Decoding Reading Goals from Eye Movements

Omer Shubi*, **Cfir Avraham Hadar*** , **Yevgeni Berzak**
Faculty of Data and Decision Sciences,
Technion - Israel Institute of Technology, Haifa, Israel
{shubi,kfir-hadar}@campus.technion.ac.il, berzak@technion.ac.il

Abstract

Readers can have different goals with respect to the text that they are reading. Can these goals be decoded from their eye movements over the text? In this work, we examine for the first time whether it is possible to distinguish between two types of common reading goals: *information seeking* and *ordinary reading* for comprehension. Using large-scale eye tracking data, we address this task with a wide range of models that cover different architectural and data representation strategies, and further introduce a new model ensemble. We find that transformer-based models with scanpath representations coupled with language modeling solve it most successfully, and that accurate predictions can be made in *real time*, shortly after the participant started reading the text. We further introduce a new method for model performance analysis based on mixed effect modeling. Combining this method with rich textual annotations reveals key properties of textual items and participants that contribute to the difficulty of the task, and improves our understanding of the variability in eye movement patterns across the two reading regimes.[1]

1 Introduction

Reading is a ubiquitously practiced skill that is indispensable for successful participation in modern society. When reading, our eyes move over the text in a sequence of *fixations*, during which the gaze is stable, and rapid transitions between fixations called *saccades*. This sequence contains rich information about language comprehension in real time and the nature of the reader's interaction with the text (Rayner, 1998; Hyönä and Kaakinen, 2019; Schotter and Dillon, 2025).

In daily life, a reader may have one or several *goals* that they pursue with respect to the text. For example, they may read the text closely or skim it to obtain the gist of the text's content, they may proofread it, or they may be seeking specific information of interest. Each such goal can impact online linguistic processing and the corresponding eye movement behavior while reading. Despite the many reading goals readers pursue in everyday life, research on eye movements in cognitive science, as well as work that integrated eye movements data in NLP and machine learning have primarily focused on one reading regime, which can be referred to as *ordinary reading*. In this regime, the reader's goal is general comprehension of the text. Although widely acknowledged, other forms of reading received much less attention and remain understudied (Radach and Kennedy, 2004).

In this work, we go beyond ordinary reading and ask whether broad reading goals can be reliably decoded from the pattern of the reader's eye movements over the text. We focus on the distinction between ordinary reading and *information seeking*, a highly common reading regime in everyday life, where the reader is interested in obtaining specific information from the text. Decoding reading goals from eye movements has practical implications in several areas. In education, it can enable real-time monitoring of students' engagement, facilitating targeted interventions to support effective reading and information-seeking strategies. For user-centric applications, it allows dynamic content adaptation, such as highlighting relevant information when users are seeking specific details. In assistive technologies, it can provide real-time support for special populations, such as helping elderly users navigate complex websites by identifying and addressing their information-seeking needs.

Prior work suggests that on average across participants and texts, there are substantial differences in eye movement patterns between information seeking and ordinary reading (Hahn and Keller, 2023; Malmaud et al., 2020; Shubi and Berzak, 2023).

*Equal contribution.

[1]Code is available at: https://github.com/lacclab/Decoding-Reading-Goals-from-Eye-Movements.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5616–5637
July 27 - August 1, 2025

However, it is currently unknown whether there is sufficient signal in the eye movement record for automatic decoding of the reading goal given eye movements of a single participant over a single textual item. Furthermore, little is known about the factors that contribute to the difficulty of this task.

In this work, we address this gap by conducting a series of experiments on reading goal decoding. Our main contributions are the following:

- **Task**: We introduce a new decoding task: given eye movements from a single participant over a passage, predict whether they engaged in ordinary reading for comprehension or in information seeking.
- **Modeling**: We adapt and apply to this task 12 different state-of-the-art predictive models for eye movements in reading. We further introduce an ensemble model which leverages the diversity of predictions from single models.
- **Evaluation** We systematically characterize the generalization ability of the models across new textual items and new participants. We find that the models that perform best are transformer architectures that use scanpath sequence representations as well as the text. We further demonstrate that it is feasible to perform the task *online* and make accurate predictions long before the participant finished reading the text.
- **Error Analysis**: We introduce mixed effects modeling of model logits as a general method for analyzing model performance as a function of different properties of the data. Differently from univariate error analyses common in the literature, this method allows examining each feature of interest while controlling for all other features, and taking into account item and subject dependencies in the data. Combining this method with rich data annotations reveals key interpretable axes of variation that contribute to task difficulty, and provides new insights on the data itself.

2 Task

We address the task of predicting whether a reader is engaged in ordinary reading for comprehension or in seeking specific information, based on their eye movements over the text. Let S be a participant, P a textual passage, and E_P^S the recording of the participant's eye movements over the passage. Given a ground-truth mapping $\mathcal{C}(S, P) \rightarrow \{\text{Information Seeking}, \text{Ordinary Reading}\}$, we aim to approximate $\mathcal{C}$ with a classifier h:

$$h : (E_S^P, P) \rightarrow \begin{cases} \text{Information Seeking} \\ \text{Ordinary Reading} \end{cases}$$

Where the passage P is an optional input, such that the classifier can be provided only with the eye movement data E_S^P or with both the eye movements and the underlying text. We assume that the participant has not read the paragraph previously.

The information seeking regime is a general framework for addressing goal based reading. It is operationalized by presenting the participant with an arbitrary question Q prior to reading the passage. This question prompts the participant to seek specific information in the text. We assume that the classifier does not receive the question nor any information on the participant, which makes the task relevant for real-world scenarios where users are anonymous and no information is available about their specific information seeking goal. Figure 1 presents the task schematically.

3 Modeling

Eye movements during reading consist of fixations and saccades (Rayner, 1998; Hyönä and Kaakinen, 2019; Schotter and Dillon, 2025), and present a highly challenging case of temporally *and* spatially aligned multimodal data, where fixations are both temporal and correspond to specific words in the text. Recently, a number of general purpose predictive models for eye movements have emerged, each typically evaluated on a different task. Here, we adjust and deploy them for a single task, which allows a systematic comparison of architectural and data representation strategies. The models can be broadly divided along three primary axes, the modalities used (eye movements-only, or eye movements and text), how eye movement information is represented (global feature averages across the text, single word, single fixation or an image of the fixation sequence), and for the multimodal approaches, the nature of the text representations and strategy for combining them with eye movements.

3.1 Eye Movements-only Models

These models use only eye movement information, without taking into account the text. Such models

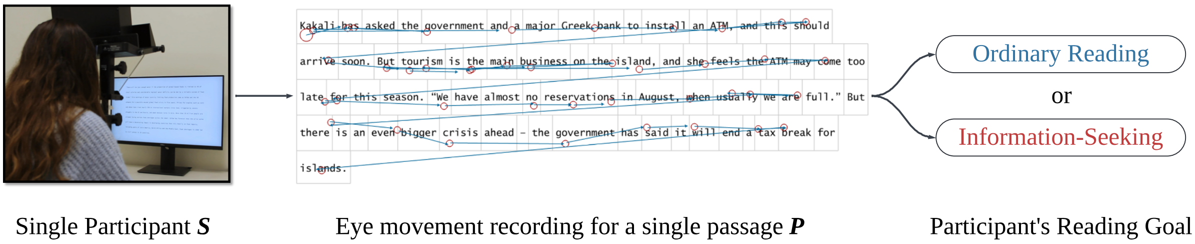


Figure 1: Proposed task: decoding whether a reader is seeking specific information or reading for general comprehension, given their eye movements over a single passage. In the eye movements image, circles represent fixations, and lines represent saccades. Bounding boxes mark word Interest Areas (fixations within the box are assigned to the respective word).

are valuable in common scenarios where the underlying text for the eye movement recording is not available. It is also the go-to approach when the eye-tracking calibration is of low quality, leading to imprecise information on the location of fixations with respect to the text. This is a highly common situation, especially with web-based eye-tracking and lower grade eye-tracking devices. Beyond practical considerations, the eye movements-only approach allows assessing the added value of textual information for our task. The models include:

- **Logistic Regression** with 9 global eye movement features capturing average fixation and saccade metrics.
- **BEyeLSTM - No Text**, similar to BEyeLSTM (Reich et al., 2022a) (see below), but without the text features.
- **Vision Models** Following the approach of Bhattacharya et al. (2020), we use two vision models, **ViT** (Dosovitskiy et al., 2021) and **ConvNext v2** (Woo et al., 2023), that represent the scanpath as an image without the underlying text, where fixations are depicted as circles with a diameter proportional to the fixation duration. See examples of input images in Figure 4 in Appendix A.1.

3.2 Eye Movements and Text Models

We further adjust a number of recent multimodal models that combine eye movements with textual information. The models encode textual information in two ways. The first is using contextual word embedding representations commonly used in NLP. The second is via linguistic word property features, including word length, word frequency and surprisal (Hale, 2001; Levy, 2008), which are motivated by their ubiquitous effects on reading times (Rayner et al., 2004; Kliegl et al., 2004; Rayner et al., 2011, among others).

The models implement three primary strategies for combining the two modalities at progressively later stages of processing: (i) in the model input, (ii) merging them within intermediate model representations, or (iii) with architectures that fuse the modalities using cross attention mechanisms after each modality has been processed separately. Furthermore, since eye movements in reading are both temporally and spatially aligned with the underlying text, the models can be categorized based on how they capture this alignment: (i) by aggregating eye movements for each *word*, thereby focusing on spatial alignment; or (ii) by aggregating eye movement information for each individual *fixation*, which explicitly encodes both spatial and temporal correspondences between eye movements and text. We use the following models:

- **RoBERTEye-W** (Shubi et al., 2024): A multimodal transformer model that combines word embeddings with word-level eye movement features at the input layer.
- **RoBERTEye-F** (Shubi et al., 2024): similar to the above, but with fixation-level representations.
- **MAG-Eye** (Shubi et al., 2024): Injects word-level eye movement features into intermediate transformer representations.
- **PLM-AS** (Yang and Hollenstein, 2023): Reorders word embeddings based on fixation sequences and processes them with an RNN.
- **Haller RNN** (Haller et al., 2022): Processes fixation-ordered word embeddings with concatenated eye movement features via an RNN.
- **BEyeLSTM** (Reich et al., 2022a): Combines fixation sequences and global features with an LSTM and a linear projection layer.

- **Eyettention** (Deng et al., 2023): Aligns word and fixation sequences using cross-attention between a RoBERTa encoder and an LSTM fixation encoder.

- **PostFusion-Eye** (Shubi et al., 2024): Combines RoBERTa word representations and convolution-based fixation features using cross-attention and shared latent projection.

In Appendix A.1 we provide additional details about each of the models, and Figures 5 and 6 in Appendix A.2 we include the model diagrams.

3.3 Logistic Ensemble

As shown in Figure 9 in Appendix F, the examined models turn out to have diverse predictive behaviors. We therefore introduce a **Logistic Ensemble**: a 12-feature logistic regression model that predicts the reading goal from the probability outputs of our 12 models.

3.4 Baselines

When examining the utility of eye movements for a prediction task, it is important to benchmark models against simpler approaches that do not require eye movement information (Shubi et al., 2024). We therefore introduce the following two baselines:

- **Majority Class** Assigns the label of the majority class in the training set to all the trials in the test set. Since our data is balanced (see below), this baseline is equivalent to random guessing.

- **Reading Time (per word)** Total reading time per word, computed by dividing the participant's total reading time of the paragraph by the number of words in the paragraph. This behavioral baseline does not require eye-tracking and is motivated by the analyses of Hahn and Keller (2023), Malmaud et al. (2020) and Shubi and Berzak (2023), which indicate that on average, reading is faster in information seeking compared to ordinary reading.

4 Experimental Setup

4.1 Data

Addressing the proposed task is made possible by OneStop Eye Movements (Berzak et al., 2025), the first dataset that contains broad coverage eye-tracking data in both ordinary reading and information seeking regimes. The textual materials of OneStop are taken from OneStopQA (Berzak et al., 2020), a multiple-choice reading comprehension dataset that comprises 30 Guardian articles from the OneStopEnglish corpus (Vajjala and Lučić, 2018). Each article is available in the original (Advanced) and simplified (Elementary) versions. Each paragraph has three multiple choice reading comprehension questions that can be answered based on any of the two paragraph difficulty level versions. Each question is paired with a manually annotated textual span, called the *critical span*, which contains the vital information for answering the question. An example of a OneStopQA paragraph along with one question and its critical span annotation is provided in Table 3 in Appendix B.

Eye movements data for OneStopQA were collected in-lab from 360 adult native English speakers using an EyeLink 1000 Plus eye tracker. Each participant read a batch of 10 articles (54 paragraphs) paragraph by paragraph. The experiment has two between-subjects reading goal tasks: information seeking and ordinary reading. In the information seeking task, participants were presented with the question (without the answers) prior to reading the paragraph. In the ordinary reading task, they did not receive the question prior to reading the paragraph. In both tasks, after having read the paragraph, participants proceeded to answer the question on a new screen, without the ability to return to the paragraph. Each paragraph was read by 120 participants: 60 in ordinary reading and 60 in information seeking (split equally between the Advanced and Elementary versions of the paragraph).

Overall, the data consists of 19,438 trials, where a trial is a recording of eye movements from a single participant over a single paragraph. The data is balanced, with 9,718 trials in ordinary reading and 9,720 in information seeking. Figure 7 in Appendix B shows example trials for both reading regimes. Additional data statistics are described in Appendix B.

4.2 Model Training and Evaluation Protocol

We use 10-fold cross validation, addressing three levels of model generalization:

- **New Textual Item**: prior eye tracking data is available for the participant but not for the paragraph.

- **New Participant**: prior eye tracking data is available for the paragraph, but not for the participant.

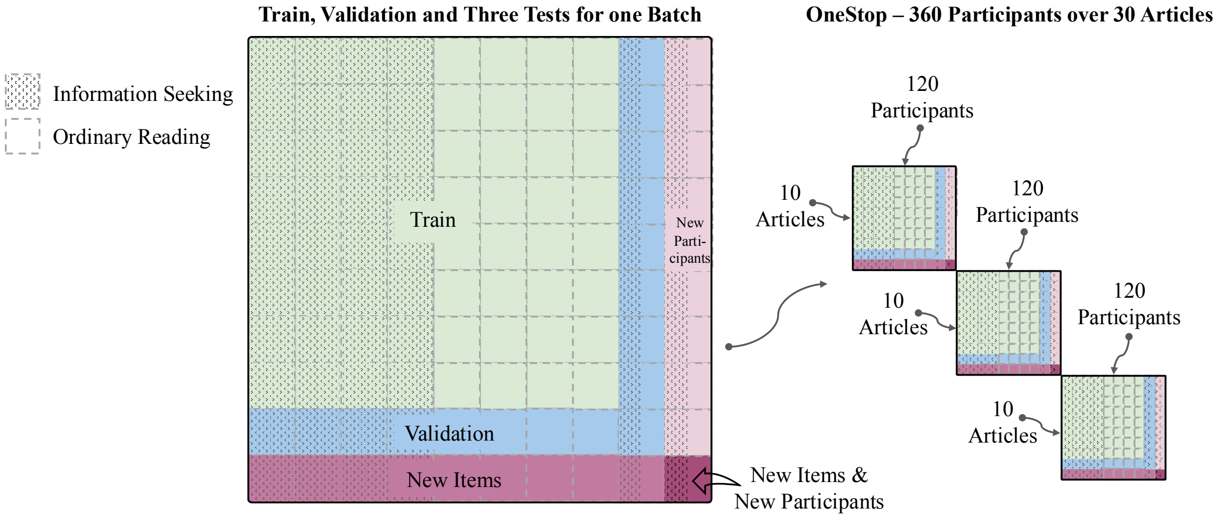


Figure 2: A schematic depiction of one of the 10 splits into train, validation, and the three test sets for one batch of 10 OneStopQA articles and 120 participants (left) and for all three batches (right). Dashed lines denote information seeking trials. The full data split consists of the union of three such batches.

- **New Item & Participant**: No prior training data for the participant nor the paragraph.
- **All**: aggregated results for the three regimes.

The New Item regime evaluates performance on unseen paragraphs using eye movement data from other texts, which is a relevant scenario for applications such as e-learning. The New Participant regime tests predictions for unseen individuals on passages for which data from other participants was already collected, reflecting scenarios such as exams where behavioral data for the given materials exist, but not from the tested participant. The New Item & New Participant evaluation, which addresses zero-shot prediction for an arbitrary unseen reader on an arbitrary unseen passage, is the most challenging and flexible regime.

The data is split into train, validation, and the three test sets separately for each batch of 10 articles and the 120 participants who read the batch. The three batch splits are then combined to form the full split of the dataset. Paragraphs are allocated to the train, validation, and test portions of each batch split at the *article level*, such that all the paragraphs of each article appear in the same portion of the split. This ensures that items in the test set are unrelated in content to items in training and validation.

Each data split contains 64% of the trials in the training set, 17% in the validation set and 19% in the test sets (9% New Item, 9% New Participant and 1% New Item & Participant). Aggregated across the 10 splits, 90% of the trials in the dataset appear in each of the New Participant and New Item evaluation regimes, and 10% in the New Item & Participant regime. Figure 2 presents this breakdown for one batch split.

Model Hyperparameters We perform hyperparameter optimization and model selection separately for each split. We assume that at test time, the evaluation regime of the trial is *unknown*. Model selection is therefore based on the entire validation set of the split. Further details regarding the training procedure, including the full hyperparameter search space for all the models are provided in Appendix C.

Statistical Testing The samples in the OneStop dataset are not i.i.d; each item is read by multiple participants, and each participant reads multiple items. To account for these dependencies when comparing model performance, we fit linear mixed-effects models with maximal random effects for items and participants (Barr et al., 2013) using the MixedModels package in Julia (Bates et al., 2024).

5 Results

Test set accuracy results are presented in Table 1. In line with prior observations of faster reading in information seeking compared to ordinary reading (Hahn and Keller, 2023; Malmaud et al., 2020; Shubi and Berzak, 2023), the Reading Time baseline yields above chance accuracies ($p < 0.01$ in all regimes), thus providing a strong benchmark for the evaluation of eye tracking-based models. Among the 12 examined models, RoBERTEye-F

Model	Gaze Representation	Text Representation	New Item	New Participant	New Item & Participant	All
Majority Class / Chance	–	–	$50.0_{\pm 0.0}{}_{+++}$	$50.0_{\pm 0.0}{}_{+++}$	$50.0_{\pm 0.0}{}_{+++}$	$50.0_{\pm 0.0}{}_{+++}$
Reading Time	–	–	$59.0_{\pm 0.4}{}_{+++}$	$58.9_{\pm 1.0}{}_{+++}$	$60.4_{\pm 1.2}{}_{+++}$	$59.0_{\pm 0.5}{}_{+++}$
Log. Regression	Global	–	$62.4_{\pm 0.3}{}^{**}_{+++}$	$60.6_{\pm 1.4}{}_{+++}$	$60.8_{\pm 1.6}{}_{+++}$	$61.5_{\pm 0.8}{}^{*}_{+++}$
BEyeLSTM (Reich et al., 2022a) No Text	Fixations	–	$71.5_{\pm 0.6}{}^{***}_{+++}$	$61.0_{\pm 1.1}{}_{+++}$	$61.5_{\pm 1.5}{}_{+++}$	$65.9_{\pm 0.4}{}^{***}_{+++}$
ConvNext v2	Scanpath Image	–	$70.4_{\pm 0.5}{}^{***}_{+++}$	$63.7_{\pm 0.8}{}^{**}_{+++}$	$64.0_{\pm 0.7}{}_{+++}$	$66.9_{\pm 0.3}{}^{***}_{+++}$
ViT	Scanpath Image	–	$70.6_{\pm 0.5}{}^{***}_{+++}$	$64.4_{\pm 0.8}{}^{***}_{+++}$	$64.4_{\pm 1.5}{}^{*}_{+++}$	$67.3_{\pm 0.4}{}^{***}_{+++}$
RoBERTEye-W (Shubi et al., 2024)	Words	Emb+LF	$64.6_{\pm 0.7}{}^{***}_{+++}$	$62.5_{\pm 1.3}{}^{*}_{+++}$	$62.0_{\pm 1.3}{}_{+++}$	$63.5_{\pm 0.9}{}^{**}_{+++}$
MAG-Eye (Shubi et al., 2024)	Words	Emb+LF	$52.1_{\pm 0.3}{}_{+++}$	$52.3_{\pm 0.4}{}_{+++}$	$51.5_{\pm 0.4}{}_{+++}$	$52.1_{\pm 0.2}{}_{+++}$
PLM-AS (Yang and Hollenstein, 2023)	Fixations Order	Emb	$58.6_{\pm 0.4}{}_{+++}$	$59.5_{\pm 0.5}{}_{+++}$	$57.5_{\pm 0.9}{}_{+++}$	$59.0_{\pm 0.4}{}_{+++}$
Haller RNN (Haller et al., 2022)	Fixations	Emb	$61.7_{\pm 0.6}{}^{*}_{+++}$	$61.2_{\pm 1.1}{}_{+++}$	$60.8_{\pm 1.5}{}_{+++}$	$61.3_{\pm 0.5}{}_{+++}$
BEyeLSTM (Reich et al., 2022a)	Fixations	LF	$71.4_{\pm 0.9}{}^{***}_{+++}$	$61.6_{\pm 1.1}{}_{+++}$	$62.2_{\pm 1.3}{}_{+++}$	$66.2_{\pm 0.7}{}^{***}_{+++}$
Eyettention (Deng et al., 2023)	Fixations	Emb+LF	$55.8_{\pm 0.8}{}_{+++}$	$55.7_{\pm 1.1}{}_{+++}$	$55.4_{\pm 1.8}{}_{+++}$	$55.8_{\pm 0.9}{}_{+++}$
PostFusion-Eye (Shubi et al., 2024)	Fixations	Emb+LF	$88.5_{\pm 0.7}{}^{***}_{+++}$	$90.3_{\pm 0.6}{}^{***}$	$86.0_{\pm 1.1}{}^{***}_{+}$	$89.3_{\pm 0.4}{}^{***}_{+++}$
RoBERTEye-F (Shubi et al., 2024)	Fixations	Emb+LF	$\mathbf{89.9}_{\pm 0.6}{}^{***}$	$\mathbf{90.9}_{\pm 0.4}{}^{***}$	$\mathbf{88.2}_{\pm 0.8}{}^{***}$	$\mathbf{90.3}_{\pm 0.3}{}^{***}$
Logistic Ensemble			$91.3_{\pm 1.7}{}^{***}_{\&\&\&}$	$91.6_{\pm 1.6}{}^{***}_{\&}$	$88.0_{\pm 3.1}{}^{***}$	$91.3_{\pm 1.2}{}^{***}_{\&\&\&}$

Table 1: Test accuracy results aggregated across 10 cross-validation splits, with 95% confidence intervals. 'Emb' stands for word embeddings and 'LF' for linguistic word features such as word length, frequency and surprisal. Model performance is compared to the Reading Time baseline using a linear mixed effects model. In R notation: $is\_correct \sim model + (model \mid participant) + (model \mid paragraph)$. Significant gains over this baseline are marked with '*' $p < 0.05$, '**' $p < 0.01$ and '***' $p < 0.001$ in superscript, and significant drops compared to the best model are marked in subscript with '+'. The best performing single model is marked in bold. Significant improvements of the Logistic Ensemble over this model are marked with the subscript '&'.

achieves the highest accuracy in all the evaluation regimes. PostFusion-Eye comes second, well ahead of the remaining 10 models. The top performing models suggest that a combination of three elements is key for our task: a transformer-based architecture, fixation-level encoding of eye movements, and explicit modeling of the text.

The 10 weaker models tend to perform better on the New Item regime compared to the New Participant regime. Due to the between-subjects design, where all the training and test examples for a given participant have the same label, this could reflect, at least in part, an ability of models to learn participant-specific reading behavior without explicit information on the participant (i.e. identify the participant), which is not directly pertinent to the task at hand. The current experimental setup makes it challenging to adjudicate between these two possibilities. In either case, it is highly non-trivial that models are able to generalize from prior participant data to new items.

Finally, we find that the Logistic Ensemble improves over the accuracy of the best performing single model RoBERTEye-F in all the regimes, with statistically significant improvements in all but the New Item and Participant evaluation. These performance improvements suggest that the information encoded by the different models is to some extent complementary, and that they likely capture different aspects of the eye movement data and the task. Further evidence for that can be obtained by examining the agreement between the models. Figure 9 in Appendix F depicts the pairwise Cohen's Kappa (Cohen, 1960) agreement rates across models in the validation data, where we observe mostly moderate agreement rates.

In Appendix D Figure 8, we present the Receiver Operating Characteristic (ROC) curves across the ten cross-validation splits and their corresponding Area Under the ROC Curve (AUROC) scores (Bradley, 1997). Validation set accuracies are reported in Table 4. The outcomes of these evaluations are consistent with the test results in Table 1.

6 How Quickly Can We Make Accurate Predictions?

Thus far we examined predictions from a complete recording of eye movements for a paragraph. Can we make accurate predictions before the participant finishes reading the paragraph?

Table 2 presents RoBERTEye-F All accuracy given the first 1%, 5%, 10%, 25% and 50% of the fixation sequence. While as can be expected, data quantity does impact performance, relatively high accuracy predictions can be obtained even with only the initial 5% of the fixations, which on av-

First % of Fixations	1%	5%	10%	25%	50%	100%
Average Time (sec)	0.5	1.5	2.7	6.3	12.4	24.3
Accuracy (All)	61.0$_{\pm 3.6}$	77.6$_{\pm 0.3}$	78.9$_{\pm 0.4}$	82.3$_{\pm 2.0}$	84.9$_{\pm 2.4}$	90.3$_{\pm 0.3}$

Table 2: RoBERTEye-Fixations accuracy with 95% confidence intervals as a function of the % of scanpath data used from the beginning of the paragraph reading.

erage corresponds to the first 1.5 seconds of the eye movements recording. This is an important outcome, which demonstrates the feasibility of performing our task successfully *online*, long before the participant finishes reading a passage.

7 What Makes the Task Easy or Hard?

Having established that the prediction task at hand can be performed with a considerable degree of success, we now leverage the best performing single model RoBERT-Eye-F to obtain insights about the task itself. To this end, we introduce a new method for analyzing model performance that uses mixed effects modeling of model logits from data features. This method enables examining which trial features, which were not given to the model explicitly, affect the ability of the model to classify trials correctly. Differently from univariate methods often used for model performance analyses, our approach allows measuring the contribution of each feature *above and beyond* all the other features, while also taking into account the *non-i.i.d* nature of the data, where multiple participants read the same paragraph and multiple paragraphs are read by the same participant. The analysis takes advantage of the rich structure and auxiliary annotations of the OneStop dataset.

We define 10 features that capture various aspects of the trial. These include the following **participant** features over the item: Reading time before, within, and after the critical span, paragraph position in the experiment (1-54), and whether after having read the paragraph, the participant answered the given reading comprehension question correctly. We further include the following **item** (paragraph and question), reader-independent features: Paragraph length (in words), paragraph difficulty level (advanced / elementary), critical span start location (relative position, normalized by paragraph length), critical span length (normalized by paragraph length), and question difficulty (percentage of participants who answered the question incorrectly). Further details about these features are presented in Appendix E.

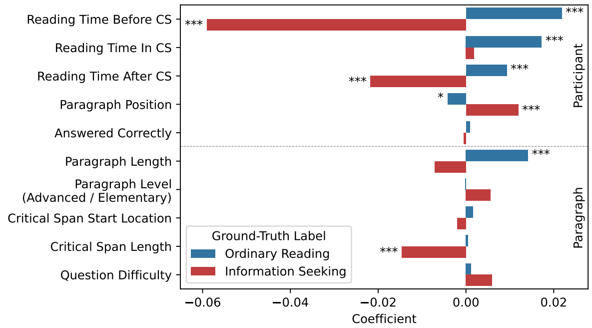


Figure 3: Coefficients from a mixed-effects model that predicts whether RoBERTEye-F's prediction for a given trial is correct from properties of the trial. CS stands for the critical span, the portion of the paragraph that contains the information that is essential for answering the question correctly. Two models are fitted separately for ordinary reading and information seeking trials. Predictors are z-normalized. Depicted are the coefficients of the fitted models after a 10x Bonferroni correction, to mitigate the risk of false positives when testing multiple hypotheses simultaneously. '*' $p < 0.05$, '**' $p < 0.01$, '***' $p < 0.001$.

To establish the relation of these features to task difficulty, we use a linear mixed effects model that uses these features to predict the probability that the model assigns to the correct label. In R notation:

$$P(\text{correct}) \sim \text{feat}_1 + \cdots + \text{feat}_{10} + (1 \mid \text{item}) \\ +(1 \mid \text{participant}) + (1 \mid \text{evaluation regime})$$

where the random effects account for correlations in predictions within participants, items and evaluation regimes[2]. We fit this model separately on the information seeking and ordinary reading trials. To make the contributions of the features to prediction accuracy comparable, we normalize each feature to be a z-score (zero mean and unit variance). We then examine feature contribution via the statistical significance, the magnitude and the sign of the corresponding coefficient. A significant coefficient for a feature indicates that it correlates with task difficulty, the absolute value determines its importance relative to other features, and the sign indicates the direction of the association.

The resulting feature coefficients are presented in Figure 3. In line with the findings of Shubi and Berzak (2023) on differences in reading speed between information seeking and ordinary reading around the critical span, we observe that prominent features for correctly classifying both information seeking and ordinary reading trials are read-

[2]Random effects structure is simplified not to include slopes due to model convergence issues.

ing times before and after the critical span. Faster readers before and after the critical span are easier to correctly classify as information seeking and harder to correctly classify as ordinary reading. Additionally, although not reflected in the reading speed analysis of Shubi and Berzak (2023), slower reading within the critical span is also beneficial for correct classification of ordinary reading trials. Longer paragraphs are also beneficial for classification of ordinary reading. We further find that shorter critical spans facilitate correct classification of information seeking trials, presumably by making information seeking more targeted and the identification of task critical information easier. Paragraph position is also significant in information seeking, suggesting that readers develop more efficient goal oriented reading strategies as they progress through the experiment. Overall, this analysis provides a highly interpretable characterization of both task difficulty and the underlying reading behavior in both reading regimes.

8 Related Work

Although the large majority of literature on the psychology of reading is concerned with ordinary reading, several studies did address goal-oriented (also referred to as task-based) reading. Most prior work focused on a small number of canonical tasks: skimming, speed reading and proofreading. Several studies found different eye movement patterns in these tasks as compared to ordinary reading (Just et al., 1982; Kaakinen and Hyönä, 2010; Schotter et al., 2014; Strukelj and Niehorster, 2018; Chen et al., 2023), and used predictive modeling to distinguish ordinary reading from skimming (Kelton et al., 2019). Rayner and Raney (1996) examined differences between ordinary reading and searching through the text for a target word. Prior work also analyzed eye movements during human linguistic annotation, often used for generating training data for NLP tools, such as annotation of named entities (Tomanek et al., 2010; Tokunaga et al., 2017). Differences in reading patterns were further found when readers were asked to take different perspectives on a given text (Kaakinen et al., 2002) or given different sets of learning goals (Rothkopf and Billington, 1979).

Our work is closest to Hahn and Keller (2023), Malmaud et al. (2020) and Shubi and Berzak (2023) who analyzed eye movement differences between ordinary reading and information seeking, where the information seeking goal is formulated using a reading comprehension question. All three studies found substantial differences in fixation and saccade patterns in information seeking as compared to ordinary reading, in particular before, within and after the text portions that are critical for the information seeking task. Here, we build on these findings, and examine whether these differences can be leveraged to automatically distinguish between these two reading regimes.

While the above studies focus primarily on descriptive data analysis, Hollenstein et al. (2023) took a predictive approach and attempted to automatically classify the reading task from eye movement features. In this study, 18 participants read single sentences from the ZuCo corpus (Hollenstein et al., 2020), and engaged either in ordinary reading or in an annotation of the presence of one of seven semantic relations in the sentence. While this benchmark is conceptually related to the current work, it is limited by the nature of the tasks, which focus on highly specialized linguistic annotations that are not performed by readers in everyday life. In the current study we take a different and more general stance on task based reading, with unrestricted questions that are more representative of the tasks commonly pursued by readers. More broadly, our work contributes to a nascent line of work which uses eye movements in reading for predicting properties of the reader's cognitive state with respect to the text, such as reading comprehension (Reich et al., 2022b; Shubi et al., 2024), as well as properties of the text itself, including document type (Kunze et al., 2013) and readability level (González-Garduño and Søgaard, 2017).

9 Summary and Discussion

Is it possible to decode reader goals from eye movements? We address this question by examining the possibility of automatically differentiating between ordinary reading and information seeking at the challenging granularity level of a single paragraph. We find that it is indeed possible to perform this task successfully, even before the participant finished reading. Model comparison reveals that the architecture, the granularity level of the eye movement representation and the inclusion of the underlying text are all important for the task. Our new error analysis method leverages the models to further reveal new insights on the factors that determine task difficulty.

10 Limitations

Our study has a number of limitations. The information seeking tasks are over individual paragraphs that span 3-10 lines of text. This leaves out shorter texts (e.g. single sentences) as well as longer texts. It is also restricted to newswire texts, and does not include texts from other genres. New datasets for the information seeking task, other types of tasks, additional populations (e.g. second language readers, younger and older participants), and datasets in languages other than English are all needed in order to study goal decoding more broadly. We further note that data collected in-lab, especially when the experiment involves frequent reading comprehension queries, may deviate from participants' reading patterns in their daily lives (Huettig and Ferreira, 2022). This can in turn limit generalization to real-world scenarios.

While the current work takes a first step in addressing the proposed task, considerable room for performance improvements remains for future work. New strategies for modeling eye movements with text are likely needed to fully exploit the potential of eye movements for this task. Furthermore, the addressed task is fundamentally limited in that it does not distinguish between different information seeking tasks. A natural direction for future work could address decoding of the specific question that was presented to the participant in the information seeking regime.

11 Ethics Statement

This work uses eye movement data collected from human participants. The data was collected by Berzak et al. (2025) under an institutional IRB protocol. All the participants provided written consent prior to participating in the eye tracking study. The data is anonymized. Analyses and modeling of eye movements in information seeking are among the main use cases for which the data was collected.

It has previously been shown that eye movements can be used for user identification (e.g. Bednarik et al., 2005; Jäger et al., 2020). We do not perform user identification in this study, and emphasize the importance of not storing information that could enable participant identification in future applications of goal decoding. We further stress that future systems that automatically infer reader goals are to be used only with explicit consent from potential users to have their eye movements collected and analyzed for this purpose.

Acknowledgments

This work was supported by ISF grant 1499/22.

References

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Douglas Bates, Phillip Alday, Dave Kleinschmidt, José Bayoán Santiago Calderón, Likan Zhan, Andreas Noack, Milan Bouchet-Valat, Alex Arslan, Tony Kelman, Antoine Baldassari, Benedikt Ehinger, Daniel Karrasch, Elliot Saba, Jacob Quinn, Michael Hatherly, Morten Piibeleht, Patrick Kofod Mogensen, Simon Babayan, Tim Holy, Yakir Luc Gagnon, and Yoni Nazarathy. 2024. JuliaStats/MixedModels.jl: v4.26.0.

Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and Pasi Fränti. 2005. Eye-movements as a biometric. In *Image Analysis: 14th Scandinavian Conference, SCIA 2005, Joensuu, Finland, June 19-22, 2005. Proceedings 14*, pages 780–789. Springer.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured Annotations for Reading Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. 2025. ;OneStop: A 360-participant english eye tracking dataset with different reading regimes. *PsyArXiv preprint*.

Nilavra Bhattacharya, Somnath Rakshit, Jacek Gwizdka, and Paul Kogut. 2020. Relevance prediction from eye-movements using semi-interpretable convolutional neural networks. In *Proceedings of the 2020 conference on human information interaction and retrieval*, pages 223–233.

Andrew P Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Xiuge Chen, Namrata Srivastava, Rajiv Jain, Jennifer Healey, and Tilman Dingler. 2023. Characteristics of deep and skim reading on smartphones vs. desktop: A comparative study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Shuwen Deng, David R. Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A. Jäger. 2023. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading.

In *Proceedings of the ACM on Human-Computer Interaction*, pages 1–24. Association for Computing Machinery.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

William Falcon and The PyTorch Lightning team. 2024. Pytorch lightning.

Ana Valeria González-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.

Michael Hahn and Frank Keller. 2023. Modeling task effects in human reading with neural network-based attention. *Cognition*, 230:105289.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Patrick Haller, Andreas Säuberli, Sarah Elisabeth Kiener, Jinger Pan, Ming Yan, and Lena Jäger. 2022. Eye-tracking based classification of mandarin chinese readers with and without dyslexia using neural sequence models. *arXiv preprint arXiv:2210.09819*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780. Conference Name: Neural Computation.

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. 2022. Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 1–15, Taipei, Taiwan. Association for Computational Linguistics.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France. European Language Resources Association.

Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A Jäger, and Nicolas Langer. 2023. The zuco benchmark on cross-subject reading task classification with eeg and eye-tracking data. *Frontiers in Psychology*, 13:1028824.

Falk Huettig and Fernanda Ferreira. 2022. The myth of normal reading. *Perspectives on Psychological Science*, page 17456916221127226.

Jukka Hyönä and Johanna K Kaakinen. 2019. Eye movements during reading. *Eye movement research: An introduction to its scientific foundations and applications*, pages 239–274.

Lena A Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler, and Tobias Scheffer. 2020. Deep eyedentification: Biometric identification using micro-movements of the eye. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 299–314. Springer.

Marcel Adam Just, Patricia A Carpenter, and MEJ Masson. 1982. What eye fixations tell us about speed reading and skimming. *Eye-lab Technical Report Carnegie-Mellon University, Pittsburgh*.

Johanna Kaakinen and Jukka Hyönä. 2010. Task Effects on Eye Movements During Reading. *Journal of experimental psychology. Learning, memory, and cognition*, 36:1561–6.

Johanna K Kaakinen, Jukka Hyönä, and Janice M Keenan. 2002. Perspective effects on online text processing. *Discourse processes*, 33(2):159–173.

Conor Kelton, Zijun Wei, Seoyoung Ahn, Aruna Balasubramanian, Samir R Das, Dimitris Samaras, and Gregory Zelinsky. 2019. Reading detection in real-time. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–5.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, 16(1-2):262–284.

Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. 2013. I know what you are reading: recognition of document types using mobile eye tracking. In *Proceedings of the 2013 international symposium on wearable computers*, pages 113–116.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. 2020. Bridging Information-Seeking Human Gaze and Machine Reading Comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 142–152, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yoav Meiri and Yevgeni Berzak. 2024. Déjà vu: Eye movements in repeated reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Diane C. Mézière, Lili Yu, Erik D. Reichle, Titus von der Malsburg, and Genevieve McArthur. 2023. Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 58(3):425–449.

Ralph Radach and Alan Kennedy. 2004. Theoretical perspectives on eye movements in reading: Past controversies, current issues, and an agenda for future research. *European journal of cognitive psychology*, 16(1-2):3–26.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Keith Rayner, Jane Ashby, Alexander Pollatsek, and Erik D Reichle. 2004. The effects of frequency and predictability on eye fixations in reading: Implications for the ez reader model. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4):720.

Keith Rayner and Gary E Raney. 1996. Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review*, 3(2):245–248.

Keith Rayner, Timothy J Slattery, Denis Drieghe, and Simon P Liversedge. 2011. Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):514.

David Robert Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A Jäger. 2022a. Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading. In *2022 Symposium on Eye Tracking Research and Applications*, pages 1–8.

David Robert Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A. Jäger. 2022b. Inferring Native and Non-Native Human Reading Comprehension and Subjective Text Difficulty from Scanpaths in Reading. In *2022 Symposium on Eye Tracking Research and Applications*, pages 1–8, Seattle WA USA. ACM.

Ernst Z Rothkopf and MJ Billington. 1979. Goal-guided learning from text: inferring a descriptive processing model from inspection times and eye movements. *Journal of educational psychology*, 71(3):310.

Elizabeth R Schotter, Klinton Bicknell, Ian Howard, Roger Levy, and Keith Rayner. 2014. Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition*, 131(1):1–27.

Elizabeth R Schotter and Brian Dillon. 2025. A beginner's guide to eye tracking for psycholinguistic studies of reading. *Behavior Research Methods*, 57(2):68.

Omer Shubi and Yevgeni Berzak. 2023. Eye movements in information-seeking reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Omer Shubi, Yoav Meiri, Cfir A. Hadar, and Yevgeni Berzak. 2024. Fine-grained prediction of reading comprehension from eye movements. In *The 2024 Conference on Empirical Methods in Natural Language Processing*.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexander Strukelj and Diederick C Niehorster. 2018. One page of text: Eye movements during regular and thorough reading, skimming, and spell checking. *Journal of Eye Movement Research*, 11(1).

Takenobu Tokunaga, Hitoshi Nishikawa, and Tomoya Iwakura. 2017. An eye-tracking study of named entity annotation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 758–764, Varna, Bulgaria. INCOMA Ltd.

Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jürgen Ziegler. 2010. A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1158–1167, Uppsala, Sweden. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142.

Duo Yang and Nora Hollenstein. 2023. Plm-as: Pretrained language models augmented with scanpaths for sentiment classification. *Proceedings of the Northern Lights Deep Learning Workshop*.

Appendix

A Models

A.1 Model Descriptions

Global Representation

Logistic Regression A logistic regression model with global eye movement measures from Mézière et al. (2023). The measures include averages of word reading times, single fixation duration, forward saccade length, the rate of regressions (saccades that go backwards), and skips (words that were not fixated) during first pass reading (i.e. before proceeding to the right of the word). All the features are standard measures from the psycholinguistic literature.

Word-based Representations

RoBERTEye-W(ords) (Shubi et al., 2024) is a RoBERTa transformer model (Liu et al., 2019) augmented with eye movements. This model concatenates word embeddings and word-level eye movement features in the model input.

MAG-Eye (Shubi et al., 2024) Integrates word-level eye movement features into a transformer-based language model by injecting them into intermediate word representations using a Multimodal Adaptation Gate (MAG) architecture (Rahman et al., 2020). The text is aligned with eye movements by duplicating each word-level eye movement feature for every sub-word token.

Fixation-based Representations

PLM-AS (Yang and Hollenstein, 2023) This model represents the eye movements sequence by reordering contextual word embeddings according to the order of the fixations over the text. This reordered sequence is processed through a Recurrent Neural Network (RNN), whose final hidden layer is used for classification.

Haller RNN (Haller et al., 2022) This model is similar to PLM-AS in that it receives word embeddings in the order of the fixations. Differently from PLM-AS, each word embedding is further concatenated with eye movement features.

RoBERTEye-F(ixations) (Shubi et al., 2024) uses the same architecture as RoBERTEye-W, but represents fixations rather than words. Each fixation input consists of a concatenation of the word embedding and eye movement features associated with the fixation.

BEyeLSTM (Reich et al., 2022a) represents both the fixation sequence and textual features, combining LSTMs (Hochreiter and Schmidhuber, 1997) and global features through a linear layer.

BEyeLSTM - No Text is a model that processes raw fixation data using an LSTM. The final hidden state of the LSTM is combined with global eye movement features to perform classification. The model is inspired by BEyeLSTM (Reich et al., 2022a), using the same eye movements feature set, without the text representations.

Eyettention (Deng et al., 2023) is a model that consists of a RoBERTa word sequence encoder and an LSTM-based fixation sequence encoder. It uses a cross-attention mechanism to align the input sequences. We use the adaptation of this model by (Shubi et al., 2024) for binary classification.

PostFusion-Eye (Shubi et al., 2024) is a model that consists of a RoBERTa word sequence encoder and a 1-D convolution-based fixation sequence encoder. It then uses cross-attention to query the word representations using the eye-movement representations, followed by concatenation and projection into a shared latent space.

Image Representations

We represent scanpaths as two-dimensional images, as illustrated in Figure 4. In this visualization, fixations are depicted as circles positioned at their original x-y coordinates from the screen display, with the entire representation cropped to maintain consistent dimensions. The diameter of each circle corresponds to the duration of the fixation. To indicate the sequential progression of the eye movements, we employ a gradient shading scheme. Additionally, we differentiate between saccade types by color-coding them according to the five categories established by Schotter and Dillon (2025) - forward saccade, skip, refixation, return sweep, regression, and another for any saccade typethat does not fall into this categorization. Note that for these features knowledge about the *existance* of text is needed, but not the textual content itself. We use the *convnextv2_base.fcmae_ft_in22k_in1k* and *vit_base_patch14_dinov2* versions of the ConvNextv2 and ViT models respectively.

A.2 Model Diagrams

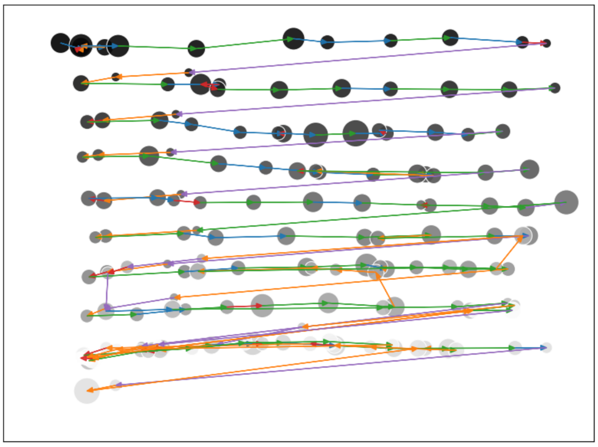
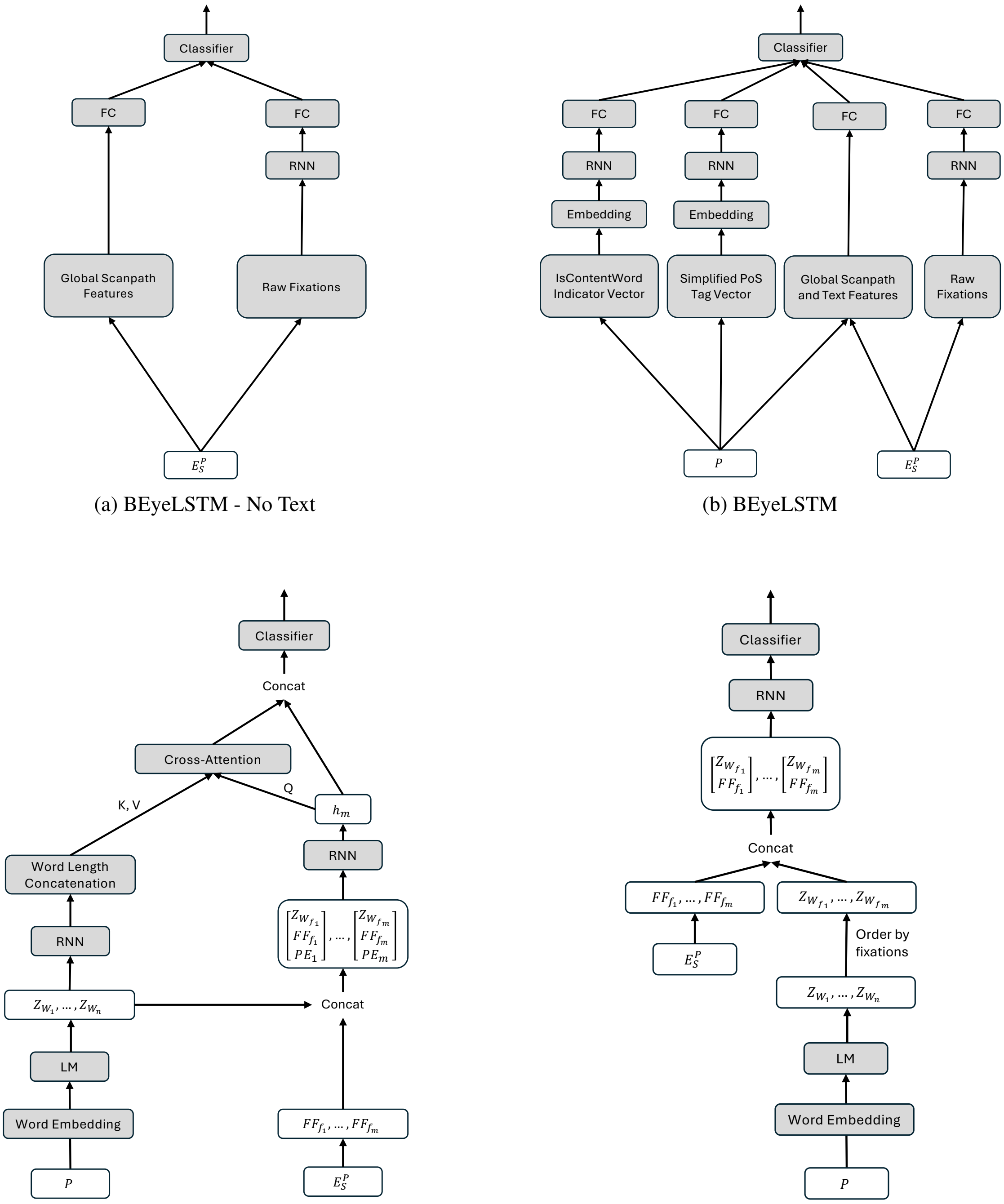


Figure 4: An example of a scanpath as an image as used for the image classification models.



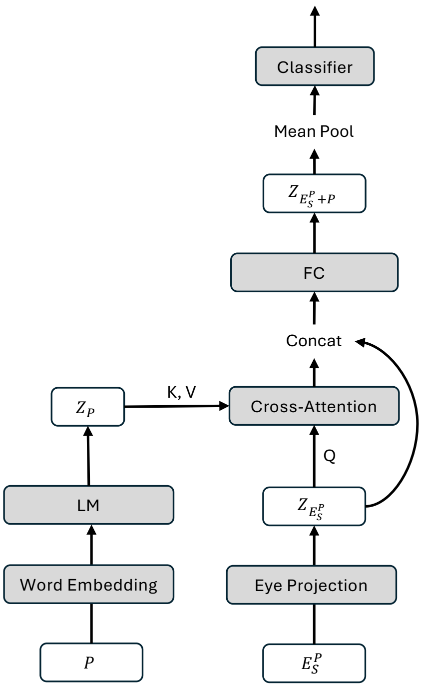
(a) BEyeLSTM - No Text

(b) BEyeLSTM

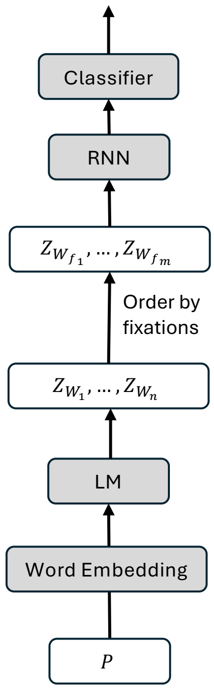
(c) Eyettention

(d) Haller RNN

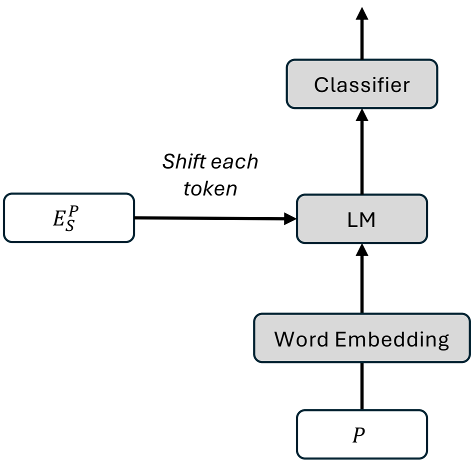
Figure 5: Visualization of the different model architectures (Part 1). P represents the paragraph, E_S^P the eye movements of participant S on P. LM stands for a language model, and FC for fully connected layers. FF_{f_i} stands for the fixation features and w_{f_i} for the word corresponding to the i-th fixation respectively.



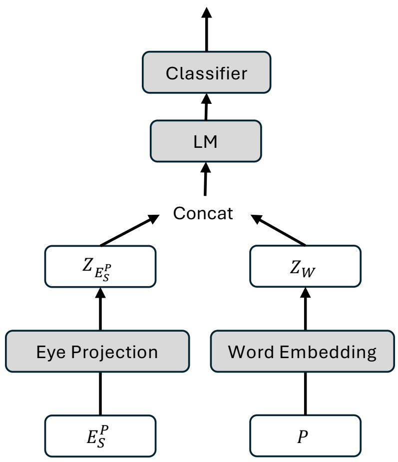
(a) PostFusion-QEye



(b) PLM-AS



(c) MAG-Eye



(d) RoBERTEye

Figure 6: Visualization of the different model architectures (Part 2). P represents the paragraph, E_S^P the eye movements of participant S on P. LM stands for a language model, and FC for fully connected layers. FF_{f_i} stands for the fixation features and w_{f_i} for the word corresponding to the i-th fixation respectively.

B OneStop Eye Movements Dataset - Additional Details

The textual data of OneStop consists of 162 paragraphs, 486 questions, and 972 unique paragraph–level–question triplets. The mean paragraph length is 109 words (min: 50; max: 165; std: 28). The mean length of Elementary paragraphs is 97 words (37 before the critical span, 30 inside it, and 30 after it), and of Advanced paragraphs 120 words (48 before the critical span, 34 inside it, and 38 after it). Each question has 20 responses, 10 for the Advanced version and 10 for the Elementary version. The mean experiment duration is approximately one hour. The raw millisecond gaze location is pre-processed into fixations and saccades using the SR Data Viewer software (`v4.3.210`).

Table 3: An example of a OneStopQA paragraph (Advanced and Elementary version) along with one of its three questions. The critical span is marked in bold red. Adapted from Berzak et al. (2020).

Advanced	A major international disagreement with wide-ranging implications for global drugs policy has erupted over the right of Bolivia's indigenous Indian tribes to chew coca leaves, the principal ingredient in cocaine. **Bolivia has obtained a special exemption from the 1961 Single Convention on Narcotic Drugs, the framework that governs international drugs policy, allowing its indigenous people to chew the leaves.** Bolivia had argued that the convention was in opposition to its new constitution, adopted in 2009, which obliges it to "protect native and ancestral coca as cultural patrimony" and maintains that coca "in its natural state ... is not a dangerous narcotic."
Elementary	A big international disagreement has started over the right of Bolivia's indigenous Indian tribes to chew coca leaves, the main ingredient in cocaine. This could have a significant effect on global drugs policy. **Bolivia has received a special exemption from the 1961 Convention on Drugs, the agreement that controls international drugs policy. The exemption allows Bolivia's indigenous people to chew the leaves.** Bolivia said that the convention was against its new constitution, adopted in 2009, which says it must "protect native and ancestral coca" as part of its cultural heritage and says that coca "in its natural state ... is not a dangerous drug."
Question	**What was the purpose of the 1961 Convention on Drugs?**
Answers	A Regulating international policy on drugs B Discussing whether indigenous people in Bolivia should be allowed to chew coca leaves C Discussing the legal status of Bolivia's constitution D Negotiating extradition agreements for drug traffickers

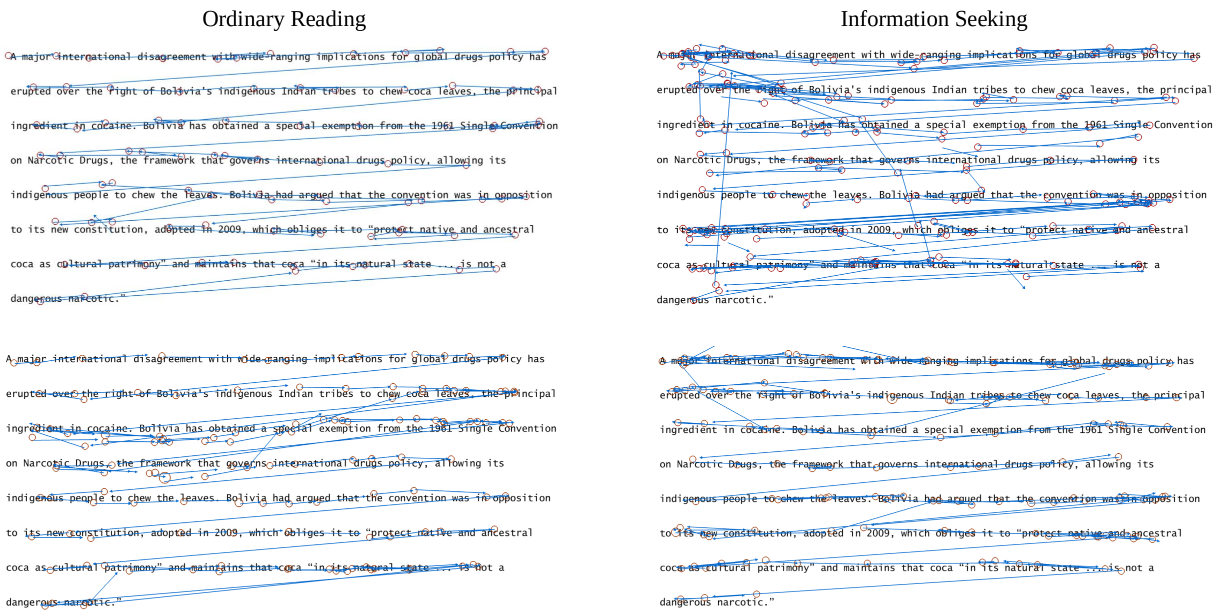


Figure 7: Examples of eye movements over a single passage; left: ordinary reading, right: information seeking. Circles represent fixations, and lines represent saccades.

C Model Training and Hyperparameters

All neural network-based models were trained using the PyTorch Lighting (Falcon and team, 2024) library on NVIDIA A100-40GB and L40S-48GB GPUs.

Since the models we use were developed for different tasks and datasets, we conducted a hyperparameter search for each model. The search space for each model is described below. In all cases, it includes the optimal parameters reported in the work that introduced the model, extended to provide a fair comparison between models.

For all neural models we train with learning rates of $\{0.00001, 0.00003, 0.0001\}$ and dropout of $\{0.1, 0.3, 0.5\}$ following Shubi et al. (2024). Additionally, for all models that make use of word embeddings, we include both frozen and unfrozen language model variants in the search space.

- For **Logistic Regression**, the search space for the regularization parameter C is $\{0.1, 5, 10, 50, 100\}$, with and without an L2 penalty.
- Following (Reich et al., 2022a), for **BEyeLSTM** and **BEyeLSTM - No Text**, the search space consists of learning rates $\{0.001, 0.003, 0.01\}$, embedding dimensions $\{4, 8\}$ and hidden dimensions $\{64, 128\}$.
- For **MAG-Eye** the search space for injection layer index is: $\{0, 11, 23\}$.
- Following Yang and Hollenstein (2023), we train **PLM-AS** and **Haller RNN** with dropout rate search space of 0.1, and for PLM-AS, we use LSTM layer counts of $1, 2$. Additionally, as in (Haller et al., 2022), we search over LSTM hidden layer sizes of $10, 40, 70$. For PLM-AS, the LSTM hidden layer size is fixed at 1024 to match the LM's dimensionality in Yang and Hollenstein (2023).
- For **Eyettention**, we also train with a learning rate of 0.001 and dropout of 0.2, as done in (Deng et al., 2023)
- For **PostFusion-Eye**, the 1D convolution layers have a kernel size of three, stride 1 and padding 1.

We train the deep-learning based models for a maximum of 40 epochs, with early stopping after 8 epochs if no improvement in the validation error is observed. Following Liu et al. (2019); Mosbach et al. (2021); Shubi et al. (2024) we use the AdamW optimizer (Loshchilov and Hutter, 2018) with a batch size of 16. MAG-Eye, RoBERTEye and PostFusion-Eye use a linear warm-up ratio of 0.06, and a weight decay of 0.1. We standardize each eye movements feature using statistics computed on the training set, to zero mean unit variance.

The code base for this project was developed with the assistance of GitHub Copilot, an AI-powered coding assistant. All generated code was carefully reviewed.

D Additional Results

Below we present the test set ROC curves across the ten cross-validation splits and their corresponding AUROC scores (mean and standard deviation). We also provide accuracy results for the validation set.

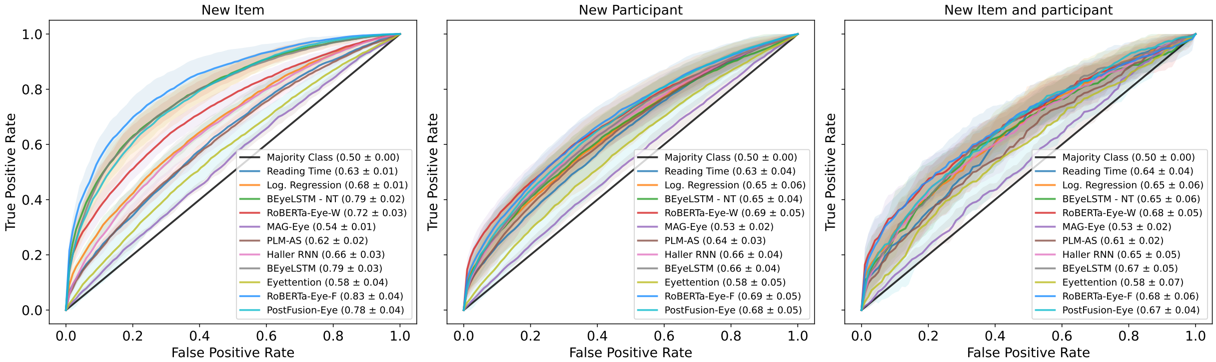


Figure 8: ROC Curves by model and evaluation regime. Each curve represents a different model across the ten cross-validation splits, with the corresponding AUROC scores (mean and standard deviation) provided in the legend.

Model	**Gaze Representation**	**Text Representation**	**New Item**	**New Participant**	**New Item & Participant**	**All**
Majority Class / Chance	–	–	$50.0_{\pm 0.0}{}_{+++}$	$50.0_{\pm 0.0}{}_{+++}$	$50.0_{\pm 0.0}{}_{+++}$	$50.0_{\pm 0.0}{}_{+++}$
Reading Time	–	–	$58.9_{\pm 0.5}{}_{+++}$	$58.9_{\pm 1.0}{}_{+++}$	$60.4_{\pm 1.3}{}_{+++}$	$58.9_{\pm 0.5}{}_{+++}$
Log. Regression (Mézière et al., 2023)	Global	–	$62.6_{\pm 0.3}{}^{**}_{+++}$	$60.6_{\pm 1.5}{}_{+++}$	$61.0_{\pm 1.8}{}_{+++}$	$61.6_{\pm 0.8}{}^{*}_{+++}$
BEyeLSTM - No Text	Fixations	–	$73.2_{\pm 0.6}{}^{***}_{+++}$	$64.9_{\pm 1.0}{}^{***}_{+++}$	$65.1_{\pm 1.4}{}^{*}_{+++}$	$68.8_{\pm 0.5}{}^{***}_{+++}$
ConvNext v2	Image of Scanpath	–	$71.2_{\pm 0.5}{}^{***}_{+++}$	$65.3_{\pm 0.9}{}^{***}_{+++}$	$65.3_{\pm 1.3}{}^{*}_{+++}$	$68.0_{\pm 0.5}{}^{***}_{+++}$
ViT	Image of Scanpath	–	$71.7_{\pm 0.3}{}^{***}_{+++}$	$65.8_{\pm 0.9}{}^{***}_{+++}$	$67.4_{\pm 0.7}{}^{***}_{+++}$	$68.6_{\pm 0.5}{}^{***}_{+++}$
RoBERTEye-W (Shubi et al., 2024)	Words	Emb+LF	$65.1_{\pm 0.6}{}^{***}_{+++}$	$64.9_{\pm 1.1}{}^{***}_{+++}$	$65.2_{\pm 1.4}{}^{*}_{+++}$	$65.1_{\pm 0.6}{}^{***}_{+++}$
MAG-Eye (Shubi et al., 2024)	Words	Emb+LF	$53.7_{\pm 0.2}{}_{+++}$	$53.5_{\pm 0.5}{}_{+++}$	$52.6_{\pm 0.6}{}_{+++}$	$53.5_{\pm 0.2}{}_{+++}$
PLM-AS (Yang and Hollenstein, 2023)	Fixations Order	Emb	$59.1_{\pm 0.5}{}_{+++}$	$61.1_{\pm 0.7}{}_{+++}$	$58.6_{\pm 0.9}{}_{+++}$	$60.1_{\pm 0.4}{}_{+++}$
Haller RNN (Haller et al., 2022)	Fixations	Emb	$62.3_{\pm 0.6}{}^{**}_{+++}$	$62.9_{\pm 1.2}{}^{**}_{+++}$	$63.4_{\pm 1.3}{}_{+++}$	$62.5_{\pm 0.6}{}^{**}_{+++}$
BEyeLSTM (Reich et al., 2022a)	Fixations	LF	$72.3_{\pm 0.6}{}^{***}_{+++}$	$65.0_{\pm 1.3}{}^{***}_{+++}$	$66.1_{\pm 1.2}{}^{**}_{+++}$	$68.5_{\pm 0.6}{}^{***}_{+++}$
Eyettention (Deng et al., 2023)	Fixations	Emb+LF	$56.4_{\pm 0.8}{}_{+++}$	$56.6_{\pm 0.9}{}_{+++}$	$58.6_{\pm 1.1}{}_{+++}$	$56.6_{\pm 0.5}{}_{+++}$
RoBERTEye-F (Shubi et al., 2024)	Fixations	Emb+LF	$90.7_{\pm 0.3}{}^{***}$	$91.9_{\pm 0.5}{}^{***}$	$88.7_{\pm 0.9}{}^{***}$	$91.2_{\pm 0.3}{}^{***}$
PostFusion-Eye (Shubi et al., 2024)	Fixations	Emb+LF	$89.2_{\pm 0.4}{}^{***}_{+++}$	$91.3_{\pm 0.5}{}^{***}$	$87.8_{\pm 0.6}{}^{***}$	$90.1_{\pm 0.4}{}^{***}_{+++}$
Logistic Ensemble			$92.3_{\pm 0.9}{}^{***}$	$93.2_{\pm 1.2}{}^{***}$	$89.6_{\pm 2.3}{}^{***}$	$92.6_{\pm 0.8}{}^{***}$

Table 4: Validation accuracy results aggregated across 10 cross-validation splits. 'Emb' stands for word embeddings, 'LF' for linguistic word features such as word length, frequency and surprisal. Model performance is compared to the Reading Time baseline using a linear mixed effects model. In R notation: $is\_correct \sim model + (model \mid participant) + (model \mid paragraph)$. Significant gains over this baseline are marked with '*' $p < 0.05$, '**' $p < 0.01$ and '***' $p < 0.001$ in superscript, and significant drops compared to the best model in each regime are marked in subscript with '+'.

E Feature Descriptions

We define 10 features that capture various aspects of the trial. These include the following reader features over the item:

1-3. **Reading time before, within, and after critical span**: These features are motivated by the findings of Malmaud et al. (2020) and (Shubi and Berzak, 2023) regarding faster reading times in information seeking compared to ordinary reading, primarily before and after the critical span, as well as the reported classification results of the Reading Time baseline.

4. **Paragraph position** (1-54): Each participant reads 54 paragraphs, in a random article order. This feature captures the position of the paragraph in the experiment's presentation sequence. It is included as reading strategies can change as the experiment progresses (e.g. Meiri and Berzak (2024) show that readers become faster as the experiment progresses).

5. **Answered correctly**: this feature encodes whether after having read the passage, the participant answered the given reading comprehension question correctly. It captures participant-specific task difficulty and the extent to which the participant read the passage attentively.

We further include the following item (paragraph and question), reader-independent features:

6. **Paragraph length** (in words): this feature is chosen as we hypothesize that more data could lead to more accurate predictions for the item.

7. **Paragraph level** (Advanced / Elementary): is chosen as eye movements could be influenced by the difficulty level of the text, for example through differences in word frequency and surprisal (Singh et al., 2016; Hollenstein et al., 2022).

8. **Critical span start location** (relative position, normalized by paragraph length): Shubi and Berzak (2023) showed skimming-like reading patterns after processing task critical information in information seeking. We thus hypothesize that earlier appearance of task critical information could facilitate the ability to correctly identify information seeking reading.

9. **Critical span length** (normalized by paragraph length): we also hypothesize that less task critical information during information seeking could further aid distinguishing it from ordinary reading.

10. **Question difficulty** (percentage of participants who answered the question incorrectly): estimated from the train data. We include this feature as it can influence eye movements in information seeking, with harder questions potentially obscuring patterns of goal oriented reading.

F Pairwise agreement between models by evaluation regime

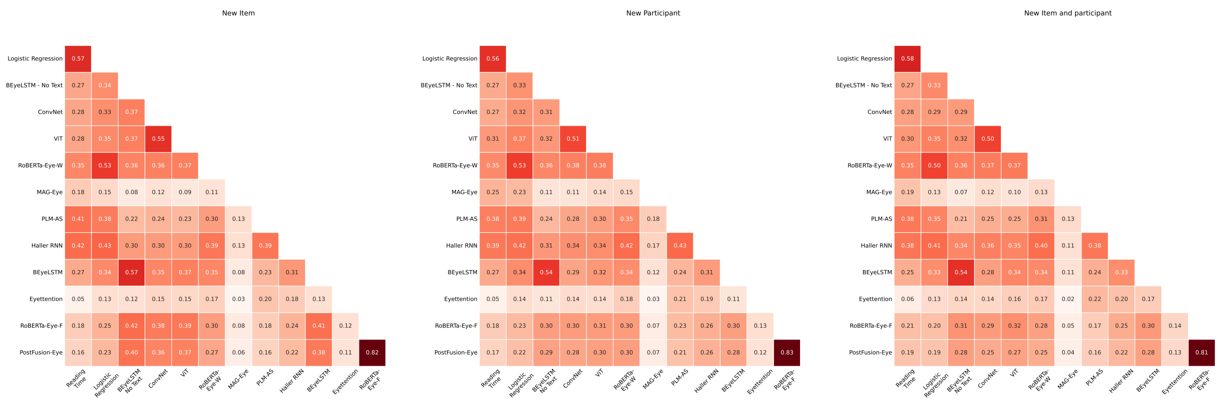


Figure 9: Pairwise Cohen's Kappa agreement between model predictions on the validation set by evaluation regime.