Attribution 4.0 International (cf. https://github.com/berzak/celer).

Provo: No information on informed consent and ethical approval of the study is available (Luke & Christianson, 2018). **In case the dataset is suitable for training the AI system, full information on ethical and privacy aspects will be required.** Fully anonymized data is available under CC-By Attribution 4.0 International (https://osf.io/sjefs/).

Valliappan et al. 2020: Each participant provided their explicit and informed consent to data collection by reading and signing a study-specific participant agreement that informed them about collecting the front-facing camera feed for research analyses purposes, and the potential risks involved in performing gaze tasks for several minutes (e.g., eye strain, fatigue). Gaze estimates (inferred x- and y-locations on screen) for the studies are available from the corresponding author upon request. To protect study participant privacy and consent, the captured full face image data are not publicly available (Valliappan et al., 2020).

Potsdam-Allahabad Hindi Eyetracking Corpus: No information on informed consent and ethical approval of the study is available (Husain et al., 2015). **In case the dataset is suitable for training the AI system, full information on ethical and privacy aspects will be required.** Dataset is available: https://rdrr.io/github/vasishth/lingpsych/man/df_hindi.html.

Russian Sentence Corpus: All participants gave written informed consent in Russian in accordance with the Declaration of Helsinki (Laurinavichyute et al., 2019). The study was approved by the local Institutional Review Board. **Further specification is needed regarding participants' consent to the reuse of their data for other research purposes**. Fully anonymized data is available under CC-By Attribution 4.0 International (https://www.hse.ru/en/neuroling/research/RSC ; https://osf.io/x5q2r/).

MQA-RC: No information on informed consent and ethical approval of the study is available (Sood et al., 2020). **In case the dataset is suitable for training the AI system, full information on ethical and privacy aspects will be required.** Data can be requested by filling out the EULA license agreement (https://collaborative-ai.org/research/datasets/MQA-RC/).

SB-SAT: No information on informed consent and ethical approval of the study is available (Ahn et al., 2020). **In case the dataset is suitable for training the AI system, full information on ethical and privacy aspects will be required.** Dataset is available without information about license: https://github.com/ahnchive/SB-SAT.

IITB-HGC: Informed consent to participate and use of data. No information on data reuse and ethical approval (Maharaj et al., 2023). **Further specification is needed regarding participants' consent to the reuse of their data for other research purposes**. Fully anonymized data is available under MIT license (https://github.com/kishanmaharaj/gaze-hallucination-detection)

InteRead: Informed consent to participate and reuse and share the data; ethical approval (n. Az. 22-018) by Ethics Committee of University of Stuttgart (Zermiani et al., 2024). Fully anonymized data is available under CC-By Attribution 4.0 International (https://osf.io/43j5f/).

Lee et al. 2024: Study approved by the University of Southampton Ethics and Research Governance Board (Lee et al., 2024). Informed consent to participate. All data and material are available at https://osf.io/dk82u/?view_only=88d54bdc1f9b4793ba4b2dfebf106085.

5. System Requirements

Reading behaviour is characterised by distinct patterns of eye movement. Two critical aspects of these movements are saccades and fixations. A saccade is the movement of the eye between fixations; a fixation is the initial position of the eye.

We focus on text-specific understanding: the aim is to predict whether a reader will correctly answer a comprehension question about a specific text based on their eye movement patterns while reading it. From a cognitive science perspective, this task facilitates the investigation of effective versus ineffective reading strategies and their manifestation in gaze behaviour.

The typical process of ET involves acquiring eye movement data, choosing indicators and selecting an evaluation technique, before analysing the data. Modern video-based ET systems can capture eye movements with great precision in terms of both time and space. This enables researchers to distinguish gaze locations at the level of individual words, or even letters. The resulting raw gaze data comprises sequences of gaze positions that are time-stamped and analysed to identify and segment fixations and saccades. These low-level events can be aggregated further to derive reading measures at the word level and summarised at broader levels, including aggregates at the sentence or paragraph level. This allows reading behaviour to be studied across multiple levels of granularity.

This section clarifies the various requirements for our ET system. These include target groups, deployment settings, relevant metrics and datasets, system functionality, suitable equipment, aspects of predictive modelling, output formats and ethical issues.

5.1. Requirements of educators

We summarize the insights of focus group discussions held during the EYE-TEACH workshop on May 14th, 2025. The discussions brought together teachers from primary, secondary, and tertiary education to explore their experiences and expectations regarding reading comprehension challenges and the use of AI-supported ET systems.

Preliminary findings from these focus groups can be categorized into the following types of requirements.

Understanding student reading processes and diagnosing reading issues, including highlighting re-reading patterns, skimming behavior, and attention shifts, providing indicators of reading pace, cognitive load, and engagement, and detecting whether issues stem from decoding or comprehension;

User-friendly data presentation and visualization, including class-level dashboards with drill-down access to individual students, providing visual summaries (heatmaps, graphs) and textual interpretations appropriate to the education level and context,

***T*eacher autonomy and control**, including allowing teachers to approve or override system suggestions, providing different levels of automation, i.e., *human-in-the-loop*, and enabling teachers to configure system autonomy per student or task, and

Training guidelines and usability, including providing onboarding materials, training modules, and ongoing support, and intuitive Interfaces which maximize sense-making while minimizing teacher cognitive load.

5.2. Relevant eye-movement and readability metrics and datasets

In the study of reading behavior, eye movement metrics provide valuable insights into the cognitive processes underlying reading (Rayner et al., 1989; Rayner, Chace, et al., 2006). These metrics can be generally categorized based on the granularity of the text units they target, which in turn reflects different levels of cognitive engagement during reading (Mézière et al., 2023).

Global eye movement measures focus on broader textual units, such as sentences or paragraphs, and are particularly informative of general reading behaviors and processing. These metrics are often employed to capture phenomena such as text difficulty and inter-individual variability among readers (Rayner, Chace, et al., 2006; Rayner, Reichle, et al., 2006). Commonly used global measures include:

- **Mean fixation duration;**
- **Mean saccade length;**
- **Reading speed.**

In contrast, *local eye movement metrics* focus on finer-grained linguistic units, such as individual words (Mézière et al., 2023). These measures are particularly useful for investigating both early (e.g., lexical access) and later (e.g., syntactic parsing and semantic integration) stages of text processing (Clifton et al., 2007). Among others, the local measures include:

- **Word-frequency effect** (Shilling et al., 1998);
- **Skipping rate;**
- **First-pass fixation duration;**
- **Gaze duration;**
- **Regression rate.**

Previous research has pointed to the predictive value of global metrics such as reading speed, mean fixation duration, and saccade length in relation to reading comprehension performance (Martínez-Gómez & Aizawa, 2014; D'Mello et al., 2020; Southwell et al., 2020; Mézière et al., 2023). However, findings in the literature remain inconsistent, likely due to variations in methodologies and the nature of reading comprehension tasks (Mézière et al., 2023). This underscores the importance of adopting a nuanced approach to metric selection and interpretation, particularly when investigating individual differences or task effects in reading comprehension.

In the context of WP2 and the ongoing meta-analysis on which eye movement metrics are the most useful predictors of reading comprehension performance (D2.1), preliminary analyses were conducted and presented at *The 3rd Workshop on Eye Movements and the Assessment of Reading Comprehension*[1]. We aimed at obtaining a first evaluation of which eye movement metrics might show a significant correlation with reading comprehension outcomes already at such an early stage. We therefore collected both the eye-movement-while-reading data and the related reading comprehension scores from the major available datasets. Specifically, we used CopCo, GECO[2], Lee et al. (2024), MECO (L1)[3], Meziérè et al. (2025)[4], OneStop, PoTeC, WebQAmGaze. While we collected data from these databases, we only analyzed the measures for which we had at least three data points, according to the related WP2 meta-analysis pre-registration. To quantify the relationship between eye

[1] The workshop was organized by the University of Stuttgart (DE) and the MultiplEYE European COST Action, and took place in Stuttgart, on June 5-7, 2025.

[2] We considered each language contained in GECO as a separate dataset, resulting in 3 datasets (i.e., English native speakers, Dutch native speakers, English-Dutch bilinguals).

[3] We considered each language contained in MECO-L1 as a separate dataset, resulting in 13 datasets (i.e., Dutch, English, Estonian, Finnish, German, Greek, Hebrew, Italian, Korean, Norwegian, Russian, Spanish, and Turkish).

[4] This dataset is not publicly available and is only accessible to one author, who also ran the preliminary analyses related to the EYE-TEACH WP2. For this reason, we only included this dataset for the purpose of the preliminary analyses, but it will be excluded from the system training and development.

movement metrics and reading comprehension, we obtained the Pearson correlation coefficient (r). Whenever the Pearson correlation coefficient was not available in the datasets, we computed it starting from other available statistical values (e.g., the means of the variables of interest). To normalize effect sizes, we subsequently applied the Fisher's Z-transformation before conducting the meta-analysis and visualizing the related forest plots. Results showed that the late eye movements measures of re-reading time, number of passes, total reading time, and regression rate significantly correlate with reading comprehension outcomes. It is important to note that these analyses did not consider any moderator effect, such as any potential influence of language, reading comprehension task and text characteristics, but rather constituted a first preliminary assessment of the relation between eye movements metrics and reading comprehension based on different datasets.

Among the above-mentioned public datasets, **CopCo, GECO, Lee et al. (2024), MECO (L1 and L2), MQA-RC, OneStop, PoTeC, and WebQAmGaze are particularly well-aligned with the goals of the present project from a scientific perspective**. They provide data from multiple languages, allowing for cross-linguistic comparison. Besides, they include extended reading materials (e.g., full text passages rather than single sentences) and comprehension assessments following reading (i.e., with the text no longer available during testing), mirroring standardized comprehension tasks used in educational contexts (Catrysse et al., 2016; Mézière et al., 2023; Salmerón et al., 2024). Such designs enhance ecological validity and provide opportunities to investigate reading comprehension under conditions that closely resemble classroom or educational environments.

Although the focus is on large-scale corpora, smaller datasets may also be considered, especially when they offer high-quality data on specific populations or under-represented languages.

Text readability formulas are aimed to provide a broad estimate of a text's difficulty. Such formulas are validated by correlating them to actual data from a group of participants. Data could come from actual text comprehension performance in a set of comprehension questions (Crossley et al., 2008), reading ease as indicated in ET measures (Klein et al., 2025), or subjective estimates of text difficulty (Crossley et al., 2019).

Different readability formulas have been proposed during the last decades. Early formulas were developed in the second half of the 20th century. Those traditional formulas, such as Flesch–Kincaid, mostly focus on surface linguistic properties of texts, such as syllable count, word or sentence length. New formulas are informed by advances in natural language processing techniques (NLP) and have a stronger focus on cognitive informed indicators (Crossley et al., 2019). As there are hundreds of possible linguistic indicators, researchers tend to select in their studies a set of indicators that can be linked to cognitive processes of text comprehension, such as lexical decoding, syntactic parsing and meaning construction (Crossley et al., 2008). In validation studies, such indicators are included in a regression model to identify its weight in predicting comprehension. For example, Crossley et al. (2008) predicted text comprehension based on three indicators: lexical co-referentiality, syntactic sentence similarity, and word frequency.

Such formulas tend to outperform traditional ones in predicting text comprehension (Crossley et al., 2019) or reading ease (Klein et al., 2025). Accordingly, our system should be informed by the new cognitively informed NLP indicators.

A major limitation for a multilingual project such as ours is that most research on readability formulas has been conducted with English texts (Linders & Louwerse, 2024). Even in the case of traditional formulas, such as Flesch–Kincaid, there is no validation study in relevant languages such as Finnish. An alternative is to develop language agnostic formulas, which can be derived empirically from large datasets of easy/difficult texts (Dalal et al., 2023). But this way we depart from the cognitively informed models, which is at odds with the emphasis on cognitive informed indicators of eye-movements.

There is another alternative to deal with non-English texts. Recently, Linders and Louwerse (2024) developed Lingualyzer, an open tool that provides hundreds of NLP indicators in more than 40 languages, including all relevant for our project. The tool can be accessed at https://lingualyzer.com/

What still needs to be done is to create the formulas for each language of interest based on the NLP indicators provided by Lingualyzer. For example, we could run regression models using the Meco dataset. This way we will be able to obtain specific readability formulas for each language of the project.

Another pressing issue is to explore if text difficulty should be used to predict performance on text comprehension tests, or/and on ET measures during reading. Those associations may differ. For example, text difficulty can be strongly related to reading speed, but not to idea density (Klein et al., 2025). In turn, idea density may be

Funded by the European Union

strongly related to text comprehension. Thus, incorporating both may increase the reliability of our system.

5.3. Human-centered AI-based requirements

We will follow existing HCAI guidelines and design considerations (Ouhaichi, et al., 2024) for the development of the proposed AI-assisted ET-analytics tool, including those listed below.

Data and Model Transparency. Data and model transparency is essential to provide clear information about the data collection and processing, and the model training and deployment. Model and data documentation approaches, including *Datasheets for Datasets* (Gebru et al., 2021) and *(Interactive) Model Cards* (Mitchell et al., 2021, Crisan et al., 2022) can provide information about data collection and preprocessing, model development and evaluation, and potential risks and biases.

Automation and interactivity levels. The developed models should be able to function at different automation levels, supporting user autonomy and decision-making. For example, alerting dashboards models should be able to automatically detect events, e.g., engagement drop in classroom but should require teacher's input for decision-making. Moreover, the developed methods should allow teachers to interact at different levels with the models (Molenaar, 2022), e.g., data exploration, XAI-based interactions, and interactive visualization. Data visualizations should facilitate teacher-system conversations which can go further than the model providing an output for a given input, while accounting for teacher's cognitive load.

Explainability. The developed models and interactions should be explainable and understandable to teachers. Scenario-based XAI design methods can be deployed early in the development process to identify requirements that might arise when explainable models are deployed into complex settings of use (Wolf, 2019). We will consider the aspect of informational transparency which focuses on what information and how it should be disclosed to support teacher understanding, e.g., through model uncertainty indicators and counterfactual explanations.

Processing. The developed solutions should process and analyze data and metrics used for reading comprehension and student behavior, as well as for human-AI interaction modeling. The models should be able to process such data in real time, as well as to aggregate information for asynchronous use, considering both privacy and processing aspects. For example, *federated learning* methods enable the training of

ML models using local devices without transferring raw data. This approach reduces data leakage risks by keeping sensitive information on local devices, preserving privacy and security (Fachola et al., 2023).

Adaptability and personalization (or Hybrid Intelligence mechanisms). The model should be able to tailor its behavior and outputs to individual user needs, preferences, and contexts. The developed models should also support continuous learning from user feedback and interactions to improve over time and adapt to new tasks or environments with minimal retraining or manual intervention.

Human-centered evaluation. The developed models should be evaluated in terms of both algorithmic performance and user-centered aspects. In terms of algorithmic evaluation, metrics such as accuracy, precision, and recall should be appropriately selected, while for user-centered evaluation, metrics can involve (teacher's) task performance and completion time, cognitive load, and perceived trust in the system.

5.4. ET equipment and software requirements

Regarding ET devices, for the development of a scalable and accessible prototype, **webcam-based ET systems** present a cost-effective solution. These systems have shown promising results in detecting major reading behaviors and offer a feasible path for broader deployment in educational settings (Guan et al., 2022; Lin et al., 2022; Hutt et al., 2023; Ribeiro et al., 2023). For validation purposes instead, it is recommended to employ **high-resolution ET systems** in laboratory settings. These provide more accurate and fine-grained measures of eye movements, essential for establishing the reliability and sensitivity of the chosen metrics.

Therefore, we are considering developing two ET solutions: a simpler (low-end), widely usable webcam-based system, and a more sophisticated (high-end) system requiring more advanced equipment. For our purposes, the **remote ET systems** seem to be the most appropriate, with the sampling frequency at least 30 Hz (Angele et al., 2025). A crucial requirement is **explainability** of our solutions that should provide the output both in the graphical way (e.g. as heatmaps) and in the natural language (LLM-based).

The *EyeLink 1000 Plus* and *EyeLink Portable Duo* have high levels of accuracy and precision, making them the suitable tools for reading research. The EyeLink 1000 Plus (https://www.sr-research.com/eyelink-1000-plus/) is a highly flexible ET solution with a range of mounting options. The EyeLink Portable Duo (https://www.sr-research.com/eyelink-portable-duo/) is ideal for taking reading research out of the lab and into settings such as schools and libraries.

Webcam based ET is supported by *iMotions WebET 3.0* (https://imotions.com/products/imotions-lab/modules/eye-tracking-webcam/). The underlying algorithm has been trained using millions of images of faces under different lighting conditions and from a variety of ethnic groups, ensuring robust performance and accurate results. This software offers a reliable way to conduct ET research online, with remote data collection capabilities.

An open-source Python package called *pymovements* was designed for processing eye movement data (https://github.com/aeye-lab/pymovements). This package provides a simple interface for downloading publicly available datasets, preprocessing gaze data, detecting oculomotor events and rendering plots for the visual analysis of results.

The following decision criteria for ET equipment and software were suggested in the EYE-TEACH project:

- *Cost:* purchase price, license fees, long-term sustainability
- *Usability:* easy to use for teachers and students, low training threshold
- *Data quality:* accuracy, compatibility with needed ET metrics
- *Scalability:* works in school and home settings
- *Legal & ethical compliance:* GDPR, EU AI Act consideration
- *Control:* customizability, data pipeline control, independence from vendors
- *Integration potential:* API availability, plug-in compatibility
- *Maintenance & support:* vendor support, community activity (for open source)

5.5. Ethical and legal requirements

The design and deployment of AI-based ET systems must align with ethical and legal standards, especially in sensitive environments like education. This section addresses the main requirements drawn from the Ethics Guidelines for Trustworthy AI (2019) by the European Commission's High-Level Expert Group, the Artificial Intelligence Act (AI Act 2024), and the General Data Protection Regulation (GDPR 2016). They include explainable models, human-centric design, fairness testing, and strict data governance, starting at the dataset selection stage for AI training. In particular, the use of only anonymized and ethically compliant ET datasets ensures that the AI system respects the rights of both data subjects and future users.

Transparency and Explainability. Since the European AI Act categorizes educational AI systems using biometric data as "high-risk", the system must comply with mandatory conformity assessments, risk management procedures, and post-market monitoring. Article 13 of the AI Act (2024) requires high-risk AI systems to be intelligible, aligning with the ethical requirements of transparency and explainability set out in the Ethics Guidelines (2019).

Transparency requires that users (educators, learners, and administrators) understand how the AI functions. The system should provide clear explanations of how ET data is interpreted and how this influences system feedback or educational decisions. Explainability must be adapted to various stakeholders' literacy levels, ensuring that students understand how their data affects outcomes. Therefore, ET systems must ensure that educators, learners, and guardians understand how AI decisions are made. In particular, the system should be provided by (i) visual and textual explanations of attention scores or performance predictions, (ii) transparent documentation of data inputs, algorithmic logic, and performance metrics, (iii) access to audit logs for verification and accountability.

Human Autonomy and Oversight. AI systems must support human autonomy and decision-making, ensuring that teachers remain in control and central in interpreting AI outputs, while learners should be empowered to contest or question AI-driven feedback. From this perspective, ET technologies should function as assistive tools rather than decision-makers. Mechanisms for human oversight must be embedded to intervene when the system's behavior contradicts pedagogical goals or student well-being. Therefore, built-in mechanisms for human-in-the-loop control and override are essential.

Non-Discrimination and Fairness. The system must be free from biases in training data or algorithmic processing and related to ethnicity, gender, neurodiversity, or disability, since they can reinforce educational disparities. Key safeguards include: (i) bias testing across demographic variables (e.g., age, gender, ethnicity), (ii) design for inclusivity, ensuring equitable performance across diverse student populations, (iii) fairness audits and bias mitigation strategies must be part of the development lifecycle.

Data Privacy and Governance. Given that ET involves sensitive biometric data, strict data protection measures are required. This includes adherence to the GDPR (2016), with informed consent, purpose limitation, data minimization, and the right to access and erase personal data. Special care should be taken to anonymize data, where possible, and ensure secure storage and processing. These provisions are in accordance with Recital 44 and Articles 10 and 11 of the AI Act and GDPR's Articles 5-9.

Only fully anonymized open datasets, obtained through ethics-approved protocols should be employed in the training phase. Therefore, among the dataset analyzed in **Section 4.6.1**, EYE-TEACH project should build on those datasets, such as WebQAmGaze, CopCo, PoTeC, MECO, OneStop, ZuCo, ZuCo 2.0, CELER, Valiapan et al., IITB-HGC, InteRead, that are publicly available under research use licenses, that have been developed in compliance with institutional review boards or equivalent ethics committee, and that have included permission for further research activities compatible with the training of AI system in the participant's informed consent.

6. Conclusion

Data obtained through ET while reading constitutes a distinctive type of spatio-temporal, multimodal information. Eye movements unfold dynamically over static linguistic input in the form of text. The reading process itself is influenced by three factors: text properties, reader characteristics, and the nature of the task. In this context, predictive modelling poses a particular challenge to ML.

Evidence demonstrating the interplay between ET metrics, ML methods and reading behaviour has been found (Arnold et al., 2025). This confirms the usefulness of advanced computational approaches in improving the accuracy of ET technologies, and highlights important areas for future research. These include overcoming limitations in mobile systems. Expanding ET research to natural, uncontrolled environments is crucial for understanding real-world reading, comprehension and attention processes and for addressing the limitations of controlled settings. Including demographics beyond university students is also essential in order to capture variations in age, language and cultural background.

The challenge lies in facilitating progress in ML models that can decode rich cognitive signals from ET data during the reading process. Future improvements should be driven by robust and versatile modelling approaches that advance the alignment of eye movements with text.

To make further progress in this area, additional data collection efforts are needed. Open, well-documented, large-scale, high-quality datasets spanning diverse languages, populations, text genres and reading interactions are especially important.

To responsibly deploy AI-based ET systems in education, it is essential to meet ethical and legal standards. Compliance with the Ethics Guidelines for Trustworthy AI, the AI Act, and GDPR ensures an ethics and privacy-based approach by design of such technologies, which will allow EYE-TEACH not only to improve educational outcomes but also to respect fundamental rights and human dignity.

In this deliverable, we summarised the requirements for AI solutions that can translate eye movements into comprehensible outputs, supporting teachers in diagnosing reading difficulties and giving them greater control in the classroom and during one-to-one tutorials. We identified the most relevant eye movement metrics, including fixation duration, saccade length, reading speed, the word-frequency effect, skipping rate, first-pass fixation duration, gaze duration and regression rate, as well as suitable public datasets such as CopCo, GECO, MECO, MQA-RC, OneStop, PoTeC, SB-SAT and WebQAmGaze. We also considered both low-end and high-end solutions for our development, enabling normal usage with webcams in classroom settings, as well as evaluation with high-resolution ET systems. Finally, we specified the human-centred, AI-based requirements alongside the ethical and legal requirements, in which transparency, explainability and human autonomy play an important role.

In the next project phase, we will focus on defining our data framework and selecting appropriate AI models to process and analyse the eye movement metrics curated by WP2. Potential tasks include analysing and predicting reading behaviour patterns, mining emotions and cognition, detecting reading competencies, and providing LLM-based assistance.

7. References

Ahn, S., Kelton, C., Balasubramanian, A., & Zelinsky, G. (2020, June). Towards predicting reading comprehension from gaze behavior. In *ACM Symposium on Eye Tracking Research and Applications* (pp. 1-5).

Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 100215.

Angele, B., Gunes Ozkan, Z., Serrano-Carot, M., & Duñabeitia, J. A. (2025). How low can you go? Tracking eye movements during reading at different sampling rates. *Behavior Research Methods*, *57*(7), 1-25.

Arnold, L., Aryal, S., Hong, B., Nitharsan, M., Shah, A., Ahmed, W., ... & Piaggio, D. (2025). A Systematic Literature Review of Eye-Tracking and Machine Learning Methods for Improving Productivity and Reading Abilities. *Applied Sciences*, *15*(6), 3308.

Berzak, Y., Malmaud, J., Shubi, O., Meiri, Y., Lion, E., & Levy, R. (2025). Onestop: A 360-participant english eye-tracking dataset with different reading regimes. *PsyArXiv preprint*.

Berzak, Y., Nakamura, C., Smith, A., Weng, E., Katz, B., Flynn, S., & Levy, R. (2022). CELER: A 365-participant corpus of eye movements in L1 and L2 English reading. *Open Mind*, *6*, 41-50.

Boels, L. (2023). Reflections on gaze data in statistics education. *Teaching Statistics*, *45*, S40-S51.

Catrysse, L., Gijbels, D., Donche, V., de Maeyer, S., van den Bossche, P., & Gommers, L. (2016). Mapping processing strategies in learning from expository text: an exploratory eye tracking study followed by a cued recall. Frontline Learning Research, 4(1), 1-16. https://doi.org/10.14786/flr.v4i1.192

Chen, Q., Li, Z., Pong, T. C., & Qu, H. (2019, April). Designing narrative slideshows for learning analytics. In 2019 IEEE pacific visualization symposium (PacificVis) (pp. 237-246). IEEE.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. G. van Gompel (Ed.), *Eye movements: A window on mind and brain* (pp. 341–374). Elsevier. https://doi.org/10.1016/B978-008044980-7/50017-3

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, *49*, 602-615.

Crisan, A., Drouhard, M., Vig, J., & Rajani, N. (2022, June). Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 427-439).

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. Tesol Quarterly, 42(3), 475-493.

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. Journal of Research in Reading, 42(3-4), 541-561.

Dalal, M., Gautam, S., Gulalkari, V., Joshi, S., Seetha, V., Tom, A., & Lanham, M. A. (2023, July). Language Agnostic Readability Assessments. In 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE) (pp. 1532-1536). IEEE.

Dass, S., Ramananda, H. S., & Savio, A. A. (2023). Eye-Tracking in Education: Analysing the relationship between Student's performance and Videonystagmography Report. In *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)* (Vol. 1, pp. 1-6). IEEE.

Dostálová, N., & Plch, L. (2023). A Scoping Review of Webcam Eye Tracking in Learning and Education. *Studia paedagogica*, *28*(3), 113-131.

Evdokimov, A., Finegan-Dollak, C., & Robbins, A. (2024, June). WEyeDS: A desktop webcam dataset for gaze estimation. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications* (pp. 1-2).

Fachola, C., Tornaría, A., Bermolen, P., Capdehourat, G., Etcheverry, L., & Fariello, M. I. (2023). Federated learning for data analytics in education. *Data*, *8*(2), 43.

Fikri, M. A., Santosa, P. I., & Wibirama, S. (2021). A review on opportunities and challenges of machine learning and deep learning for eye movements classification.

In *2021 IEEE International Biomedical Instrumentation and Technology Conference (IBITeC)* (pp. 65-70). IEEE.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Guan, X., Lei, C., Huang, Y., Chen, Y., Du, H., Zhang, S., & Feng, X. (2022). An analysis of reading process based on real-time eye-tracking data with web-camera — Focus on English reading at higher education level. In *Proceedings of the 4th Workshop on Predicting Performance Based on the Analysis of Reading Behavior*.

Hagihara, H., Zaadnoordijk, L., Cusack, R., Kimura, N., & Tsuji, S. (2024). Exploration of factors affecting webcam-based automated gaze coding. *Behavior Research Methods*, *56*(7), 7374-7390.

Hollenstein, N., Barrett, M. J., & Björnsdóttir, M. (2022). The Copenhagen Corpus of Eye Tracking Recordings from Natural Reading of Danish Texts. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)* (pp. 1712-1720).

Hollenstein, N., & Beinborn, L. (2021). Relative Importance in Sentence Processing. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021* (pp. 141-150). Association for Computational Linguistics (ACL).

Hollenstein, N., Chersoni, E., Jacobs, C. L., Oseki, Y., Prévot, L., & Santus, E. (2021, June). CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 72-78).

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data*, *5*(1), 1-13.

Hollenstein, N., Troendle, M., Zhang, C., & Langer, N. (2020). ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 138-146). European Language Resources Association.

Hoque, M. N., Shin, S., & Elmqvist, N. (2024). Visualization for Human-Centered AI Tools. *arXiv preprint arXiv:2404.02147*.

Hosp, B., Eivazi, S., Maurer, M., Fuhl, W., Geisler, D., & Kasneci, E. (2020). RemoteEye: An open-source high-speed remote eye tracker: Implementation insights of a pupil- and glint-detection algorithm for high-speed remote eye tracking. *Behavior research methods*, *52*, 1387-1401.

Husain, S., Vasishth, S., & Srinivasan, N. (2015). Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, *8*(2).

Hutt, S., & D'Mello, S. K. (2022). Evaluating calibration-free webcam-based eye tracking for gaze-based user modeling. In *Proceedings of the 2022 international conference on multimodal interaction* (pp. 224-235).

Hutt, S., Wong, A., Papoutsaki, A., Baker, R. S., Gold, J. I., & Mills, C. (2023). Webcam-based eye tracking to detect mind wandering and comprehension errors. *Behavior Research Methods*, 1–17. https://doi.org/10.3758/s13428-023-02261-1

Jain, T., Bhatia, S., Sarkar, C., Jain, P., & Jain, N. K. (2024, June). Real-Time Webcam-Based Eye Tracking for Gaze Estimation: Applications and Innovations. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.

Jaiswal, A., Venkatesh, J., & Nanda, G. (2023). Understanding student engagement during an experiential learning task using eye tracking: A case study. In *2023 IEEE Frontiers in Education Conference (FIE)* (pp. 1-7). IEEE.

Jakobi, D. N., Kern, T., Reich, D. R., Haller, P., & Jäger, L. A. (2024). PoTeC: A German naturalistic eye-tracking-while-reading corpus. *arXiv preprint arXiv:2403.00506*.

Jamil, N., & Belkacem, A. N. (2024). Advancing real-time remote learning: A novel paradigm for cognitive enhancement using EEG and eye-tracking analytics. *IEEE Access*.

Kar, A. (2020). MLGaze: Machine learning-based analysis of gaze error patterns in consumer eye tracking systems. *Vision*, *4*(2), 25.

Kędras, M., & Sobecki, J. (2023). What is hidden in clear sight and how to find it—A survey of the integration of Artificial Intelligence and eye tracking. *Information*, *14*(11), 624.

Kennedy, A., Pynte, J., Murray, W. S., & Paul, S. A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology*, *66*(3), 601-618.

Klein, K. G., Frenkel, S., Shubi, O., & Berzak, Y. (2025). Surprisal Takes It All: Eye Tracking Based Cognitive Evaluation of Text Readability Measures. arXiv preprint arXiv:2502.11150.

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., & Kliegl, R. (2019). Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior research methods*, *51*, 1161-1178.

Lee C.E., Godwin H.J., & Drieghe D. (2024) The jingle fallacy in comprehension tests for reading. PLoS ONE 19(7): e0306466. https://doi.org/10.1371/journal.pone.0306466

Ley, T., Tammets, K., Pishtari, G., Chejara, P., Kasepalu, R., Khalil, M., ... & Wasson, B. (2023). Towards a partnership of teachers and intelligent learning technology: A systematic literature review of model-based learning analytics. *Journal of Computer Assisted Learning*, *39*(5), 1397-1417.

Lin, Z., Liu, Y., Wang, H., Liu, Z., Cai, S., Zheng, Z., Zhou, Y., & Zhang, X. (2022). An eye tracker based on webcam and its preliminary application evaluation in Chinese reading tests. *Biomedical Signal Processing and Control, 74*, 103521. https://doi.org/10.1016/j.bspc.2022.103521

Linders, G. M., & Louwerse, M. M. (2024). Lingualyzer: A computational linguistic tool for multilingual and multidimensional text analysis. Behavior research methods, 56(6), 5501-5528.

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, *50*, 826-833.

Maharaj, K., Saxena, A., Kumar, R., Mishra, A., & Bhattacharyya, P. (2023, December). Eyes show the way: Modelling gaze behaviour for hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 11424-11438).

Makowski, S., Jäger, L. A., Abdelwahab, A., Landwehr, N., & Scheffer, T. (2019). A discriminative model for identifying readers and assessing text comprehension from eye movements. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18* (pp. 209-225). Springer International Publishing.

Mézière, D. C., Yu, L., Reichle, E. D., Von Der Malsburg, T., & McArthur, G. (2023). Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly, 58*(3), 425–449. https://doi.org/10.1002/rrq.468

Mézière, D. C., Yu, L., von der Malsburg, T., Reichle, E. D., & McArthur, G. (2025). Using Eye Movements From a "Read-Only" Task to Predict Text Comprehension. *Reading Research Quarterly*, *60*(3), e70023.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. European Journal of Education, 57(4), 632-645.

Özer, E., & Özdemir, S. (2021). The relation between reading performance and eye movement parameters of high-skilled and low-skilled readers. *Egitim ve Bilim*, *46*(208), 395-412.

Ouhaichi, H., Bahtijar, V., & Spikol, D. (2024, July). Exploring design considerations for multimodal learning analytics systems: an interview study. In *Frontiers in Education* (Vol. 9, p. 1356537). Frontiers Media SA.

Pozdniakov, S., Martinez-Maldonado, R., Tsai, Y. S., Echeverria, V., Swiecki, Z., & Gasevic, D. (2025). Investigating the Effect of Visualization Literacy and Guidance on Teachers' Dashboard Interpretation. *Journal of Learning Analytics*, 1-25.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10*(3), 241–255. https://doi.org/10.1207/s1532799xssr1003_3

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging, 21*(3), 448–465. https://doi.org/10.1037/0882-7974.21.3.448

Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., & Clifton, C. Jr. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes, 4*(SI), SI21–SI49. https://doi.org/10.1080/01690968908406362

Ribeiro, T., Brandl, S., Søgaard, A., & Hollenstein, N. (2023). WebQAmGaze: A multilingual webcam eye-tracking-while-reading dataset. *arXiv*. https://doi.org/10.48550/arXiv.2303.17876

Salmerón, L., Altamura, L., Delgado, P., Karagiorgi, A., & Vargas, C. (2024). Reading comprehension on handheld devices versus on paper: A narrative review and meta-analysis of the medium effect and its moderators. *Journal of Educational Psychology, 116*(2), 153–170. https://doi.org/10.1037/edu0000829

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition, 26*(6), 1270–1281. https://doi.org/10.3758/BF03201199

Sedmidubsky, J., Dostalova, N., Svaricek, R., & Culemann, W. (2024, October). ETDD70: Eye-Tracking Dataset for Classification of Dyslexia Using AI-Based Methods. In *International Conference on Similarity Search and Applications* (pp. 34-48). Cham: Springer Nature Switzerland.

Shubi, O., Meiri, Y., Hadar, C., & Berzak, Y. (2024, November). Fine-Grained Prediction of Reading Comprehension from Eye Movements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 3372-3391).

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H. D., Alexeeva, S., Amenta, S., ... & Kuperman, V. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior research methods*, *54*(6), 2843-2863.

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N. S., ... & Tsiknakis, M. (2021). Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*, *16*, 260-277.

Soh, O. K. (2016). Examining the reading behaviours and performances of sixth-graders for reading instruction: evidence from eye movements. *Journal of e-Learning and Knowledge Society*, *12*(4).

Šola, H. M., Qureshi, F. H., & Khawaja, S. (2024). AI Eye-Tracking Technology: A New Era in Managing Cognitive Loads for Online Learners. *Education Sciences*, *14*(9), 933.

Sood, E., Tannert, S., Frassinelli, D., Bulling, A., & Vu, N. T. (2020, November). Interpreting Attention Models with Human Visual Attention in Machine Reading

Comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 12-25).

Špakov, O., Siirtola, H., Istance, H., & Kari-Jouko, R. (2017). Visualizing the reading activity of people learning to read. *Journal of Eye Movement Research*, *10*(5), 10-16910.

Topali, P., Ortega-Arranz, A., Rodríguez-Triana, M. J., Er, E., Khalil, M., & Akçapınar, G. (2025). Designing human-centered learning analytics and artificial intelligence in education solutions: a systematic literature review. *Behaviour & Information Technology*, *44*(5), 1071-1098.

Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., ... & Navalpakkam, V. (2020). Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature communications*, *11*(1), 4553.

Van Leeuwen, A. (2019). Teachers' perceptions of the usability of learning analytics reports in a flipped university course: When and how does information become actionable knowledge?. *Educational Technology Research and Development*, *67*, 1043-1064.

Wolf, C. T. (2019, March). Explainability scenarios: towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 252-257).

Wong, A. Y., Bryck, R. L., Baker, R. S., Hutt, S., & Mills, C. (2023). Using a webcam based eye-tracker to understand students' thought patterns and reading behaviors in neurodivergent classrooms. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 453-463).

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, *2*, 100008.

Yao, N., Brewer, J., D'Angelo, S., Horn, M., & Gergle, D. (2018, April). Visualizing gaze information from multiple students to support remote instruction. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems* (pp. 1-6).

Yüksel, D. (2023). Investigation of web-based eye-tracking system performance under different lighting conditions for neuromarketing. *Journal of Theoretical and Applied Electronic Commerce Research*, *18*(4), 2092-2106.

Zdarsky, N., Treue, S., & Esghaei, M. (2021). A deep learning-based approach to video-based eye tracking for human psychophysics. *Frontiers in human neuroscience*, *15*, 685830.

Zermiani, F., Dhar, P., Sood, E., Kögel, F., Bulling, A., & Wirzberger, M. (2024, May). InteRead: an eye tracking dataset of interrupted reading. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 9154-9169).

Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., & Hilliges, O. (2020). ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16* (pp. 365-381). Springer International Publishing.