Déjà Vu? Decoding Repeated Reading from Eye Movements

Yoav Meiri, Omer Shubi, Cfir Avraham Hadar,
Ariel Kreisberg Nitzav and Yevgeni Berzak
Faculty of Data and Decision Sciences,
Technion - Israel Institute of Technology, Haifa, Israel
{meiri.yoav,shubi,kfir-hadar,ariel.kr}@campus.technion.ac.il
berzak@technion.ac.il

Abstract

Be it your favorite novel, a newswire article, a cooking recipe or an academic paper – in many daily situations we read the same text more than once. In this work, we ask whether it is possible to automatically determine whether the reader has previously encountered a text based on their eye movement patterns during reading. We introduce two variants of this task and address them using both feature-based and neural models. We further introduce a general strategy for enhancing these models with machine generated simulations of eye movements from a cognitive model. Finally, we present an analysis of model performance which on the one hand yields insights on the information used by the models, and on the other hand leverages predictive modeling as an analytic tool for better characterization of the role of memory in repeated reading. Our work advances the understanding of the extent and manner in which eye movements in reading capture memory effects from prior text exposure, and paves the way for future applications that involve predictive modeling of repeated reading.[1]

1 Introduction

Reading is a widely practiced skill that occupies many hours of our daily lives. During these hours, there are various ways in which we interact with texts. While reading is often thought of as an interaction with new linguistic material, in many daily scenarios we read texts more than once. This can happen in the framework of educational curricula that involve repeated reading, because we want to understand or recall the text better, re-examine specific parts of interest, or simply because we enjoyed reading the text.

The importance of studying repeated reading for understanding human language processing has long been recognized in psychology and psycholinguistics. In these areas of study, it was shown that when reading a text for a second time, the way our eyes move over the text and the extent to which the eye movements depend on the linguistic characteristics of the text tend to differ compared to the first reading. In essence, eye movements in repeated reading reflect reading facilitation: for example, readers tend to read faster and skip more words compared to the first reading (Hyönä and Niemi, 1990; Raney and Rayner, 1995; Schnitzer and Kowler, 2006; Meiri and Berzak, 2024). Although the precise differences can depend on the experimental setup, and some are debated, the presence of facilitation effects comes as no surprise, as when encountering a text for the second time readers already have knowledge of its content, and can more easily foresee what comes next at any given time.

Despite the advances in the study of eye movements in repeated reading, prior work has been limited to *descriptive* analyses of overall effects, averaged across texts and participants. Consequently, it is currently unknown how much information can be extracted regarding the type of interaction of a specific reader with a specific text. Addressing this question is important both for improving the scientific understanding of the extent and manner in which eye movements reflect the reader's memory of the text, and for building the foundations for practical applications in areas such as e-learning and educational settings more broadly, where it can be beneficial to infer in real time whether the reader has already encountered the text.

In this work, we tackle this challenge using a *predictive* modeling approach for determining the interaction of a single reader with a specific text from their eye movements. We pose the following question: is it possible to decode whether the reader is reading a text for the first or the second time from their eye movements over the text? Ad-

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

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19460–19482
July 27 - August 1, 2025

dressing this question is made possible by OneStop Eye Movements (Berzak et al., 2025), the first publicly available dataset that contains eye movement recordings of both first and repeated reading.

We operationalize this question via a prediction task in two variants. In the first variant, the task is to determine whether a single eye movement sample is a first or a repeated reading. In the second, less challenging variant, given two eye movement samples from the same participant over the same text, the goal is to determine which is a first reading and which is a repeated reading. We address these tasks with feature-based models and with multimodal neural models that combine eye movements with text, and further augment the models with synthetic eye movement trajectories from a cognitive model of eye movements in reading. Finally, we demonstrate how such predictive models can be used to obtain insights on the determinants of eye movements in repeated reading.

The contributions of this work are the following:

- **Tasks**: We introduce a new prediction task - automatically determine whether the reader has previously encountered a text, based on their eye movements during reading. We address this task in two variants of decreasing difficulty: (i) a single eye movement sample; (ii) a pair of first and second reading samples for the same text, from the same participant.
- **Modeling**: We experiment with two types of predictive approaches: (i) feature-based models; (ii) neural multimodal language models. We further introduce a strategy for integrating into the prediction pipeline synthetic data for first reading.
- **Analyses**: We present analyses of model performance as a function of article location in the experiment and the amount of intervening material between readings. These analyses provide insights on the information used by models and on the role of memory in repeated reading effects.

2 Related Work

When we read, the eye movement trajectory, or *scanpath* over the text is divided into *fixations*, prolonged periods of time during which the gaze location is relatively fixed, and *saccades*, fast transitions between fixations (Rayner, 1998; Hyönä and Kaakinen, 2019; Schotter and Dillon, 2025). Prior work in psycholinguistics has consistently demonstrated that this trajectory differs in repeated reading compared to first reading, with large *facilitation* effects marked by shorter text reading times, fewer fixations, shorter fixation durations, longer saccades and fewer regressions (backward saccades) (Hyönä and Niemi, 1990; Raney and Rayner, 1995; Schnitzer and Kowler, 2006; Meiri and Berzak, 2024). Several studies have also examined the interaction between repeated reading and the effect of linguistic word properties such as word length, frequency and surprisal on reading times, mostly finding less sensitivity to word properties in repeated reading (Raney and Rayner, 1995; Foster et al., 2013; Zawoyski et al., 2015; Meiri and Berzak, 2024). In line with these studies, Hyönä and Niemi (1990) further demonstrated a reduction in the sensitivity of eye movements to the introduction of new topics in repeated reading. All the above studies examined individual features aggregated across participants and texts, and it is currently unknown whether first and repeated reading can be effectively distinguished using predictive modeling at the level of an individual participant and a single text.

In machine learning and NLP, a nascent line of work focuses on decoding properties of the reader and their interaction with the text, from eye movements in reading. These include, among others, decoding of linguistic knowledge (Berzak et al., 2017, 2018; Skerath et al., 2023), reading comprehension (Ahn et al., 2020; Reich et al., 2022; Mézière et al., 2023; Shubi et al., 2024b), subjective text difficulty (Reich et al., 2022) and the reader's goals (Hollenstein et al., 2023; Shubi et al., 2024a). The current study falls broadly within this area, but introduces and addresses a new task of decoding repeated reading. Following Sood et al. (2020), our work leverages the output of E-Z Reader, a computational cognitive model for automatic generation of reading scanpath trajectories (Reichle et al., 1998, 2003, 2009; Veldre et al., 2023).

3 Problem Formulation

We ask whether it is possible to accurately distinguish between first and second readings from an eye movement recording of a single participant over a single textual item. We assume a setup in which a participant S reads a textual item W, optionally reads k other items $\{W'\}^k$, and then reads W again. The parameter k can range from 0 for consecutive repeated reading to any $k > 0$ for non-consecutive repeated reading[2]. Hence, a reading $r \in \{1, 2\}$ (first or repeated) of W has a distinct eye movement recording $E_S^{W,r}$. We define a *decoding task* where the goal is to distinguish between eye movements of a single participant over a single text in first reading ($r = 1$) and repeated reading ($r = 2$). The task has two variants:

Single Trial Task

$$(W,\ E_S^{W,r}) \longrightarrow \hat{r}$$

In this task, the input is an eye movement recording $E_S^{W,r}$ for text W, and optionally the text itself. The output $\hat{r} \in \{1, 2\}$ corresponds to whether the eye movements E are from a first or a repeated reading of W.

Paired Trials Task

$$(W,\ E_S^{W,r},\ E_S^{W,r'}) \longrightarrow (\widehat{r, r'})$$

Here, the input consists of two eye movement recordings $E_S^{W,r}$ and $E_S^{W,r'}$ of the same participant S for the same text W in an unknown presentation order, and optionally the text itself. The output is $(\widehat{r, r'}) \in \{(1, 2), (2, 1)\}$, i.e., which recording corresponds to the first reading of W, and which to the second.

4 Data

We use OneStop Eye Movements (Berzak et al., 2025), an eye tracking dataset collected with an Eyelink 1000 Plus eye tracker, where native (L1) speakers of English read Guardian newswire articles in English. The textual materials are taken from the OneStopQA dataset (Berzak et al., 2020). OneStop Eye Movements includes 180 participants who read for comprehension, each reading a 10-article batch in a randomized article order, where each article contains between 4 and 7 paragraphs. Participants read each paragraph on a single page, and then answer a reading comprehension question about the paragraph on a new page, without the ability to return to previous pages.

After reading a 10-article batch, participants read two articles for a second time. In repeated reading, the paragraphs are identical to the first reading, while the questions are different. The article in position 11 is a consecutive second presentation of the article in position 10. The article in position 12 is a non-consecutive second presentation of an article in one of the positions 1-9. Thus, half of the repeated reading data captures immediate consecutive rereading of the same article, and the other half is rereading with intervening reading material, ranging from 2 to 10 articles. Figure 1 shows example trials for first reading and repeated reading. Figure 2 presents the experimental design schematically.

Overall, there are 360 second presentations of articles, 180 in consecutive rereading in position 11

Figure 1: Example of eye movements from the same participants for a single passage; top: first reading, bottom: repeated reading. Circles represent fixations, and lines represent saccades.

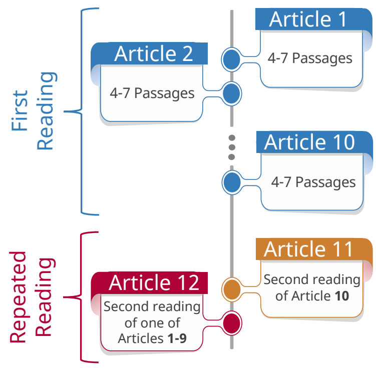

Figure 2: First and repeated reading for one participant. After reading a 10-article batch in a random order of articles, there is a consecutive repeated reading of the last article in position 10, and then a non-consecutive repeated reading of one of the articles in positions 1-9.

[2]In this work, $k = 0$ or $2 \leq k \leq 10$ (see Section 4).

and 180 in a non-consecutive rereading in position 12. The first reading of position 12 articles occurs 36 times in position 1 and 18 times in each of the positions 2-9. The 360 repeated article readings correspond to 1,944 paragraph trials with a total of 105,540 word tokens over which eye movements were collected, split equally between positions 11 and 12.

5 Modeling

We experiment with both feature-based and neural language modeling approaches for representing eye movements and their interaction with the text. For the single trial variant, we present methods that directly predict first versus repeated reading. We also propose a general approach for leveraging synthetic scanpaths as an additional first reading reference, which enables using the same input representations as in the paired trials task.

5.1 Single Trial Modeling

Feature-Based Model

For feature-based modeling, we use XGBoost tree-boosting models (Chen and Guestrin, 2016) with the following global eye movement features, resulting in an input feature vector $e_S^{\text{global}} \in \mathbb{R}^{d_{\text{global}}}$ where $d_{global} = 36$. The features are motivated by the psycholinguistic literature in general, and work on differences between first and repeated reading in particular.

- **Standard Eye Movement Measures** 9 trial-level features from the psycholinguistic literature: Skip Rate, Reading Speed, Regression Rate per word, proportion of refixated words, and per-word means of Total Fixation Duration, First Fixation Duration, Gaze Duration, number of fixations and extra dwell time beyond the first run (zero if no refixations). See Appendix B for definitions. Part of these features have been used in prior prediction tasks (e.g. Mézière et al., 2023) and differ between first and repeated reading (see Section 2).

- **Word Property Coefficients** 20 features that measure the responsiveness of reading measures to linguistic word properties: frequency, surprisal and length. Building on Berzak et al. (2018), the features are coefficients from linear models that predict the participant's speed-normalized eye movement measures from these three word properties. This feature-set is motivated by prior work that has demonstrated that the responsiveness of eye movements to linguistic word properties varies across reading scenarios and readers (e.g. Reichle et al., 2010; Berzak and Levy, 2023; Shubi and Berzak, 2023). In repeated reading, this responsiveness is weaker compared to first reading (Meiri and Berzak, 2024). See Appendix B.1 for further details on the models.

- **Saccade Network Measures** Following Zhu and Feng (2015) and Ma et al. (2023), we define a directed graph that encodes the scanpath of eye movements over the paragraph $G = \{V, T\}$ such that V is the set of words in the paragraph, and for all $u, v \in V$:

$$T = \{(u, v) : \text{there is a saccade from } u \text{ to } v\}$$

  We extract 7 features which capture connectivity, centrality and clustering measures of this graph. Additional details and definitions of the measures, along with network visualization examples are provided in Appendix B.2.

Neural Models

We use two variants of the **RoBERTEye** multimodal language model (Shubi et al., 2024b), which is a state-of-the-art approach in predictive modeling using eye movements in reading. This model was previously applied to the prediction of reading comprehension (Shubi et al., 2024b) and reading goals (Shubi et al., 2024a) from eye movements, outperforming in most cases prior models from the literature. RoBERTEye extends the RoBERTa model (Liu et al., 2019) by incorporating eye movement information. It does so by projecting an input eye movement feature vector for each word or fixation into the embedding space of the language model, aligning these projections with their corresponding words, and then concatenating the projections with the word embedding sequence.

The model has two variants, with **word-level** and **fixation-level** eye movement representations. In **RoBERTEye-Words** the eye movements input consists of $\left(e_S^{\text{word}_j}\right)_{j=1}^{N_{\text{words}}}$ where each $e_S^{\text{word}_j} \in \mathbb{R}^{d_{\text{word}}}$ is an eye movement feature vector for the word j, with $d_{word} = 13$ features.

In **RoBERTEye-Fixations**, both fixation-level and word-level features are used. Each fixation i on word j has a fixation vector $e_S^{\text{fix}_{i,j}} \in \mathbb{R}^{d_{\text{fix}}}$ with $d_{fix} = 6$ features of the fixation. This vector is concatenated with the word-level feature vector $e_S^{\text{word}_j}$:

$$\left(e_S^{\text{fix}_{i,j}} \oplus e_S^{\text{word}_j}\right)_{i=1}^{N_{\text{fixations}}}$$

where $\oplus$ denotes the concatenation operation along the feature dimension. To help the model distinguish between eye movement and textual information, two special token vectors are added, one to all the text embeddings, and the other to all the projected eye movement embeddings. Descriptions of all the fixation-level and word-level features are provided in Appendix A.

5.2 Single Trial Modeling with Synthetic Scanpath References

We introduce a new method, where in addition to the human eye movement data, the model input further includes a *synthetic scanpath* reference $E_M^{W,1}$ of eye movements E generated for each text W from an external model M for scanpath generation. As all existing computational models for scanpath generation assume a first reading, in this work we focus on the generation of first reading reference scanpaths. In essence, this reference provides an external source of information on how a first reading of the text by an average reader should look like. This reference addition yields an input structure that resembles that of the paired trials task, only that one of the eye movement inputs is now machine generated, and its reading interaction is assumed to be first reading:

$$(W,\ E_M^{W,1}\ E_S^{W,r}) \longrightarrow \hat{r}.$$

We then obtain the following three types of feature representations of eye movements:

Global representations This representation is a concatenation of the human features, and the difference between the synthetic and the human features. Formally,

$$e_S^{\text{global}} \oplus (e_M^{\text{global}} - e_S^{\text{global}})$$

Word-level representations For each word, we concatenate the human features with the difference between machine-generated synthetic features and the human features. Formally,

$$\left(e_S^{\text{word}_j} \oplus (e_M^{\text{word}_j} - e_S^{\text{word}_j})\right)_{j=1}^{N_{words}}$$

Fixation-level representations Differently from global and word-level representations, in fixation-level feature representations the number of features for two different trials can differ, and there is no direct alignment between the representations. We therefore construct the input as follows:

$$\left(e_S^{\text{fix}_{i,j}} \oplus e_S^{\text{word}_j}\right)_{i=1}^{N_{\text{fixations}}} \parallel \left(e_M^{\text{fix}_{i,j}} \oplus e_M^{\text{word}_j}\right)_{i=1}^{\tilde{N}_{\text{fixations}}}$$

where $\parallel$ denotes concatenation along the sequence dimension, and $\tilde{N}_{\text{fixations}}$ is the length of the scanpath generated by M. To help RoBERTEye distinguish between human and synthetic scanpath features, in addition to the word and human eye movement special tokens, we introduce a third token that marks machine generated scanpaths.

E-Z Reader Scanpaths The synthetic scanpaths are generated using E-Z Reader (Reichle et al., 1998, 2003, 2009; Veldre et al., 2023), a prominent computational cognitive model for eye movements generation. The full details of the generation process and the adaptations made to the original model are described in Appendix C. As we expect the effectiveness of the augmentation approach to depend on the quality of the generated scanpaths, we perform an evaluation of E-Z Reader outputs in the context of our task. To this end, we compare E-Z Reader outputs with human eye movements in both first and repeated reading. Our analysis examines the overall similarity of the scanpaths, which we expect to be greater in first reading, as well as the direction of the deviations. A necessary condition for E-Z Reader outputs to be effective as approximations of first reading behavior, is that on average, they should be more similar to human first reading than to human repeated reading.

Measure	First Reading	Repeated Reading
Fixation Count	$+0.03_{\pm 0.01}$	$-0.31_{\pm 0.01}$
Mean TF (ms)	$+27_{\pm 2.3}$	$-29_{\pm 1.7}$
Regression Rate	$+0.2_{\pm 0.003}$	$+0.1_{\pm 0.002}$
Skip Rate	$+0.2_{\pm 0.003}$	$+0.3_{\pm 0.003}$

Table 1: Trial-level mean differences between human first/repeated reading and E-Z Reader-generated measures for four standard eye movement measures, with 95% confidence intervals.

Table 1 suggests that this indeed tends to be the case. It presents four measures for which robust differences between first and repeated reading were previously observed (Meiri and Berzak, 2024). Using mixed-effects models with text-level bootstrapping (see Appendix C.2), we compared the absolute differences of E-Z Reader from first and repeated reading. For Fixation Count and Skip Rate, E-Z Reader is significantly closer to first reading ($p < 0.001$), whereas Regression Rate is significantly closer to repeated reading ($p < 0.001$). Although mean Total Fixation Duration (TF) does not show a significant difference ($p \approx 0.53$), the direction of the difference favors first reading. Overall, these findings support the viability of E-Z Reader as an approximation of human first reading scanpath trajectories.

E-Z Reader is a probabilistic model that samples scanpaths for a given text. We generate 1000 synthetic scanpaths for each paragraph, and then augment the human eye movement data with reference features derived from these first reading simulated scanpaths. To obtain global and word-level representations, we first average measures across all the generated scanpaths. The averaging aims to enhance the robustness of the representations by reducing noise inherent to a single scanpath. For the fixation-level representation, averaging scanpaths is not applicable, and we therefore follow Mézière et al. (2024) in selecting a *prototype scanpath* that minimizes the mean scanpath distance to all other scanpaths, using the Scasim scanpath similarity metric (Von der Malsburg and Vasishth, 2011).

5.3 Paired Trials Modeling

In the paired trials task, we use the same feature representation as in the single trial task with machine generated scanpaths described above, only that now both eye movement samples are from a human participant. We shuffle the order of the two inputs such that it is randomized, and the output of the model is the probability of the repeated reading trial being second in order. Further, differently from the single-augmented setting, here the third special token in RoBERTEye marks the second human input.

5.4 Baselines

- **Majority Class**: the most frequent class in the training set. As our data is balanced, this baseline is equivalent to a random choice.
- **Reading Speed**: the number of words read per second. Note that when the text is available, this measure can be calculated from the total reading time of the trial, and therefore does not require eye tracking. Prior work has consistently shown that reading is faster in repeated reading compared to first reading (see Section 2). We therefore expect this to be a strong baseline. Crucially, it enables determining the added value of eye tracking information for our decoding tasks.

6 Experimental Setup

Evaluation Regimes

We use 10-fold cross validation with three evaluation regimes:

- **New Participant** eye movement data is available for the given paragraph, but no prior eye movement data was collected for the participant.
- **New Item** prior eye movement data is available for the participant, but not for the paragraph.
- **New Participant and Item** no prior data is available for the participant nor for the paragraph.
- **All** the union of the above three regimes.

Data Splits

To allow complete matching of participants across the first and repeated reading of each article, out of the 10 articles read by each participant during first reading, we use only the 2 articles that were read twice. Further, we leverage the counterbalancing properties of OneStop to obtain data splits that fulfill the following properties: 1) the three test regimes are balanced in number of participants 2) the three validation regimes are balanced to the extent possible in number of participants 3) there is an equal number of consecutive and non-consecutive repeated readings in each portion of the split.

We define a constrained combinatorial problem that has an algorithmic solution that satisfies these constraints. We provide further details on the solution in Appendix D. All resulting splits satisfy that the training set has 264 participant-article pairs, the validation set has 48 pairs, and the test set has 54 pairs, where each test regime has exactly 18 pairs, all balanced with respect to consecutive and non-consecutive repeated reading. In Figure 3 we present an example of one split.

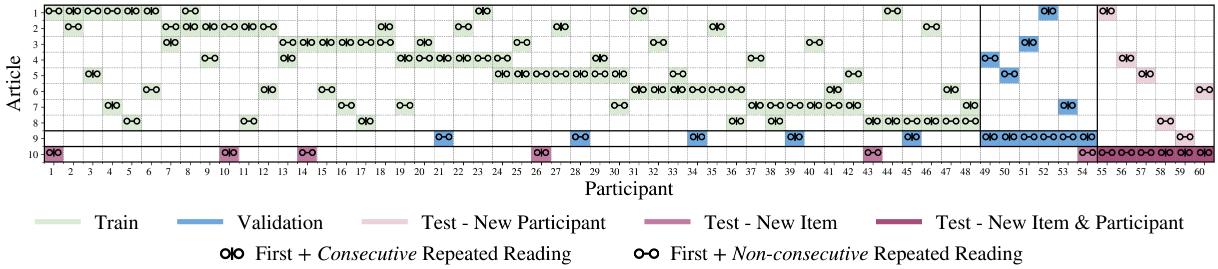

Figure 3: Visualization of a 10-article, 60-participant data split, divided into train, validation, and three test regimes. Each non-empty cell represents a participant-article pair, comprising the first and repeated readings of an article by the same participant. '⚯' denotes consecutive repeated reading and '⚬-⚬' denotes non-consecutive repeated reading (i.e. with intervening articles between the first and second readings).

Model Training and Selection

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 library (Falcon and The PyTorch Lightning team, 2024). Further details on the training procedure, including the full hyperparameter search space for all models are provided in Appendices E and F.

7 Results

Below, we summarize our main experimental findings for both the paired and single-trial variants of the task. Table 2 presents the quantitative results on classification accuracy.

Single Trial In this task all models outperform the reading speed baseline, demonstrating the added value of eye movement information. The highest All Accuracy of 70.2 is achieved with the XGBoost model augmented with E-Z Reader Scanpaths. However, it does not outperform the other models statistically, and different models come first on different evaluations, again, with no statistically significant differences from the other models. Within each model, performance is relatively stable across the three evaluation regimes.

Single Trial with Synthetic References While the best performing model in the 'All' evaluation regime includes a synthetic scanpath augmentation, the gains over the non-augmented model counterparts are not statistically significant and not consistent within models and evaluation regimes.

Paired Trials In the paired trials setup, the model's output is an ordering of two trials. To make the evaluation of these predictions comparable to the single-trial task, we "unaggregate" the model's predictions so that predicting the correct order counts as two correct single-trial classifications (and vice versa for an incorrect prediction). The reading speed baseline achieves a high All Accuracy of 87.7. While the neural models exhibit baseline-level results, XGBoost substantially outperforms the baseline and the neural models in all the evaluation regimes, reaching an overall Accuracy of 91.2.

In Appendix G we present validation and test results for three complementary evaluation measures: Precision, Recall, and F1. Across all three measures and evaluation regimes, the best-performing model in the single trial evaluations tends to be one of the E-Z Reader augmented models.

8 Fine-Grained Analysis of Model Performance

The controlled experimental design of OneStop enables going beyond aggregated model performance evaluation metrics and analyzing model behavior as a function of trial characteristics. Prior work with OneStop observed that individual eye movement measures in first and repeated reading vary systematically across different item and participant characteristics (Meiri and Berzak, 2024). Here, we analyze this variability through model classification performance. This enables a detailed characterization of model behavior across data characteristics, and further leverages models as analytic tools for data analysis.

We focus on the more challenging single trial task, and analyze the assigned probabilities and prediction accuracy of the best performing model, XGBoost augmented with E-Z Reader scanpaths. Figure 4 presents the mean probability assigned

Task Variant	Model	Eye Movements Input	New Item Seen Participant	New Participant Seen Item	New Item & Participant	All	
Single Trial	Majority Class	-	50.0	50.0	50.0	50.0	-
	Reading Speed	$E_S^{W,r}$	$66.9_{\pm 2.0}$	$67.1_{\pm 2.1}$	$66.8_{\pm 2.1}$	$66.9_{\pm 1.2}$	-
		$E_{EZ}^{W,1}, E_S^{W,r}$	$67.3_{\pm 2.2}$	$67.3_{\pm 2.0}$	$66.5_{\pm 2.1}$	$67.1_{\pm 1.2}$	n.s
	XGBoost	$E_S^{W,r}$	$69.5_{\pm 2.0}$	$70.7_{\pm 2.0}$	$68.7_{\pm 2.0}$	$69.6_{\pm 1.2}$	***
		$E_{EZ}^{W,1}, E_S^{W,r}$	$70.1_{\pm 2.1}$	$\mathbf{71.2}_{\pm \mathbf{1.9}}$	$69.3_{\pm 2.0}$	$\mathbf{70.2}_{\pm \mathbf{1.2}}$	***
	RoBERTEye-Fixations	$E_S^{W,r}$	$\mathbf{70.7}_{\pm \mathbf{2.1}}$	$69.4_{\pm 2.0}$	$\mathbf{70.3}_{\pm \mathbf{2.0}}$	$70.1_{\pm 1.1}$	**
		$E_{EZ}^{W,1}, E_S^{W,r}$	$69.9_{\pm 2.0}$	$\mathbf{71.2}_{\pm \mathbf{2.0}}$	$69.4_{\pm 2.1}$	$70.1_{\pm 1.2}$	***
	RoBERTEye-Words	$E_S^{W,r}$	$69.1_{\pm 2.0}$	$71.0_{\pm 2.0}$	$69.0_{\pm 2.1}$	$69.7_{\pm 1.2}$	*
		$E_{EZ}^{W,1}, E_S^{W,r}$	$70.0_{\pm 2.0}$	$70.4_{\pm 2.1}$	$70.0_{\pm 2.1}$	$70.1_{\pm 1.2}$	***
Paired Trials	Majority Class	-	50.0	50.0	50.0	50.0	-
	Reading Speed	$E_S^{W,r}, E_S^{W,r'}$	$88.0_{\pm 2.0}$	$88.1_{\pm 2.1}$	$87.2_{\pm 2.1}$	$87.7_{\pm 1.2}$	-
	XGBoost		$\mathbf{91.5}_{\pm \mathbf{1.7}}$	$\mathbf{92.2}_{\pm \mathbf{1.7}}$	$\mathbf{90.5}_{\pm \mathbf{1.8}}$	$\mathbf{91.2}_{\pm \mathbf{1.0}}$	***
	RoBERTEye-Fixations		$88.8_{\pm 1.9}$	$87.5_{\pm 2.0}$	$87.3_{\pm 2.2}$	$87.9_{\pm 1.2}$	n.s
	RoBERTEye-Words		$90.1_{\pm 1.8}$	$89.9_{\pm 1.9}$	$89.9_{\pm 1.5}$	$89.5_{\pm 1.1}$	**

Table 2: Test accuracy aggregated across 10 cross-validation splits, with 95% confidence intervals. $E_S^{W,r}$ and $E_{EZ}^{W,1}$ are human and E-Z Reader-synthesized eye movements respectively. Differences in performance across models are tested using a linear mixed effects model. In R notation: $is\_correct \sim model + (model \mid participant) + (model \mid paragraph)$. Significant gains over the reading speed baseline in the All regime are marked with '*' $p < 0.05$, '**' $p < 0.01$ and '***' $p < 0.001$. Within each task and evaluation regime, the best-performing model is in bold.

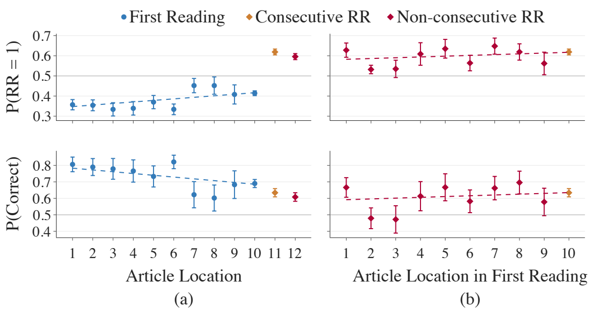

Figure 4: Analysis of the E-Z Reader-augmented XGBoost model's behavior as a function of item position in the experiment. Depicted are mean probability assignment for repeated reading (RR) (top) and mean classification accuracy (bottom) with 95% confidence intervals. (a) First and repeated reading as a function of article position in the experiment. (b) Repeated reading as a function of the article position in the first reading. Dashed lines are predictions obtained from linear mixed-effects models[3].

to trials being a repeated reading trial (top) and the mean probability of classifying trials correctly (bottom) as a function of the position of the article in the experiment, see Figure 2.

In **first reading** trials (blue) in Figure 4 (a), as the experiment progresses, the model exhibits decreasing confidence in classifying trials as first reading ($p < 10^{-8}$) and goes down in prediction accuracy ($p \approx 0.001$). This outcome mirrors a decrease in reading times during first reading as the experiment progresses observed in Meiri and Berzak (2024). When adding reading speed as a predictor in models that predict the probability assignments for repeated reading and accuracy[3], we find that for both the effect of position is no longer significant ($p \approx 0.7$ for probabilities and $p \approx 0.5$ for accuracy). This suggests that the model heavily relies on reading speed or correlates thereof. Consequently, as reading speed increases during the experimental session and comes closer to the reading speed in repeated reading, the model becomes worse at correctly classifying first reading items.

In **repeated reading** trials in Figure 4 (a), the model assigns higher probabilities to consecutive repeated readings compared to non-consecutive ones ($p \approx 0.02$). Accuracy is also numerically lower compared to consecutive repeated readings, but the difference is not significant ($p \approx 0.16$). These outcomes are again in line with the analysis of Meiri and Berzak (2024) who found lower

[3]In R notation: $outcome \sim fixed\_terms + (fixed\_terms|subject) + (fixed\_terms|paragraph)$ where $outcome \in \{P(RR = 1), P(Correct)\}$, and $fixed\_terms \in \{position, position + reading\_speed\}$.

reading times and less skipping in non-consecutive versus consecutive reading. Taken together with these results, our analysis strengthens Meiri and Berzak (2024) interpretation that reading facilitation is greater in consecutive reading, potentially due to better memory retention of the first reading.

In Figure 4 (b) we examine repeated reading item probabilities and accuracy as a function of the position of the first reading. Reflecting again reading speed and other single measure analyses in Meiri and Berzak (2024), we find no evidence for a better model's classification accuracy with fewer articles between the two readings ($p \approx 0.16$). However, we do find an effect in model probabilities for repeated reading which increases with article position ($p \approx 0.03$), hinting at a favorable effect of the recency of the first reading. This suggests that model performance analysis can unveil more fine grained data patterns than traditional analysis of individual measures.

9 Conclusion and Discussion

Our study presents the first attempt to decode the number of prior interactions of a reader with a text from their eye movements. We demonstrate that it is feasible to perform this task, with various degrees of success depending on the difficulty of the task variant. In addition, we propose and experiment with a general method for leveraging synthetic ordinary reading data to improve predictive modeling of non-ordinary reading. Overall, the results indicate that there is an informative signal for the presence or absence of a prior text interaction in eye movements at the level of a single paragraph and a single reader. This signal tends to be better captured by feature-based models than by neural language models. Our work extends prior literature on predictive modeling from eye movements in reading, providing further evidence for the informativeness of eye movements regarding the reader's cognitive state during online language processing.

Our findings also open the door for practical, user facing applications. For example, in education, repeated reading is used as a pedagogical technique during the acquisition of reading skills in children, and a large body of work supports its effectiveness in improving reading fluency (Meyer and Felton, 1999; Faulkner and Levy, 1999; Teigen et al., 2001; Kuhn, 2004; Ardoin et al., 2008, among others). Varying levels of certainty in automatic identification of prior exposure to a text could enhance educational and e-learning platforms which involve repeated reading by serving as an indicator of information uptake during the initial reading. It can further facilitate special assistance to individuals and populations that struggle with reading comprehension, which can potentially be diagnosed via repeated reading. Another possible use case is real-time content adaptation, where online content delivery platforms, such as news outlets, could personalize recommendations by detecting a reader's familiarity with the current content and suggesting novel material. While these examples are hypothetical, they illustrate the broader potential for reader–text interaction modeling to support user goals in naturalistic reading environments.

In addition to practical applications, we believe that the ability to predict repeated reading using modern day machine learning models, and using them as analysis tools for understanding reading behavior during repeated reading will pave the way to further scientific advances in psycholinguistics and the psychology of reading. Such advances will include more comprehensive accounts of reading interactions in daily life, as well as new possibilities for studying the similarities and differences between human and machine language processing.

10 Limitations

Our work has a number of limitations that are related to the experimental design and the eye tracking data. Consecutive repeated reading occurs at the level of a full article, such that there are minimally 3 intervening paragraphs between two readings of the same paragraph. This setup does not address immediate repeated reading that involves working memory. The maximal amount of intervening material between two readings of the same article is 10 articles, leaving out larger time intervals between the readings. Repeated reading always occurs after participants have already read 10 articles, and not at earlier stages of the experiment, which limits the generality of the results. The experiment is also restricted to two readings of any given text, while in daily life the same text can be read more than twice. Further, repeated reading is always of the exact same text, leaving out reading of texts that are similar but not identical in content to previously read material. Finally, the underlying texts are all newswire articles, and while they include a wide range of topics, other textual domains are not covered. We intend to collect data and investigate both shorter and longer repetition intervals, repeated reading at earlier stages of the experiment, repeated reading of paraphrased texts, multiple repeated readings, as well as additional textual domains in future work.

In the current work we use E-Z Reader for scanpath augmentation. We use this model due to its prominence in the psycholinguistic literature, as a proof-of-concept for the viability of generating first reading scanpaths for our task. However, there are other cognitive models of scanpath generation, such as SWIFT (Engbert et al., 2005), and in recent years, neural-net based models of scanpath generation have been introduced (Deng et al., 2023; Bolliger et al., 2025). In future work, we plan to examine the potential of such models for improving generation quality and downstream classification performance for our task.

An additional limitation is the experimental procedure, where reading occurs in-lab, and the presence of a reading comprehension question after each paragraph. Both aspects can negatively affect the ecological validity of the data and lead to reading behavior that is not fully representative of everyday life. Relatedly, while we use the term ordinary reading to refer to reading for general text comprehension, we acknowledge that this term, and similar terms such as "normal reading", are not without faults (Huettig and Ferreira, 2022).

Further limitations concern the participants, the experiment language and the equipment used. Although OneStop (Berzak et al., 2025) is the first public dataset that enables studying repeated reading, it is restricted to adult L1 speakers, with no cognitive impairments, and mostly with no eye conditions. This pool of participants does not cover multiple populations, including L2 speakers, children, elderly, participants with cognitive and physical impairments and others. Moreover, the eye tracking data and modeling work is restricted to English. These factors limit the scope and the generality of the results. Both data collection and model development work is required to include additional languages and populations. Finally, our approach has currently only been tested using a state-of-the-art eye tracker (Eyelink 1000 Plus) at a sampling rate of 1000Hz. This eye tracker allows extracting gaze position and duration at a very high temporal resolution and character-level precision. Such equipment is generally not available for end users, limiting the application potential of the current work. The feasibility of using lower spatial and temporal resolution eye tracking systems, as well as standard front-facing cameras on devices such as laptops, tablets and phones should be examined in future work.

11 Ethical Considerations

The eye tracking data used in our experiments was collected under an institutional IRB protocol (Berzak et al., 2025). All the participants provided written consent prior to taking part in the eye tracking experiment and received monetary compensation for their participation. The dataset is anonymized. Analyses and modeling of eye movements in repeated reading are among the main use cases for which the data was collected.

As mentioned above, decoding of repeated reading can be valuable in applications for monitoring reading acquisition, reading comprehension and retention of learned material. However, such technologies also pose a potential for inaccurate predictions and biases that may put various individuals and populations at a disadvantage. These include L2 learners, participants with cognitive impairments, participants with eye conditions and others. Additional data collection, modeling and analysis work for these groups is required before considering the deployment of such technology.

Finally, it is important to consider the issues of privacy and consent in the scope of eye tracking technologies. It was previously shown that eye movements contain information that 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 point out the importance of not storing information that could enable participant identification in future studies on repeated reading and other reading regimes. We further stress that future systems that perform prediction of repeated reading 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

Seoyoung Ahn, Conor Kelton, Aruna Balasubramanian, and Greg Zelinsky. 2020. Towards Predicting Reading Comprehension From Gaze Behavior. In *ACM Symposium on Eye Tracking Research and Applica-*

tions, ETRA '20 Short Papers, pages 1–5, New York, NY, USA. Association for Computing Machinery.

Phillip M. Alday and Douglas Bates. 2025. Mixedmodels.jl.

Scott P Ardoin, Tanya L Eckert, and Carolyn AS Cole. 2008. Promoting generalization of reading: A comparison of two fluency-based interventions for improving general education student's oral reading rate. *Journal of Behavioral Education*, 17:237–252.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models usinglme4. *J. Stat. Softw.*, 67(1).

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, Boris Katz, and Roger Levy. 2018. Assessing Language Proficiency from Eye Movements in Reading. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1986–1996, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yevgeni Berzak and Roger Levy. 2023. Eye movement traces of linguistic knowledge in native and non-native reading. *Open Mind*, 7:179–196.

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*.

Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting Native Language from Gaze. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 541–551, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lena S. Bolliger, David Robert Reich, and Lena Ann Jäger. 2025. Scandl 2.0: A generative model of eye movements in reading synthesizing scanpaths and fixation durations. *Proceedings of the ACM on Human-Computer Interaction*, 9:1 – 29.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Anthony Christopher Davison and David Victor Hinkley. 1997. *Bootstrap methods and their application*. 1. Cambridge university press.

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. *Proceedings of the ACM on Human-Computer Interaction*, 7(ETRA):1–24.

Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. 2005. SWIFT: a dynamical model of saccade generation during reading. *Psychological review*, 112(4):777.

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

Heather J Faulkner and Betty Ann Levy. 1999. Fluent and nonfluent forms of transfer in reading: Words and their message. *Psychonomic Bulletin & Review*, 6:111–116.

Tori E Foster, Scott P Ardoin, and Katherine S Binder. 2013. Underlying changes in repeated reading: An eye movement study. *School Psychology Review*, 42(2):140–156.

H2O.ai. 2022. *h2o: Python Interface for H2O*. Python package version 3.42.0.2.

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*.

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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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.

Jukka Hyönä and Pekka Niemi. 1990. Eye movements during repeated reading of a text. *Acta psychologica*, 73(3):259–280.

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.

Melanie Kuhn. 2004. Helping students become accurate, expressive readers: Fluency instruction for small groups. *The Reading Teacher*, 58(4):338–344.

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*.

Xiaochuan Ma, Yikang Liu, Roy Clariana, Chanyuan Gu, and Ping Li. 2023. From eye movements to scanpath networks: A method for studying individual differences in expository text reading. *Behavior research methods*, 55(2):730–750.

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.

Marianne S Meyer and Rebecca H Felton. 1999. Repeated reading to enhance fluency: Old approaches and new directions. *Annals of dyslexia*, 49:283–306.

Diane C Mézière, Lili Yu, Genevieve McArthur, Erik D Reichle, and Titus von der Malsburg. 2024. Scanpath regularity as an index of reading comprehension. *Scientific Studies of Reading*, 28(1):79–100.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. ArXiv:2006.04884 [cs, stat].

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.

Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. 2022. TorchMetrics - Measuring Reproducibility in PyTorch.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Gary E Raney and Keith Rayner. 1995. Word frequency effects and eye movements during two readings of a text. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 49(2):151.

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

David Robert Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A. Jäger. 2022. 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.

Erik D. Reichle, Simon P. Liversedge, Denis Drieghe, Hazel I. Blythe, Holly S.S.L. Joseph, Sarah J. White, and Keith Rayner. 2013. Using E-Z Reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review*, 33(2):110–149.

Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125.

Erik D Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476.

Erik D. Reichle, Andrew E. Reineberg, and Jonathan W. Schooler. 2010. Eye movements during mindless reading. *Psychological Science*, 21(9):1300–1310. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

Erik D Reichle, Tessa Warren, and Kerry McConnell. 2009. Using ez reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic bulletin & review*, 16:1–21.

Hildur EH Schilling, Keith Rayner, and James I Chumbley. 1998. Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & cognition*, 26(6):1270–1281.

Brian S Schnitzer and Eileen Kowler. 2006. Eye movements during multiple readings of the same text. *Vision research*, 46(10):1611–1632.

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.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Omer Shubi and Yevgeni Berzak. 2023. Eye movements in information-seeking reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Omer Shubi, Cfir Avraham Hadar, and Yevgeni Berzak. 2024a. Decoding reading goals from eye movements. *arXiv preprint arXiv:2410.20779*.

Omer Shubi, Yoav Meiri, Cfir Avraham Hadar, and Yevgeni Berzak. 2024b. Fine-grained prediction of reading comprehension from eye movements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3372–3391, Miami, Florida, USA. Association for Computational Linguistics.

Lina Skerath, Paulina Toborek, Anita Zielińska, Maria Barrett, and Rob Van Der Goot. 2023. Native language prediction from gaze: a reproducibility study. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 152–159.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33:6327–6341.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Tana Teigen, Paul R Malanga, and William J Sweeney. 2001. Combining repeated readings and error correction to improve reading fluency. *Journal of Precision Teaching and Celeration*, 17(2):58–67.

Aaron Veldre, Erik D Reichle, Lili Yu, and Sally Andrews. 2023. Understanding the visual constraints on lexical processing: New empirical and simulation results. *Journal of Experimental Psychology: General*, 152(3):693.

Titus Von der Malsburg and Shravan Vasishth. 2011. What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, 65(2):109–127.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Andrea M Zawoyski, Scott P Ardoin, and Katherine S Binder. 2015. Using eye tracking to observe differential effects of repeated readings for second-grade students as a function of achievement level. *Reading Research Quarterly*, 50(2):171–184.

Mengxiao Zhu and Gary Feng. 2015. An exploratory study using social network analysis to model eye movements in mathematics problem solving. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, pages 383–387.

A Word and Fixation Level Features

This section describes the features that constitute both e^{word} and e^{fix} in human and generated trials. The full lists of eye movement features at both word and fixation levels appear in Table 3. Additionally, the linguistic word properties that, together with word-level eye movement features, constitute e^{word} are listed in Table 4.

Here we provide definitions for the standard eye movement measures presented in Section 5.1 that are averaged per-word:

Total Fixation Duration (TFD) The total duration of fixations on a word.

Number of Fixations The total number of fixations on a word.

Gaze Duration (GD) The total duration of all fixations on a word from the first time entering it to the first time exiting it. See Appendix C for details on GD implementation in E-Z Reader.

First Fixation Duration (FFD) The duration of the first fixation on a word.

Regression Rate The proportion of saccades per word that move regressively (backwards).

B Trial Level Feature Sets

For all tasks, we computed all features for each trial independently of other trials.

In all feature-based models, we address the strong co-linearity observed among several features by applying Principal Component Analysis (PCA). A PCA model is fit on the training set, and the training, validation, and test features are transformed using the trained PCA. The number of PCA components is determined as the minimum required to maintain a specified fraction of explained variance. This fraction is optimized during hyperparameter tuning (see Appendix E for the search space we use).

In addition to the per-word averaged standard eye movement features listed in the previous section, we also computed additional measures for each trial:

- **Skip Rate** the proportion of skipped words in a trial (i.e., the fraction of words where `total_skip = 1`; see Table 3).
- **Reading Speed** the total reading time of the paragraph divided by the number of words in the paragraph. Implemented as `PARAGRAPH_RT` / Paragraph Length; see Table 3.
- **`num_of_words_with_TFD_GD_diff`:** This is the proportion of fixated words for which TFD > GD, indicating refixations on a word after the first pass.
- **`mean_without_first_run_dwell_time`:** For words fixated more than once (including first pass only), this feature represents the average extra fixation duration per additional fixation (i.e., TFD minus GD, divided by the number of additional fixations). If no word is fixated more than once, the value is set to 0.

B.1 Word Property Coefficients

The formula for the linear model is:

$$Measure \sim 1 + Surp + Freq + Length + \\ Freq : Length + normalized\_word\_index$$

For each trial, we fit a linear model using the OLS function from the `Statsmodels` library (Seabold and Perktold, 2010). Before fitting the model, we normalize all measures. In order to maintain the assumptions of the linear model, we exclude zero values for the measures TFD, FFD, GD (their original distribution is normally-shaped with a point mass at zero due to the high number of skips). Surprisal (Hale, 2001; Levy, 2008) is defined as $-\log_2(p(word|context))$ for each *word* given the preceding textual content of the textual item as *context*, probabilities extracted from the GPT-2-small language model (Radford et al., 2019; Wolf et al., 2020). Frequency is based on the Wordfreq package (Speer, 2022), formulated as $\log_2(p(word))$. Length is defined by the number of characters, ignoring punctuation. We also include $normalized\_word\_index$ following the results presented in (Shubi and Berzak, 2023), which show general decrease in reading times for later words within each paragraph in OneStop.

B.2 Saccade Network Measures

As described in Section 5.1, we define the directed graph $G = \{V, T\}$ such that V is the set of words in W, and for all $u, v \in V$:

$$T = \{(u, v) : \text{there is a saccade from } u \text{ to } v\}$$

A visualization example of two such networks appears in Figure 5.

Feature Name	Description
Word-Level Eye Movement Features	
IA_DWELL_TIME	(**TFD**) The sum of the duration across all fixations that fell in the current interest area
IA_DWELL_TIME_%	Percentage of trial time spent on the current interest area (IA_DWELL_TIME / PARAGRAPH_RT).
IA_FIXATION_COUNT	Total number of fixations falling in the interest area.
IA_REGRESSION_IN_COUNT	(**Regression Rate**) Number of times interest area was entered from a higher IA_ID (from the right in English).
IA_REGRESSION_OUT_FULL_COUNT	Number of times interest area was exited to a lower IA_ID (to the left in English).
IA_FIRST_FIX_PROGRESSIVE	Checks whether the first fixation in the interest area is a first-pass fixation.
IA_FIRST_FIXATION_DURATION	(**FFD**) Duration of the first fixation event that was within the current interest area
IA_FIRST_RUN_DWELL_TIME	(**GD**) Dwell time of the first run (i.e., the sum of the duration of all fixations in the first run of fixations within the current interest area).
IA_TOP	Y coordinate of the top of the interest area.
IA_LEFT	X coordinate of the left-most part of the interest area.
normalized_Word_ID	Position in the paragraph of the word interest area, normalized from zero to one.
IA_REGRESSION_OUT_COUNT	Number of times interest area was exited to a lower IA_ID (to the left in English) before a higher IA_ID was fixated in the trial.
PARAGRAPH_RT	Reading time of the entire paragraph.
total_skip	Binary indicator whether the word was fixated on.
Fixation-level Eye Movement Features	
CURRENT_FIX_INDEX	The position of the current fixation in the trial.
CURRENT_FIX_DURATION	Duration of the current fixation.
CURRENT_FIX_X	X coordinate of the current fixation.
CURRENT_FIX_Y	Y coordinate of the current fixation.
CURRENT_FIX_INTEREST_AREA_INDEX	The word index (IA_ID) on which the current fixation occurred.
NEXT_FIX_INTEREST_AREA_INDEX	The word index (IA_ID) on which the next fixation occurred.

Table 3: Word-level and fixation-level eye movement features, defined and extracted by SR Data Viewer.

Feature Name	Description
Surprisal	(Hale, 2001; Levy, 2008), formulated as $-\log_2(p(word\|context))$ for each *word* given the preceding textual content of the paragraph as *context*, probabilities extracted from the GPT-2-small language model (Radford et al., 2019; Wolf et al., 2020).
Wordfreq_Frequency	Frequency of the word based on the Wordfreq package (Speer, 2022), formulated as $-\log_2(p(word))$.
Length	Length of the word in characters.
start_of_line	Binary indicator of whether the word appeared at the beginning of a line.
end_of_line	Binary indicator of whether the word appeared at the end of a line.
Is_Content_Word	Binary indicator of whether the word is a content word. A content word is defined as a word that has a part-of-speech tag of either PROPN, NOUN, VERB, ADV, or ADJ.
n_Lefts	The number of leftward immediate children of the word in the syntactic dependency parse.
n_Rights	The number of rightward immediate children of the word in the syntactic dependency parse.
Distance2Head	The number of words to the syntactic head of the word.

Table 4: Linguistic word properties and their descriptions. POS tags and parse trees were obtained using SpaCy (Honnibal et al., 2020).

The following measures are computed for each saccade network instance:

1. **Average Degree**:

$$\text{Avg Degree} = \frac{\sum_{v \in V} \deg(v)}{|V|}$$

where $\deg(v)$ is the degree of vertex v, and $|V|$ is the number of vertices in G.

2. **Density**:

$$\text{Density} = \frac{2|T|}{|V|(|V|-1)}$$

for an undirected graph G, where $|T|$ is the number of edges in G.

3. **Average Clustering Coefficient**:

$$\text{Avg CC} = \frac{1}{|V|} \sum_{v \in V} C(v)$$

where $C(v)$ is the clustering coefficient of vertex v, defined as the fraction of pairs of neighbors of v that are connected.

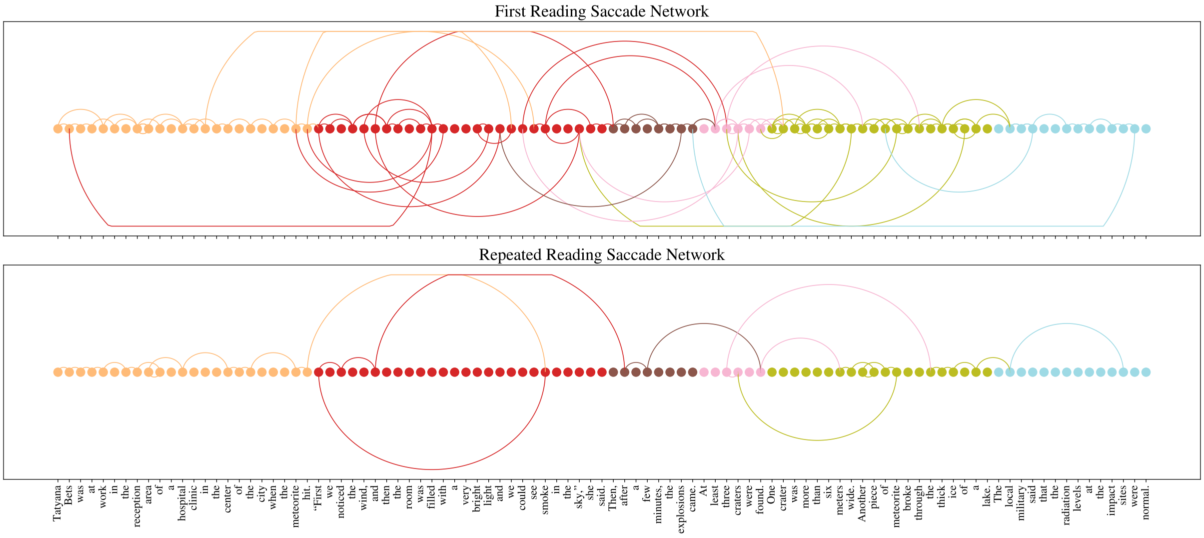

Figure 5: Visualization of two Saccade Networks as defined in Section 5.1. Circles represent words and arcs represent saccades between words. Different colors indicate different sentences within the paragraph. The top network represents the first reading, while the bottom network corresponds to the repeated reading of the same paragraph by the same participant.

4. **Average Betweenness Centrality**:

$$\text{Avg Betweenness} = \frac{1}{|V|} \sum_{v \in V} b(v)$$

where $b(v)$ is the betweenness centrality of vertex v, defined as the fraction of shortest paths in G that pass through v.

5. **Average Closeness Centrality**:

$$\text{Avg Closeness} = \frac{1}{|V|} \sum_{v \in V} c(v)$$

where $c(v)$ is the closeness centrality of vertex v, defined as the reciprocal of the average shortest path length from v to all other vertices in G.

6. **Transitivity**:

$$\text{Transitivity} = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets}}$$

where a triangle is a set of three mutually connected vertices, and a triplet is a set of three vertices where at least two are connected.

7. **Number of Bridges**:

$$\text{num\_bridges} = |\{e \in T \mid G \setminus \{e\} \\ \text{has more connected components than } G\}|$$

where a bridge is an edge whose removal increases the number of connected components in G.

C Scanpath Generation

For eye movement data generation we use E-Z Reader 10.4 (Veldre et al., 2023) with the default parameters estimated from (Schilling et al., 1998):

$$\begin{aligned}
&A = 25.0, \quad \alpha_1 = 124.0, \quad \alpha_2 = 11.1, \\
&\alpha_3 = 76.0, \quad \delta = 1.68, \quad \epsilon_1 = 0.1, \\
&\epsilon_2 = 0.5, \quad \epsilon_3 = 1.0, \quad \eta_1 = 0.5, \quad \xi = 0.5 \\
&\eta_2 = 0.1, \quad I = 50.0, \quad \lambda = 0.25, \\
&M_1 = 150.0, \quad M_2 = 25.0, \quad \omega_1 = 6.0, \\
&\omega_2 = 3.0, \quad p_F = 0.01, \quad \psi = 7.0, \\
&S = 25.0, \quad \sigma_\gamma = 20.0, \quad V = 60.0
\end{aligned}$$

Parameter definitions appear in (Reichle et al., 2013). We set the `includeRegressionTrials` parameter to `True` to allow inter-word regressions. Out of the 1000 generated scanpaths per text, we choose the prototype scanpath to be the one which minimizes the mean Scasim (Von der Malsburg and Vasishth, 2011) distance to all other scanpaths. We use Scasim with the following formula

$$\texttt{CURRENT\_FIX\_DURATION} \sim \texttt{CURRENT\_FIX\_X} \\ + \texttt{CURRENT\_FIX\_Y}$$

with the parameters `center_x=1280`, `center_y=720`, `distance=77`, `unit_size=1/60` and `normalize=False` because all scanpaths correspond to the same text and model parameters.

C.1 Modifications in Word and Fixation Level Features

For each text, we compute the same set of word- and fixation-level features used in human trials, as detailed in Appendix A. Here, we highlight differences in feature definitions between human trials and those generated by the E-Z Reader model.

In OneStop, each trial corresponds to a single reading of a textual item by one subject. As a result, the word-level measures for a trial are extracted from a single scanpath. In contrast, the E-Z Reader model derives its word-level measures by aggregating scanpaths from 1000 statistical subjects. This aggregation introduces several differences compared to human-derived features:

- **`IA_SKIP`:**
  In human trials (see Table 3), `IA_SKIP` is binary, indicating whether a word was skipped. In E-Z Reader trials, however, this variable takes on a continuous value between 0 and 1, representing the proportion of statistical subjects who skipped the word.

- **Reading Time Measures for Skipped Words:**
  For human data, skipped words (i.e., those with `total_skip = 1`) are assigned a value of zero for all reading-time features (e.g., TF, GD, FFD). In the E-Z Reader model, reading-time measures are summed over all statistical subjects and then normalized by the number of subjects who fixated the word. Consequently, these measures reflect only the fixation cases. To address this discrepancy, we excluded non-fixated words from human trials in the analysis presented in Table 1 and from the trial-level feature extraction used in the feature-based models.

- **`PARAGRAPH_RT` and `IA_DWELL_TIME_%`:**
  For E-Z Reader trials, these measures were computed using an "expected `IA_DWELL_TIME`," which is calculated as

  $$\texttt{IA\_DWELL\_TIME} \times (1 - \texttt{total\_skip}) \,.$$

  This adjustment ensures that the measures account for the proportion of fixated versus skipped words.

Gaze Duration (GD) We retain the original implementation of GD, which differs from the version provided in SR Data Viewer. In E-Z Reader, GD is computed at the word level as:

$$GD(w) = \frac{\sum_{S_{FP}} \text{First run dwell time}}{|S|}$$

Where S represents the subset of statistical subjects who fixated on word w, and S_{FP} represents the subset of statistical subjects who fixated on word w **during first pass**. This formulation can result in GD being smaller than FFD in some cases.

Fixation Location on Screen For human trials, the features `CURRENT_FIX_X` and `CURRENT_FIX_Y` specify the coordinates of each fixation on the screen. As E-Z Reader is inherently incapable of providing such features, we approximate them by using the center of each word.

C.2 Synthetic Data Analysis

For the comparison between human and E-Z Reader generated trials, presented in Table 1, we use the same set of trial pairs as in model training (both first reading and repeated reading). For each human trial, we first extract all measures listed in Table 1, yielding a single value per measure type and human trial (in total: 4 measures $\times$ (1944 $\times$ 2) trials). To obtain the values presented in Table 1, for each combination of eye movement measure and comparison type (either First Reading versus E-Z Reader or Repeated Reading versus E-Z Reader) we fit a linear mixed model formulated as $measure\_diff \sim 1 + (1|text)$, and extract the mean and standard error of the intercept.

For comparing the absolute differences between first or repeated reading and E-Z Reader, we apply the following procedure:

1. **Modeling:** For each measure (Fixation Count, Mean TFD, Regression Rate, and Skip Rate), we fit a mixed-effects model:

   $$\texttt{measure\_diff} \sim \texttt{comparison\_type} + (1 \mid \texttt{text\_id})$$

   where `comparison_type` is a binary indicator being 1 for first reading minus E-Z Reader differences and 0 for repeated reading minus E-Z Reader differences.

2. **Test Statistic:** We define

   $$d = |\beta_{FR-EZ}| - |\beta_{FR-EZ} + \beta_{RR-EZ}|,$$

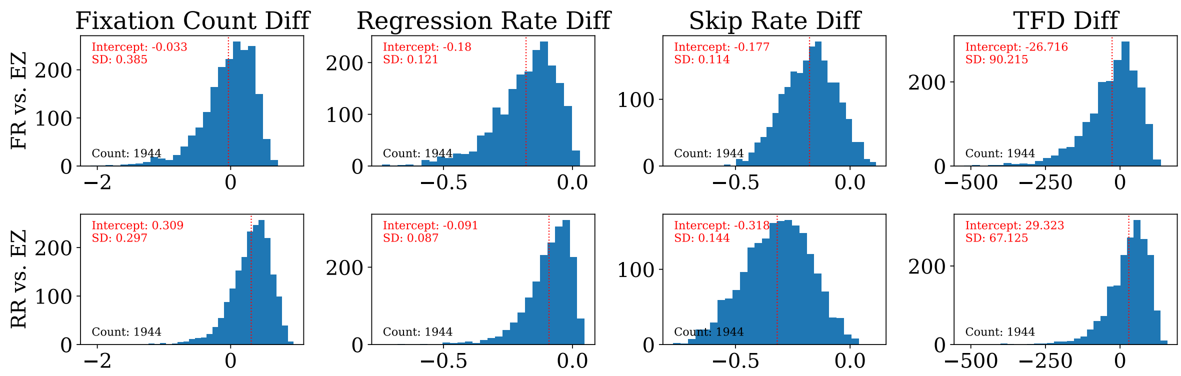

Figure 6: This visualization displays histograms representing the trial-level differences between human trials and the corresponding E-Z Reader synthesized trials. The arrangement of histograms mirrors the transposed rows and columns of Table 1. In each histogram, the top left corner shows the mean and standard deviation of the values.

where β_{FR-EZ} is the intercept (first reading) and $\beta_{FR-EZ} + \beta_{RR-EZ}$ is the mean for E-Z Reader vs. repeated reading differences.

3. **Bootstrapping:** We perform 1000 text-level bootstrap iterations (sampling with replacement) to derive the sampling distribution of d and compute one-sided p-values for all measures other than Mean TFD (for which we compute two-sided).

In addition, in Figure 6, we provide histograms that illustrate the distribution of trial-level differences for each measure.

D Cross-Validation and Participant-to-Article Assignment

For our cross-validation procedure, we split the data separately for each combination of article batch (1–3) and reading regime (information-seeking, ordinary reading). Each such combination consists of 10 articles and 60 participants, resulting in 60 "article-participant" pairs including both first and repeated reading version of each aritlce-participant pair.

Data Splitting Rationale. Our data splitting is derived from an assignment of 6 participants to each article (for a total of 60 participants). This assignment must satisfy the following:

1. All 6 participants assigned to article i have reread it (repeated reading).
2. Of these 6 participants, exactly half (3) performed *consecutive* rereading and the other half performed *nonconsecutive* rereading.
3. All 6 participants assigned to article i have read different *other* articles in the repeated reading.

Section D.1 below details the algorithmic procedure used to achieve this participant-to-article assignment.

Illustration of the Splits. Figure 3 illustrates one of the cross-validation folds for a single combination of article batch and reading regime (60 participants out of the 180). Participants are grouped by columns (from $6 \cdot (i-1)$ to $6 \cdot i$) according to the article to which they are assigned. Each table entry corresponds to the first and repeated readings of article i by participant j. Since each participant reads exactly *two* articles in repeated reading, each column has exactly two non-empty entries (one for consecutive and one for nonconsecutive repeated reading).

To create the data splits for the i-th fold, we select article $(10-i)$ and its corresponding participants as the unseen item and participants for the validation split, and $(10+1-i)$ as the unseen item and participants for the test split. Figure 3 demonstrates this process for fold 1.

Balanced Evaluation Regimes. Thanks to the properties of our participant-to-article assignment, we ensure:

1. *Balanced evaluation regimes for validation and test sets.* For each set, we balance "new item," "new participant," and "new item & participant" splits. Specifically, in the test set, each regime has 6 "article readings." In the validation set, the new item and new par-

ticipant splits have 5 “article readings” each, while the new item & participant split has 6.

2. *Balanced rereading conditions.* In the test split, all evaluation regimes are further balanced with respect to consecutive and non-consecutive repeated reading (3 of each per regime).

D.1 Participant-to-Article Assignment Algorithm

To assign participants to articles under the constraints described above, we used cvxpy (Diamond and Boyd, 2016) to solve the optimization problem detailed in Algorithm 1. This procedure is applied independently for each combination of article batch (3 batches) and reading regime (information-seeking, ordinary reading). For each combination, we extract matrices P and Q as defined in Algorithm 1, then solve to obtain a feasible assignment of participants to articles.

The optimization problem includes several key constraints to ensure the assignment is valid and meets the desired balance:

- **Constraint 1:** Each participant is assigned to exactly one article.

- **Constraints 2 and 3:** Guarantee that:

  1. Exactly 6 participants are assigned to each article.
  2. All 6 assigned participants had reread that article.
  3. Among them, half had read it in a consecutive rereading and half in a nonconsecutive rereading.

  These constraints explain the “diagonal” structure in Figure 3 and the distribution of symbols 'oo' and 'o-o' within each diagonal cell.

- **Constraint 4:** Among participants assigned to article i, at most one participant has reread any other article. This maximizes the coverage of articles in the unseen participant regime and ensures exactly 5 different participants in the unseen participant regime for the validation split.

This assignment directly underlies the cross-validation splits described earlier in this appendix (see Figure 3).

E Hyperparameter Tuning

We apply standardization for each feature in both word and fixation level representations. Mean and standard deviation are computed on the train set and applied to the validation and test sets, separately for each split. Feature normalization is performed using Scikit-learn (Pedregosa et al., 2011).

For all the neural models, we use the AdamW optimizer (Loshchilov and Hutter, 2018) with a batch size of 16, a linear warmup ratio of 0.1, and a weight decay of 0.1, following best practice recommendations from Liu et al. (2019) and Mosbach et al. (2021). In most RoBERTEye models, The search space for learning rates is $\{0.00001, 0.00003, 0.0001\}$ and for dropout $\{0.1, 0.3, 0.5\}$. To address convergence and overfitting issues in the RoBERTEye-fixation models, both for the paired-trials task variant and the augmented single-trial task variant, we search over learning rates of $0.0001, 0.001, 0.003$ and set the weight decay parameter to 10^{-5}. For RoBERTEye models, we allow the backbone RoBERTa weights to either be frozen or trainable. For all XGBoost based models, we searched over learning rate $\in \{0.3, 0.01, 0.001\}$, number of estimators $\in \{10, 100, 1000\}$, maximal tree depth $\in \{4, 6, 10\}$ and a regularization parameter $\alpha \in \{0, 0.1, 1, 10\}$. The XGBoost search space is a subset of the default search space used in (H2O.ai, 2022) and includes the default hyperparameters implemented in (Chen and Guestrin, 2016). In addition, in all feature based models we optimize the lower bound of the fraction of explained variance after a PCA transformation which constrains the number of components taken $\in \{0.8, 0.9, 1\}$.

F Hardware and Software

All neural networks are trained using the Pytorch Lighting library (Falcon and The PyTorch Lightning team, 2024; Paszke et al., 2019) and evaluated using torch-metrics (Nicki Skafte Detlefsen et al., 2022) on NVIDIA A100-40GB or comparable GPUs. For RoBERTEye we adapt the code from Shubi et al. (2024b).The baselines described in Section 5.4 are reimplemented in this framework as well. A single training epoch took approximately 3 minutes. We train for a maximum of 30 epochs, stopping after 10 epochs without improvement on the validation set. Due to convergence issues, based on early experiments, we trained the E-Z Reader augmented RoBERTEye-Fixations in the single

Algorithm 1 Participant-to-articles assignment algorithm

Input: Parameters and variables:

- m: number of items (articles).
- n: number of participants.
- $P \in \{0,1\}^{m\times n}$: $P_{i,j} = 1$ if participant j read article i in a *consecutive* repeated reading, 0 otherwise.
- $Q \in \{0,1\}^{m\times n}$: $Q_{i,j} = 1$ if participant j read article i in a *nonconsecutive* repeated reading, 0 otherwise.

Variable: $B \in \{0,1\}^{n\times m}$

Objective: Minimize a constant (null optimization problem):

$$\text{Minimize: } 0$$

Constraints:

$$\text{Constraint 1: } \forall j \in [n] : \sum_{i\in[m]} B_{j,i} = 1$$
$$\text{Constraint 2: } \forall i \in [m] : \text{row}_i(P) \cdot \text{col}_i(B) = 3$$
$$\text{Constraint 3: } \forall i \in [m] : \text{row}_i(Q) \cdot \text{col}_i(B) = 3$$
$$\text{Constraint 4: } (P + Q) \cdot B - 6 \cdot I \leq 1$$

Output: Solve this constrained null optimization problem using `cvxpy` to obtain a feasible solution B which satisfies all constraints. The assignment for each participant $j \in [n]$ is then

$$\text{assigned\_article}(j) = \arg\max(\text{row}_j(B)).$$

trial task variant for 50 epochs, with early stopping after 15 epochs without improvements .

The number of trainable model parameters for RoBERTEye is either between 2-3M parameters (depending on RoBERTEye-Words or RoBERT-Eye Fixations) when the backbone RoBERTa is frozen, otherwise 355M.

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.

We utilized the `lme4` package in `R` (Bates et al., 2015) and the `MixedModels` package in `Julia` (Alday and Bates, 2025) for fitting linear mixed-models.

G Results

In this section, we present additional results for the two task variations. Validation accuracy is reported in Table 5, F1 scores for both validation and test partitions are shown in Table 6, and Recall and Precision for both validation and test partitions are provided in Table 7. For all result tables, 95% confidence intervals are standard normal bootstrap confidence interval (Davison and Hinkley, 1997) with $B = 1000$. In addition, when comparing between models, we also experimented with adding a random effect for the fold number, but the low variance between folds prevented the model from converging.

Eval Type	Task Variant	Model	Eye Movements Input	New Item Seen Participant	New Participant Seen Item	New Item & Participant	All
Validation	Paired Trials	Reading Speed	$E_S^{W,r}, E_S^{W,r'}$	$87.7_{\pm 2.3}$	$88.8_{\pm 2.2}$	$87.1_{\pm 2.1}$	$87.9_{\pm 1.3}$
		Feature-Based	$E_S^{W,r}, E_S^{W,r'}$	$92.4_{\pm 1.8}$	$92.5_{\pm 1.8}$	$90.7_{\pm 1.9}$	$91.8_{\pm 1.0}$
		RoBERTEye Fixations	$E_S^{W,r}, E_S^{W,r'}$	$92.3_{\pm 1.8}$	$91.1_{\pm 1.9}$	$90.4_{\pm 1.9}$	$91.2_{\pm 1.0}$
		RoBERTEye Words	$E_S^{W,r}, E_S^{W,r'}$	$92.3_{\pm 1.8}$	$91.1_{\pm 1.9}$	$89.9_{\pm 1.9}$	$91.2_{\pm 1.0}$
	Single Trial	Reading Speed	$E_S^{W,r}$	$66.7_{\pm 2.2}$	$66.8_{\pm 2.2}$	$66.5_{\pm 2.2}$	$66.7_{\pm 1.3}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$67.7_{\pm 2.2}$	$66.9_{\pm 2.2}$	$66.0_{\pm 2.1}$	$66.8_{\pm 1.3}$
		Feature-Based	$E_S^{W,r}$	$71.1_{\pm 2.2}$	$71.4_{\pm 2.1}$	$69.8_{\pm 2.1}$	$70.7_{\pm 1.3}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$71.1_{\pm 2.2}$	$70.1_{\pm 2.2}$	$69.4_{\pm 2.1}$	$70.1_{\pm 1.2}$
		RoBERTEye Fixations	$E_S^{W,r}$	$72.2_{\pm 2.2}$	$74.3_{\pm 2.0}$	$72.2_{\pm 2.0}$	$72.8_{\pm 1.2}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$72.0_{\pm 2.2}$	$73.5_{\pm 2.0}$	$71.4_{\pm 2.0}$	$72.2_{\pm 1.3}$
		RoBERTEye Words	$E_S^{W,r}$	$72.7_{\pm 2.1}$	$72.0_{\pm 2.2}$	$70.9_{\pm 2.0}$	$71.8_{\pm 1.2}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$73.3_{\pm 2.1}$	$74.3_{\pm 2.1}$	$72.0_{\pm 2.0}$	$73.1_{\pm 1.2}$

Table 5: Validation accuracy results for the two variants of the first vs. second reading prediction task with 95% confidence intervals, aggregated across 10 cross-validation splits. $E_{EZ}^{W,1}$ denotes synthesized eye movements generated using E-Z Reader (Reichle et al., 2003).

Eval Type	Task Variant	Model	Eye Movements Input	New Item Seen Participant	New Participant Seen Item	New Item & Participant	All
Validation	Paired Trials	Reading Speed	$E_S^{W,r}, E_S^{W,r'}$	$87.9_{\pm 2.4}$	$88.5_{\pm 2.3}$	$87.3_{\pm 2.2}$	$87.9_{\pm 1.3}$
		Feature-Based	$E_S^{W,r}, E_S^{W,r'}$	$92.5_{\pm 1.9}$	$92.4_{\pm 1.8}$	$90.7_{\pm 2.0}$	$91.8_{\pm 1.1}$
		RoBERTEye Fixations	$E_S^{W,r}, E_S^{W,r'}$	$92.1_{\pm 1.9}$	$91.2_{\pm 2.0}$	$89.8_{\pm 2.0}$	$91.0_{\pm 1.1}$
		RoBERTEye Words	$E_S^{W,r}, E_S^{W,r'}$	$92.4_{\pm 1.9}$	$90.9_{\pm 2.0}$	$90.4_{\pm 2.0}$	$91.2_{\pm 1.1}$
	Single Trial	Reading Speed	$E_S^{W,r}$	$65.1_{\pm 2.7}$	$64.7_{\pm 2.7}$	$65.3_{\pm 2.5}$	$65.1_{\pm 1.5}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$65.9_{\pm 2.7}$	$64.7_{\pm 2.6}$	$64.6_{\pm 2.5}$	$65.0_{\pm 1.5}$
		Feature-Based	$E_S^{W,r}$	$70.5_{\pm 2.5}$	$70.2_{\pm 2.5}$	$69.0_{\pm 2.3}$	$69.9_{\pm 1.4}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$70.4_{\pm 2.5}$	$68.3_{\pm 2.6}$	$68.8_{\pm 2.5}$	$69.1_{\pm 1.5}$
		RoBERTEye Fixations	$E_S^{W,r}$	$71.6_{\pm 2.5}$	$72.8_{\pm 2.4}$	$71.5_{\pm 2.4}$	$72.0_{\pm 1.4}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$71.9_{\pm 2.5}$	$73.6_{\pm 2.2}$	$71.4_{\pm 2.3}$	$72.2_{\pm 1.4}$
		RoBERTEye Words	$E_S^{W,r}$	$72.1_{\pm 2.4}$	$69.9_{\pm 2.6}$	$70.0_{\pm 2.4}$	$70.6_{\pm 1.4}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$72.3_{\pm 2.5}$	$72.4_{\pm 2.5}$	$71.2_{\pm 2.3}$	$71.9_{\pm 1.3}$
Test	Paired Trials	Reading Speed	$E_S^{W,r}, E_S^{W,r'}$	$87.8_{\pm 2.2}$	$87.9_{\pm 2.2}$	$87.2_{\pm 2.1}$	$87.6_{\pm 1.3}$
		Feature-Based	$E_S^{W,r}, E_S^{W,r'}$	$91.5_{\pm 1.8}$	$92.1_{\pm 1.8}$	$90.6_{\pm 1.9}$	$91.4_{\pm 1.1}$
		RoBERTEye Fixations	$E_S^{W,r}, E_S^{W,r'}$	$88.9_{\pm 2.0}$	$87.6_{\pm 2.1}$	$87.6_{\pm 2.2}$	$88.0_{\pm 1.2}$
		RoBERTEye Words	$E_S^{W,r}, E_S^{W,r'}$	$90.0_{\pm 2.0}$	$89.8_{\pm 2.0}$	$88.5_{\pm 2.1}$	$89.4_{\pm 1.2}$
	Single Trial	Reading Speed	$E_S^{W,r}$	$64.8_{\pm 2.4}$	$65.3_{\pm 2.6}$	$65.2_{\pm 2.5}$	$65.1_{\pm 1.4}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$65.4_{\pm 2.6}$	$65.6_{\pm 2.5}$	$65.2_{\pm 2.6}$	$65.4_{\pm 1.5}$
		Feature-Based	$E_S^{W,r}$	$68.6_{\pm 2.4}$	$69.8_{\pm 2.4}$	$67.8_{\pm 2.5}$	$68.7_{\pm 1.4}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$68.7_{\pm 2.4}$	$70.3_{\pm 2.3}$	$68.6_{\pm 2.4}$	$69.2_{\pm 1.4}$
		RoBERTEye Fixations	$E_S^{W,r}$	$69.7_{\pm 2.4}$	$68.2_{\pm 2.5}$	$69.5_{\pm 2.3}$	$69.0_{\pm 1.4}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$70.0_{\pm 2.3}$	$71.1_{\pm 2.3}$	$70.2_{\pm 2.4}$	$70.4_{\pm 1.4}$
		RoBERTEye Words	$E_S^{W,r}$	$67.5_{\pm 2.4}$	$69.6_{\pm 2.4}$	$68.1_{\pm 2.5}$	$68.3_{\pm 1.4}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$69.0_{\pm 2.4}$	$68.7_{\pm 2.5}$	$69.0_{\pm 2.4}$	$68.9_{\pm 1.4}$

Table 6: F1 results for the two variants of the first vs. second reading prediction task with 95% confidence intervals, aggregated across 10 cross-validation splits, and presented for both test and validation partitions. $E_{EZ}^{W,1}$ denotes synthesized eye movements generated using E-Z Reader (Reichle et al., 2003).

Eval Type	Task Variant	Model	Eye Movements Input	New Item Seen Participant		New Participant Seen Item		New Item & Participant		All	
				Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
Validation	Paired Trials	Reading Speed	$E_S^{W,r}, E_S^{W,r'}$	$87.1_{\pm 3.2}$	$88.7_{\pm 3.1}$	$89.0_{\pm 3.1}$	$88.1_{\pm 3.2}$	$87.5_{\pm 2.9}$	$87.2_{\pm 2.9}$	$87.8_{\pm 1.8}$	$88.0_{\pm 1.7}$
		XGBoost	$E_S^{W,r}, E_S^{W,r'}$	$92.0_{\pm 2.7}$	$93.1_{\pm 2.4}$	$91.2_{\pm 2.7}$	$93.7_{\pm 2.3}$	$92.1_{\pm 2.5}$	$89.5_{\pm 2.8}$	$91.8_{\pm 1.5}$	$91.9_{\pm 1.5}$
		RoBERTEye Fixations	$E_S^{W,r}, E_S^{W,r'}$	$90.7_{\pm 2.7}$	$93.5_{\pm 2.5}$	$90.6_{\pm 2.8}$	$91.8_{\pm 2.6}$	$88.6_{\pm 2.7}$	$91.1_{\pm 2.5}$	$89.9_{\pm 1.6}$	$92.1_{\pm 1.5}$
		RoBERTEye Words	$E_S^{W,r}, E_S^{W,r'}$	$92.0_{\pm 2.6}$	$92.9_{\pm 2.5}$	$91.0_{\pm 2.7}$	$90.8_{\pm 2.7}$	$92.2_{\pm 2.4}$	$88.6_{\pm 2.8}$	$91.7_{\pm 1.5}$	$90.6_{\pm 1.6}$
	Single Trial	Reading Speed	$E_S^{W,r}$	$68.5_{\pm 3.3}$	$62.0_{\pm 3.2}$	$69.2_{\pm 3.3}$	$60.8_{\pm 3.2}$	$67.7_{\pm 3.1}$	$63.0_{\pm 3.0}$	$68.4_{\pm 1.9}$	$62.0_{\pm 1.9}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$69.9_{\pm 3.2}$	$62.4_{\pm 3.3}$	$69.2_{\pm 3.3}$	$60.7_{\pm 3.2}$	$67.4_{\pm 3.0}$	$62.0_{\pm 3.1}$	$68.7_{\pm 1.9}$	$61.7_{\pm 1.8}$
		XGBoost	$E_S^{W,r}$	$72.1_{\pm 3.1}$	$68.9_{\pm 3.2}$	$73.2_{\pm 3.0}$	$67.5_{\pm 3.2}$	$70.8_{\pm 2.8}$	$67.4_{\pm 2.9}$	$72.0_{\pm 1.8}$	$67.9_{\pm 1.7}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$72.3_{\pm 3.1}$	$68.6_{\pm 3.2}$	$72.5_{\pm 3.3}$	$64.6_{\pm 3.2}$	$70.2_{\pm 3.0}$	$67.4_{\pm 3.1}$	$71.5_{\pm 1.8}$	$66.9_{\pm 1.8}$
		RoBERTEye Fixations	$E_S^{W,r}$	$73.0_{\pm 3.2}$	$70.3_{\pm 3.1}$	$77.1_{\pm 3.0}$	$69.1_{\pm 3.1}$	$73.4_{\pm 2.9}$	$69.7_{\pm 3.0}$	$74.4_{\pm 1.7}$	$69.7_{\pm 1.7}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$72.2_{\pm 3.1}$	$71.7_{\pm 3.0}$	$73.1_{\pm 2.9}$	$74.1_{\pm 2.9}$	$71.4_{\pm 2.8}$	$71.5_{\pm 2.9}$	$72.2_{\pm 1.7}$	$72.3_{\pm 1.7}$
		RoBERTEye Words	$E_S^{W,r}$	$73.7_{\pm 3.1}$	$70.6_{\pm 3.1}$	$75.4_{\pm 3.2}$	$65.2_{\pm 3.2}$	$72.2_{\pm 3.0}$	$67.9_{\pm 3.0}$	$73.7_{\pm 1.8}$	$67.9_{\pm 1.7}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$75.1_{\pm 3.0}$	$69.8_{\pm 3.1}$	$77.9_{\pm 3.0}$	$67.6_{\pm 3.3}$	$73.4_{\pm 2.8}$	$69.1_{\pm 2.9}$	$75.3_{\pm 1.7}$	$68.8_{\pm 1.8}$
Test	Paired Trials	Reading Speed	$E_S^{W,r}, E_S^{W,r'}$	$88.4_{\pm 2.9}$	$87.2_{\pm 3.0}$	$88.3_{\pm 2.9}$	$87.6_{\pm 3.0}$	$88.4_{\pm 2.8}$	$86.1_{\pm 3.0}$	$88.3_{\pm 1.7}$	$86.9_{\pm 1.8}$
		XGBoost	$E_S^{W,r}, E_S^{W,r'}$	$90.5_{\pm 2.6}$	$92.5_{\pm 2.3}$	$91.4_{\pm 2.5}$	$93.0_{\pm 2.3}$	$91.9_{\pm 2.4}$	$89.3_{\pm 2.7}$	$91.2_{\pm 1.5}$	$91.6_{\pm 1.5}$
		RoBERTEye Fixations	$E_S^{W,r}, E_S^{W,r'}$	$87.2_{\pm 3.0}$	$90.6_{\pm 2.6}$	$85.8_{\pm 3.0}$	$89.4_{\pm 2.8}$	$87.3_{\pm 2.9}$	$88.1_{\pm 2.8}$	$86.7_{\pm 1.7}$	$89.4_{\pm 1.6}$
		RoBERTEye Words	$E_S^{W,r}, E_S^{W,r'}$	$89.7_{\pm 2.7}$	$90.4_{\pm 2.6}$	$90.0_{\pm 2.7}$	$89.6_{\pm 2.8}$	$88.5_{\pm 2.0}$	$87.1_{\pm 3.0}$	$89.9_{\pm 1.5}$	$89.0_{\pm 1.6}$
	Single Trial	Reading Speed	$E_S^{W,r}$	$69.2_{\pm 2.9}$	$60.9_{\pm 3.1}$	$69.1_{\pm 3.0}$	$61.9_{\pm 3.1}$	$68.4_{\pm 3.1}$	$62.4_{\pm 3.2}$	$68.9_{\pm 1.8}$	$61.7_{\pm 1.7}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$69.4_{\pm 3.2}$	$61.8_{\pm 3.1}$	$69.4_{\pm 3.1}$	$62.2_{\pm 2.9}$	$68.0_{\pm 3.1}$	$62.6_{\pm 3.1}$	$68.9_{\pm 1.8}$	$62.2_{\pm 1.8}$
		XGBoost	$E_S^{W,r}$	$70.7_{\pm 2.9}$	$66.7_{\pm 3.0}$	$72.2_{\pm 2.9}$	$67.6_{\pm 3.0}$	$69.7_{\pm 2.9}$	$66.0_{\pm 3.1}$	$70.8_{\pm 1.7}$	$66.7_{\pm 1.2}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$72.2_{\pm 3.0}$	$65.5_{\pm 2.9}$	$72.7_{\pm 2.9}$	$68.0_{\pm 2.8}$	$70.1_{\pm 2.9}$	$67.2_{\pm 3.0}$	$71.6_{\pm 1.7}$	$70.1_{\pm 2.1}$
		RoBERTEye Fixations	$E_S^{W,r}$	$72.2_{\pm 3.0}$	$67.4_{\pm 3.0}$	$71.1_{\pm 3.1}$	$65.5_{\pm 3.0}$	$71.4_{\pm 3.0}$	$67.6_{\pm 2.9}$	$71.5_{\pm 1.7}$	$66.8_{\pm 1.7}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$69.8_{\pm 3.0}$	$70.4_{\pm 2.8}$	$71.5_{\pm 2.9}$	$70.6_{\pm 2.8}$	$68.5_{\pm 2.9}$	$71.9_{\pm 2.9}$	$69.8_{\pm 1.7}$	$71.0_{\pm 1.7}$
		RoBERTEye Words	$E_S^{W,r}$	$71.1_{\pm 3.0}$	$64.3_{\pm 3.0}$	$73.1_{\pm 2.9}$	$66.4_{\pm 3.0}$	$70.2_{\pm 3.1}$	$66.1_{\pm 3.0}$	$71.4_{\pm 1.7}$	$65.5_{\pm 1.8}$
			$E_{EZ}^{W,1}, E_S^{W,r}$	$73.0_{\pm 3.0}$	$66.5_{\pm 3.0}$	$73.0_{\pm 3.0}$	$65.0_{\pm 3.1}$	$71.4_{\pm 3.0}$	$66.8_{\pm 3.0}$	$71.9_{\pm 1.8}$	$66.1_{\pm 1.7}$

Table 7: Precision and Recall (repeated reading being positive and first reading being negative) results for the two variants of the first vs. second reading prediction task with 95% confidence intervals, aggregated across 10 cross-validation splits, and presented for both test and validation partitions. $E_{EZ}^{W,1}$ denotes synthesized eye movements generated using E-Z Reader (Reichle et al., 2003).