Decoding Reading Goals from Eye Movements

Omer Shubi[*,1]**, Cfir Avraham Hadar**[*,1]**, Yevgeni Berzak**[1,2]

[1]Faculty of Data and Decision Sciences,
Technion - Israel Institute of Technology, Haifa, Israel
[2]Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, Cambridge, USA
{shubi,kfir-hadar}@campus.technion.ac.il,berzak@technion.ac.il

Abstract

Readers can have different goals with respect to the text they are reading. Can these goals be decoded from the pattern of their eye movements over the text? In this work, we examine for the first time whether it is possible to decode two types of reading goals that are common in daily life: *information seeking* and *ordinary reading*. Using large scale eye-tracking data, we apply to this task a wide range of state-of-the-art models for eye movements and text that cover different architectural and data representation strategies, and further introduce a new model ensemble. We systematically evaluate these models at three levels of generalization: new textual item, new participant, and the combination of both. We find that eye movements contain highly valuable signals for this task. We further perform an error analysis which builds on prior empirical findings on differences between ordinary reading and information seeking and leverages rich textual annotations. This analysis reveals key properties of textual items and participant eye movements that contribute to the difficulty of the task.[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 saccadic fashion, where there are periods of time in which the gaze is stable at a specific location, called *fixations*, and rapid transitions between fixations called *saccades* (Rayner, 1998). This sequence of fixations and saccades is generally hypothesized to contain rich information about how readers interact with text. Automatic decoding of such information is currently a growing area of research (e.g. Reich et al., 2022; Shubi et al., 2024).

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 have a profound impact on online linguistic processing and on 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 Natural Language Processing (NLP) and machine learning (ML) have primarily focused on one reading regime, which can be referred to as *ordinary reading*. In this regime, the reader's goal is typically general comprehension of the text. Nearly all broad coverage datasets used in research on eye movements in reading, such as Dundee (Kennedy et al., 2003), MECO (Siegelman et al., 2022) and CELER (Berzak et al., 2022) were collected in this regime. Other forms of reading, although widely acknowledged (Radach & Kennedy, 2004), received much less attention and remain understudied.

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. Prior work suggests that on average across participants and texts, there are substantial differences in eye movement
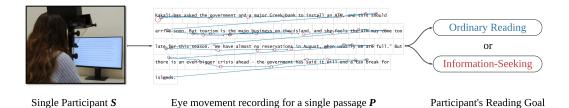

Figure 1: Proposed task: decoding whether a reader is seeking specific information or reading for general comprehension, given their eye movements over a single passage. In the eye movements image, circles represent fixations, and lines represent saccades. Bounding boxes mark word Interest Areas (fixations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

patterns between these two reading regimes (Hahn & Keller, 2023; Shubi & Berzak, 2023). However, it is currently unknown whether there is sufficient signal in eye movements 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. We use a range of recent models that implement different strategies for representing eye movements and combining them with text, and perform systematic evaluations of their individual and ensemble generalization abilities. We demonstrate that it is indeed possible to perform this task with relatively high accuracy rates. We further perform a systematic error analysis which leverages rich data annotations to reveal key axes of variation that contribute to task difficulty.

The main contributions of this work 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 or in information seeking.

- **Modeling and Evaluation**: We adapt and apply to this task 10 different state-of-the-art predictive models for eye movements in reading. We further introduce an ensemble model which outperforms the best single model. We characterize the generalization ability of all the models across textual items and participants.

- **Error Analysis**: We use statistical modeling and detailed textual annotations to perform an error analysis which reveals interpretable axes of variation in item and participant-related properties of trials that are indicative of classification difficulty. This analysis provides new insights on the nature of the task.

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 $\mathcal{C}(S, P) \rightarrow \{\text{Information Seeking, Ordinary Reading}\}$, we aim to approximate $\mathcal{C}$ with a classifier h:

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

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 P. We assume that the participant has not read the paragraph previously.

The information seeking regime is operationalized by presenting the participant with a 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. Figure 1 presents the task schematically.

3 MODELING

Eye movements during reading present a highly challenging case of temporally *and* spatially aligned multimodal data, where fixations are temporal and correspond to specific words in the text. We use a wide range of state-of-the-art architectures for processing this data. While the models were originally applied to other prediction tasks, such as predicting participant's reading comprehension and various participant properties, they offer general purpose mechanisms for processing eye movement data during reading that can be applied to other tasks, and cover different 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), the granularity at which eye movement information is represented (global averages across the text, single word, or a single fixation), and for the multimodal approaches, the nature of the text representations and strategy for combining them with eye movements. These models are directly applicable, in some cases with small adaptations, to the task of reading goal prediction.

3.1 EYE MOVEMENTS-ONLY MODELS

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

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.

FIXATION-BASED REPRESENTATION

BEyeLSTM - No Text is our adaptation of the BEyeLSTM model (Reich et al., 2022) described below, which uses only eye movements input without the underlying text. This model is more expressive compared to the logistic regression model, in that rather than averaging eye movement measures for the entire trial, it explicitly represents each fixation in the eye movement trajectory. The final hidden state of an LSTM (see below), combined with global eye movement features, is then used for classification.

3.2 EYE MOVEMENTS AND TEXT MODELS

We examine 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.

WORD-BASED REPRESENTATIONS

RoBERTa-Eye-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 & Hollenstein, 2023) This model represents the eye movements sequence by re-ordering 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.

RoBERTa-Eye-F(ixations) (Shubi et al., 2024) uses the same architecture as RoBERTa-Eye-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., 2022) represents both the fixation sequence and textual features, combining LSTMs (Hochreiter & Schmidhuber, 1997) and global features through a linear layer.

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.

3.3 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). Here, we introduce two such 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 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 Malmaud et al. (2020), Hahn & Keller (2023) and Shubi & 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 (Malmaud et al., 2020), 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 & 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. The mean paragraph length is 109 words. An example of a OneStopQA paragraph along with one question and its critical span annotation is provided in Table 2 in Appendix A.

Eye movements data for OneStopQA were collected in-lab from 360 adult native English speakers using an EyeLink 1000+ eye tracker (SR Research), a state-of-the-art eyetracking device for reading research. Each participant read a batch of 10 articles (54 paragraphs) paragraph by paragraph and answered a comprehension question about each paragraph. Each paragraph was shown to a participant either in the Advanced or Elementary version, along with one of the three possible questions. 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.

In sum, each participant read 54 paragraphs, all of which are either in information seeking or in ordinary reading. 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 5 in Appendix A shows example trials for each of the two reading regimes. Additional data statistics are described in Appendix A.

4.2 MODEL TRAINING AND EVALUATION PROTOCOL

We adopt the training and evaluation procedure of Shubi et al. (2024) for prediction of reading comprehension from eye movements, with several modifications for our task. Similarly to Shubi et al. (2024), we use 10-fold cross validation, addressing three levels of model generalization:

- New Item (paragraph): eye-tracking data is available during training for the participant but not for the paragraph.

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

- New Item & Participant: Neither the participant nor the paragraph are in the training data.

We further report aggregated results across all three regimes.

The data is split 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 split further guarantees an equal number of participants from each OneStopQA batch in each data portion.

Shubi et al. (2024) predicted reading comprehension separately in information seeking and ordinary reading, and thus defined separate data splits for the information seeking and ordinary reading trials. They further stratified items by the selected answers to the reading comprehension questions. As these steps are not relevant for our task, we use the union of the information seeking and ordinary reading trials in each part of each split, and stratify trials by reading goal rather than by reading comprehension answers.

Each data split contains approximately 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, approximately 90% of the trials in the dataset appear in each of the
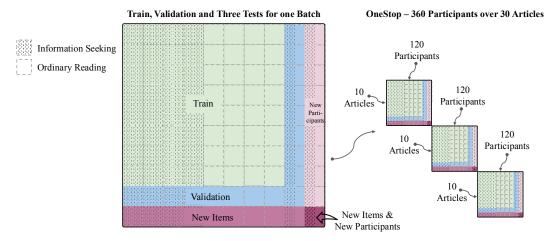
Figure 2: A schematic depiction of one of the 10 splits into train, validation, and the three test sets for one batch of 10 OneStopQA articles and 120 participants. Dashed lines denote information seeking trials. The full data split consists of the union of three such splits. Adapted from (Shubi et al., 2024).

New Participant and New Item evaluation regimes, and 10% in the New Item & Participant regime. Figure 2 presents this breakdown for one batch split.

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. All neural network-based models were trained using the PyTorch Lighting (Falcon & team, 2024) library on NVIDIA A100-40GB and L40S-48GB GPUs. Further details regarding the training procedure, including the full hyperparameter search space for all the models are provided in Appendix B.

4.3 STATISTICAL TESTS FOR MODEL PERFORMANCE COMPARISONS

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 performing statistical tests, we fit linear mixed-effects models with maximal random effects for items and participants (Barr et al., 2013). To this end, we use the MixedModels package in Julia (Bates et al., 2024).

5 RESULTS

Test set accuracy results are presented in Table 1. In line with prior findings regarding faster reading in information seeking compared to ordinary reading (Malmaud et al., 2020; Hahn & Keller, 2023; Shubi & Berzak, 2023), the Reading Time baseline yields above chance accuracies ($p < 0.01$ in all evaluation regimes), thus providing a strong benchmark for the eyetracking-based models. Among the 10 examined models, RoBERTa-Eye-F achieves the highest accuracy in the New Item, New Item & Participant and All evaluations, while RoBERTa-Eye-W slightly outperforms it in the New Participant regime. Differently from the New Item and the New Participant regimes, where multiple models outperform the Reading Time baseline, in the most challenging New Item & Participant regime, the improvements over this baseline are not statistically significant.

In addition to RoBERTa-Eye-F, three other fixation based models, PostFusion-Eye and the two BEyeLSTM models, excel in the New Item regime. One possible explanation for this outcome is better generalization of fixation based models to new items compared to new participants. Note however, that due to the between-subjects design where all the training and test examples for a given participant have the same label, it could alternatively reflect, at least in part, an ability of fixation based models to learn participant specific reading behavior without explicit information on the participant (i.e. identify the participant), an ability that is not directly pertinent to the task at hand. The current experimental setup makes it challenging to adjudicate between these two possibilities.

Table 1: Test accuracy results aggregated across 10 cross validation folds. 'Emb' stands for word embeddings, 'Ling. Feat.' for linguistic word properties 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 '***' $p < 0.001$.

Model	Gaze Representation	Text Representation	New Item	New Participant	New Item & Participant	All
Majority Class / Chance	None	None	50.0	50.0	50.0	50.0
Reading Time	None	None	59.0	58.9	60.4	59.0
Log. Regression (Mézière et al., 2023)	Global	None	62.4**	60.6	60.8	61.4*
BEyeLSTM - No Text	Fixations	None	71.5***	61.0	61.5	66.0***
RoBERTa-Eye-W (Shubi et al., 2024)	Words	Emb + Ling. Feat.	65.9***	63.2**	62.3	64.4***
MAG-Eye (Shubi et al., 2024)	Words	Emb + Ling. Feat.	52.1	52.3	51.5	52.1
PLM-AS (Yang & Hollenstein, 2023)	Fixations Order	Emb	58.6	59.5	57.5	59.0
Haller RNN (Haller et al., 2022)	Fixations	Emb	61.3	61.5	60.3	61.3
BEyeLSTM (Reich et al., 2022)	Fixations	Ling. Feat.	71.4***	61.6	62.2	66.2***
Eyettention (Deng et al., 2023)	Fixations	Emb + Ling. Feat.	55.8	55.7	55.4	55.8
RoBERTa-Eye-F (Shubi et al., 2024)	Fixations	Emb + Ling. Feat.	**74.7***	63.1*	**62.6**	**68.5***
PostFusion-Eye (Shubi et al., 2024)	Fixations	Emb + Ling. Feat.	70.7***	62.2*	61.4	66.1***

In either case, it is highly non-trivial that models are able to generalize from prior participant data to new items. In Appendix C, we further include ROC curves and AUC results (Bradley, 1997) in Figure 6, and validation set accuracies in Table 3. The outcomes of these evaluations are consistent with the test results reported in Table 1.
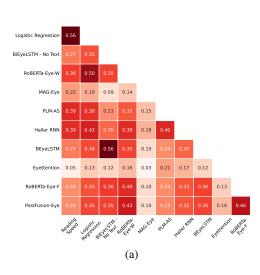
5.1 LOGISTIC ENSEMBLE

Ensemble models combine predictions from multiple individual models, and often yield better results than single models, especially when models have diverse predictive behaviors (Sagi & Rokach, 2018). To examine the appropriateness of this approach for our case, we compute the agreement between models' predictions, quantified with Cohen's Kappa (Cohen, 1960). Figure 3a depicts the pairwise agreement rates across models in the validation data, where we observe mostly moderate agreement rates. Figure 7 in Appendix D further suggests consistency across the three evaluation regimes. These outcomes indicate that while the examined models do not always differ substantially in accuracy, they likely capture different aspects of the eye movement data and the task.

Motivated by the results of this analysis, we introduce a simple **Logistic Ensemble**: a 10-feature logistic regression model that predicts the reading goal from the probability outputs of our 10 models. The model is trained for each fold on the validation set and tested on the corresponding three test sets. In Figure 3b we present the performance of the Logistic Ensemble, against the best performing model, RoBERTa-Eye-F, and the Reading Time baseline. We find that the Logistic Ensemble improves over the accuracy of RoBERTa-Eye-F in all the regimes, with an accuracy of 77.3 for New Items, 64.6 for New Participants, 64.3 for New Items & Participants, and 70.5 overall. The improvements over RoBERTa-Eye-F are statistically significant in all but the New Item & Participant regime, in which it is the only model that is statistically better than the Reading Time baseline. These performance improvements suggest that the information encoded by the different models is to some extent complementary.

6 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 model to obtain insights about the task itself. We do so by examining which trial features, which were not given to the model explicitly, affect the ability of the model to classify trials correctly. This 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 reader features over the item:
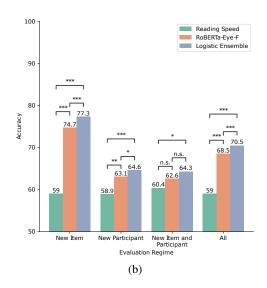
(a)   (b)

Figure 3: (a) Pairwise Cohen's Kappa agreement between model predictions on the validation set. (b) Test accuracy of the Logistic Ensemble model compared to the best single model RoBERTa-Eye-F and the Reading Time baseline for the ordinary reading vs. information-seeking goal decoding task, using 10-fold cross-validation. Statistical significance is tested with a linear mixed effects model. In R notation: $is\_correct \sim model + (model \mid participant) + (paragraph \mid parag)$. '*' $p < 0.05$, '**' $p < 0.01$ and '***' $p < 0.001$.

1-3. **Reading time before, within, and after critical span**: These features are motivated by the findings of Malmaud et al. (2020) and Shubi & 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 & 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 & 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): based on data from the Prolific QA experiment in Berzak et al. (2020). We include this feature as it can influence eye movements in information seeking, with harder questions potentially obscuring patterns of goal oriented reading.
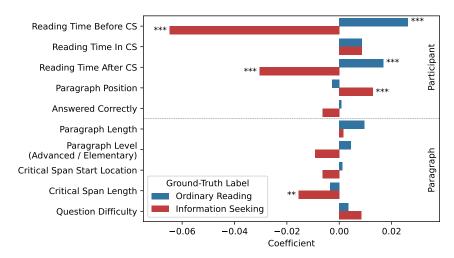
Figure 4: Coefficients from a mixed-effects model that predicts whether RoBERTa-Eye-F's prediction for a given trial is correct from properties of the trial. CS stands for the critical span, the portion of the paragraph that contains the information that is essential for answering the question correctly. 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. '*' $p < 0.05$, '**' $p < 0.01$, '***' $p < 0.001$.

In order to establish the relation of these features to task difficulty, we use a linear mixed effect model that uses these features to predict whether the trial prediction of the best performing model, RoBERTa-Eye-F, was correct. In R notation:

$$\text{is\_correct} \sim 1 + \text{feat}_1 + \text{feat}_2 + \cdots + \text{feat}_{10} + (1 \mid \text{participant}) + (1 \mid \text{item}) + (1 \mid \text{evaluation regime})$$

where the random effects account for correlations in predictions within participants, items and evaluation regimes[2]. The advantage of a multiple regression model over separate univariate analyses of each feature, is the ability to tease apart the contribution of each feature above and beyond all other features. 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 statistical significance, magnitude and sign of the corresponding coefficient. A significant coefficient for a feature indicates that it correlates with task difficulty, the absolute value determines its importance rank relative to other features, and the sign indicates the direction of the association.

The resulting feature coefficients are presented in Figure 4. In line with the findings of Shubi & Berzak (2023), we observe that the most prominent features for correctly classifying both information seeking and ordinary reading trials are reading time before and after the critical span. Slower readers in these regions are easier to correctly classify as ordinary reading and harder to correctly classify as information seeking. We also find, as hypothesized, that shorter critical spans facilitate correct classification of information seeking trials, presumably by making information seeking more targeted. Finally, paragraph position is 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 task difficulty, with insights on reading behavior in both reading regimes.

7 RELATED WORK

The vast majority of the literature on the psychology of reading and its interfaces with computational modeling is concerned with ordinary reading. However, several studies did address goal-oriented (also referred to as task-based) reading. Most prior work focused on a small number of canonical

---

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

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 & Hyönä, 2010; Schotter et al., 2014). Rayner & 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 also 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 & Billington, 1979).

Our work is closest to Malmaud et al. (2020), Hahn & Keller (2023) and Shubi & 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 for automatic discrimination 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, participants read single sentences from the ZuCo corpus (Hollenstein et al., 2020), and engaged either in ordinary reading or in an annotation task where they had to determine the presence of one of seven semantic relations in the sentence. While this benchmark is conceptually related to the current work, it is restricted in scope and ecological validity. First, the benchmark contains only 18 participants, which makes generalization and reliable evaluations highly challenging. More fundamentally, the framework is limited by the nature of the tasks, which focus on highly specialized linguistic annotations that are not performed by readers in everyday life. If viewed through the lens of question answering, the setup comprises only seven possible questions that concern the presence of a semantic relation in the sentence. 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.

8 SUMMARY AND DISCUSSION

Is it possible to decode reader goals from eye movements? We address this question by examining the possibility to automatically discriminate 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 with considerable success, and show that eye movements provide valuable information for this task, above and beyond reading time. We further find that different models capture different aspects of the signal and the task, and can be successfully combined with ensembling, which yields the strongest results for our task. Our error analysis leverages the models to further reveal new insights on the factors that determine task difficulty.

While the current work takes a first step in addressing the proposed task, ample room for performance improvements remains for future work, especially in the new participant regimes. 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. Finally, 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.

9 ETHICS STATEMENT

This work uses eye movement data collected from human participants. The data was collected by Malmaud et al. (2020) under an institutional IRB protocol. All the participants provided written consent prior to participating in the eyetracking study. The data is anonymized.

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.

10 REPRODUCIBILITY STATEMENT

We describe the model training and selection procedure, the evaluation protocol, and the hardware and software specifications in Section 4.2. The hyperparameter search space for each model is described in Appendix B. The dataset used in the experiments is described in Section 4.1 and appendix A. The source code, which implements the models, experimental procedures, and analysis, is available at: `https://github.com/lacclab/Goal-Decoding-from-Eye-Movements`. Data will be made publicly available.

REFERENCES

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, April 2013. ISSN 0749-596X. doi: 10.1016/j.jml.2012.11.001. URL `https://www.sciencedirect.com/science/article/pii/S0749596X12001180`.

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. JuliaStats/MixedModels.jl: v4.26.0, September 2024. URL `https://zenodo.org/records/13754074`.

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

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

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. CELER: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*, pp. 1–10, apr 2022.

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

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

Shuwen Deng, David R. Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A. Jäger. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. In *Proceedings of the ACM on Human-Computer Interaction*, pp. 1–24. Association for Computing Machinery, may 2023. doi: 10.1145/3591131. URL `https://doi-org.ezproxy.uzh.ch/10.1145/3591131`.

William Falcon and The PyTorch Lightning team. Pytorch lightning, August 2024. URL `https://doi.org/10.5281/zenodo.13254264`.

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

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

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

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `https://ieeexplore.ieee.org/abstract/document/6795963`. Conference Name: Neural Computation.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 138–146, Marseille, France, May 2020. European Language Resources Association. URL `https://aclanthology.org/2020.lrec-1.18`.

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. Patterns of text readability in human and predicted eye movements. In Michael Zock, Emmanuele Chersoni, Yu-Yin Hsu, and Enrico Santus (eds.), *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pp. 1–15, Taipei, Taiwan, November 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.cogalex-1.1`.

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

Lena A Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler, and Tobias Scheffer. 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*, pp. 299–314. Springer, 2020.

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

Johanna Kaakinen and Jukka Hyönä. Task Effects on Eye Movements During Reading. *Journal of experimental psychology. Learning, memory, and cognition*, 36:1561–6, November 2010. doi: 10.1037/a0020693.

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

Alan Kennedy, Robin Hill, and Joël Pynte. The dundee corpus. In *European conference on eye movement*, 2003.

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

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

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

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, September 2018. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. Bridging Information-Seeking Human Gaze and Machine Reading Comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 142–152, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.11. URL https://www.aclweb.org/anthology/2020.conll-1.11.

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

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nzpLWnVAyah.

Diane C. Mézière, Lili Yu, Erik D. Reichle, Titus von der Malsburg, and Genevieve McArthur. Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 58 (3):425–449, 2023. doi: 10.1002/rrq.498. URL https://ila.onlinelibrary.wiley.com/doi/abs/10.1002/rrq.498.

Ralph Radach and Alan Kennedy. 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, 2004.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.214. URL https://aclanthology.org/2020.acl-main.214.

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

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

Keith Rayner, Jane Ashby, Alexander Pollatsek, and Erik D Reichle. 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, 2004.

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

David Robert Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A Jäger. 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*, pp. 1–8, 2022.

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

Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 8(4):e1249, 2018.

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

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

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

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, 54(6):2843–2863, 2022.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. Quantifying sentence complexity based on eye-tracking measures. In Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Thomas François, and Philippe Blache (eds.), *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pp. 202–212, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL `https://aclanthology.org/W16-4123`.

Takenobu Tokunaga, Hitoshi Nishikawa, and Tomoya Iwakura. An eye-tracking study of named entity annotation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pp. 758–764, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_097. URL `https://doi.org/10.26615/978-954-452-049-6_097`.

Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jürgen Ziegler. A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1158–1167, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `https://aclanthology.org/P10-1118`.

Sowmya Vajjala and Ivana Lučić. 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*, pp. 297–304, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0535. URL `https://aclanthology.org/W18-0535`.

Duo Yang and Nora Hollenstein. Plm-as: Pre-trained language models augmented with scanpaths for sentiment classification. *Proceedings of the Northern Lights Deep Learning Workshop*, 2023. URL `https://api.semanticscholar.org/CorpusID:256262367`.

APPENDIX

A   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 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 2: 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).

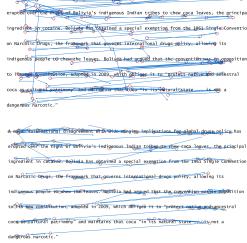|	
Advanced	A major international disagreement with wide-ranging implications for global drugs policy has erupted over the right of Bolivia's indigenous Indian tribes to chew coca leaves, the principal ingredient in cocaine. **Bolivia has obtained a special exemption from the 1961 Single Convention on 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 Answers	**What was the purpose of the 1961 Convention on Drugs?** 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 5: Examples of eye movements over a single passage; left: ordinary reading, right: information seeking. Circles represent fixations, and lines represent saccades.

B   MODEL TRAINING AND HYPERPARAMETERS

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., 2022), 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 & 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 & 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 & Hutter (2018) with a batch size of 16. MAG-Eye, RoBERTa-Eye 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. Models are implemented based on code from (Shubi et al., 2024).

C   ADDITIONAL RESULTS

Below we present test set ROC curves and AUC. We also provide accuracy results for the validation set.
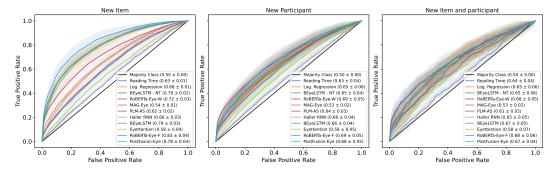

Figure 6: ROC Curves by model and evaluation regime. Each curve represents a different model, with the corresponding Area Under the Curve (AUC) values provided in the legend.

Table 3: Validation accuracy results aggregated across 10 cross-validation folds. 'Emb' stands for word embeddings, 'Ling. Feat.' for linguistic word properties. 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 '***' $p < 0.001$.

Model	Gaze Representation	Text Representation	New Item	New Participant	New Item & Participant	All
Majority Class / Chance	None	None	50.0	50.0	50.0	50.0
Reading Time	None	None	58.9	58.9	60.4	58.9
Logistic Regression	Global	None	62.6**	60.6	61.0	61.6*
BiEyeLSTM - No Text	Fixations	None	73.2***	64.9	65.1	68.8***
RoBERTa-Eye-W	Words	Emb + Ling. Feat.	66.3***	**65.4****	64.1	65.8***
MAG-Eye	Words	Emb + Ling. Feat.	53.7	53.5	52.6	53.5
PLM-AS	Fixations Order	Emb	59.1	61.1	58.6	60.1
Haller RNN	Fixations	Emb	62.5	62.4	61.4	62.3
BEyeLSTM	Fixations	Ling. Feat.	72.3***	65.0	66.1	68.5***
Eyettention	Fixations	Emb + Ling. Feat.	56.4	56.6	58.6	56.6
RoBERTa-Eye-F	Fixations	Emb + Ling. Feat.	**76.5*****	64.8*	**65.9**	**70.3*****
PostFusion-Eye	Fixations	Emb + Ling. Feat.	72.8***	64.6*	63.6	68.4***

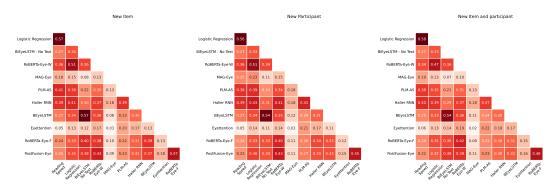# D   PAIRWISE AGREEMENT BETWEEN MODELS BY EVALUATION REGIME

Figure 7: Pairwise Cohen's Kappa agreement between model predictions on the validation set by evaluation regime.