Fine-Grained Prediction of Reading Comprehension from Eye Movements

Omer Shubi[1], Yoav Meiri[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,meiri.yoav,kfir-hadar}@campus.technion.ac.il, berzak@technion.ac.il

Abstract

Can human reading comprehension be assessed from eye movements in reading? In this work, we address this longstanding question using large-scale eyetracking data. We focus on a cardinal and largely unaddressed variant of this question: predicting reading comprehension of a single participant for a *single question* from their eye movements over a *single paragraph*. We tackle this task using a battery of recent models from the literature, and three new multimodal language models. We evaluate the models in two different reading regimes: ordinary reading and information seeking, and examine their generalization to new textual items, new participants, and the combination of both. The evaluations suggest that the task is *highly challenging*, and highlight the importance of benchmarking against a strong text-only baseline. While in some cases eye movements provide improvements over such a baseline, they tend to be small. This could be due to limitations of current modelling approaches, limitations of the data, or because eye movement behavior does not sufficiently pertain to finegrained aspects of reading comprehension processes. Our study provides an infrastructure for making further progress on this question.[1]

1 Introduction

Reading comprehension is an indispensable skill for successful participation in modern society. Consequently, many efforts and resources are invested in the development of reading comprehension assessments by educational institutions and commercial companies. The standard, and to date the only practical way to assess reading comprehension is through behavioral tasks, most commonly reading comprehension questions. However, despite its clear value and ubiquitous use, this approach is extremely time-consuming and costly, which severely

limits the volume and public availability of reading comprehension tests. Further, this testing methodology relies on *offline* behavioral signals – the end responses to a few select reading comprehension questions, and has no ability to trace the rich *online* reading comprehension processes as they unfold over time.

An alternative vision for assessing reading comprehension has been emerging in psycholinguistics and the psychology of reading. It posits that reading comprehension may be decoded in realtime directly from eye movements in reading. This vision is rooted in literature that suggests a tight correspondence between eye movements and real time language comprehension (Just and Carpenter, 1980; Rayner, 1998; Rayner et al., 2016, among others). With the rise of modern machine learning and NLP, multiple studies over the past decade attempted to use eye movement data to predict reading comprehension (Copeland et al., 2014; Ahn et al., 2020; Reich et al., 2022; Mézière et al., 2023b, among others). This line of work suggests that in some cases various aspects of reading comprehension can be predicted from eye movements with above-chance performance. However, despite the advances so far, predictive modeling of reading comprehension from gaze is still in its infancy.

A number of factors have been hindering progress in this area. One is the paucity and small size of reading comprehension data paired with eye movements. Second, the task of reading comprehension prediction has thus far been predominantly formulated as prediction of *aggregated scores across multiple questions* rather than prediction of comprehension at the resolution of an individual question. Further, reading comprehension has been primarily studied when the reader has no specific goals with respect to the text beyond general comprehension, a regime that we refer to as *ordinary reading*. Many other reading regimes common in daily life, such as explicit information

---

[1]Code is available at https://github.com/lacclab/Reading-Comprehension-Prediction.

seeking, remain largely unaddressed. Finally, despite the dramatic progress in machine learning and NLP in recent years, effective joint modeling of text and eye movements remains a nascent and challenging domain of investigation.

In this work, we take a step forward in advancing the state-of-the-art in eye movement-based prediction of reading comprehension by combining new models, new data, and systematic evaluations. Our primary contributions are the following:

- **Task**: we introduce the challenging and largely unaddressed task of predicting the reading comprehension of a single reader with respect to a *single reading comprehension question over one passage*. This task is enabled by OneStop Eye Movements (Malmaud et al., 2020), the largest eyetracking for reading comprehension dataset to date with 486 multiple-choice questions and 19,440 question responses from 360 participants.

- **Modeling**: we develop three new models that combine text and eye movements based on the transformer encoder architecture: RoBERTa-QEye, MAG-QEye, and PostFusion-QEye. These models address both test format-agnostic and multiple-choice specific variants of the task.

- **Reading Regimes**: we study reading comprehension not only in ordinary reading but also in information seeking, a highly common but understudied reading scenario.

- **Evaluation**: we evaluate our models against a battery of existing models for prediction of reading comprehension from eye movements, and a strong text-only baseline. To this end, we use a detailed evaluation protocol targeting three different levels of model generalization: new participant, new textual item, and the combination of both.

2 Related Work

Our study contributes to an existing body of work on the prediction of reading comprehension from eye movements in reading. To address various aspects of this task, prior studies used a wide range of models, including linear models (Mézière et al., 2023b,a), kernel methods (Makowski et al., 2019), feed-forward networks (e.g. Copeland et al., 2014), CNNs (Ahn et al., 2020) and RNNs (e.g. Ahn et al.,

2020; Reich et al., 2022). These were typically applied to the prediction of aggregated comprehension scores over multiple items. In this work, we evaluate multiple models from prior work on the single-item reading comprehension task.

While transformer models (Vaswani et al., 2017), have been used for joint modeling of eye movements and text (e.g. Deng et al., 2023; Yang and Hollenstein, 2023), they have not been applied to the problem of reading comprehension prediction from eye movements. In this work we introduce three new transformer models which draw on multimodal transformers, in particular MAG (Rahman et al., 2020) which integrated text, speech and vision for sentiment analysis, and language vision models such as VisualBERT (Li et al., 2019) (see Zhu et al. (2023); Xu et al. (2023) for reviews).

Most prior studies on reading comprehension prediction from eye movements relied solely on eye movement features (Copeland et al., 2014; Southwell et al., 2020; Ahn et al., 2020; Mézière et al., 2023b,a), while a few combined eye movements with properties of the underlying text (Martínez-Gómez and Aizawa, 2014; Makowski et al., 2019; Reich et al., 2022). In the current work, we take the latter, under-explored approach. The importance of combining eye movements with attributes of the text is motivated by a large literature in the psychology of reading which points to systematic effects of linguistic properties of the text on reading times (Rayner, 1998; Rayner et al., 2004; Kliegl et al., 2004; Demberg and Keller, 2008; Smith and Levy, 2013, among others), in particular in the context of reading comprehension (Just and Carpenter, 1980) and linguistic proficiency (Berzak et al., 2018; Berzak and Levy, 2023).

While highly informative, existing work is critically limited by small data, especially with respect to the number of available questions and participants. For example, Copeland et al. (2014) have 9 text pages, 18 questions and 39 participants. SB-SAT (Ahn et al., 2020), the only publicly available eyetracking dataset for reading comprehension, has 22 text pages, 20 questions, and 95 participants. The small size of previously used datasets severely limits the potential of NLP and machine learning approaches for reading comprehension prediction. At the same time, the reading comprehension component of broad coverage eyetracking datasets such as MECO (Siegelman et al., 2022) and CELER (Berzak et al., 2022) comprises only simple comprehension questions that serve as attention checks,

and as such are not well suited for studying reading comprehension. OneStop, used here, has a large number of items, participants and questions, enabling meaningfully addressing item-level prediction of comprehension.

Prior work varies in experimental designs. In several studies, multiple questions are presented after reading a multi-screen text without the ability to return to the text (Makowski et al., 2019; Ahn et al., 2020; Reich et al., 2022). This design is advantageous in the separation of text reading and question answering, but can lead to loose relations between eye movements and question-answering behavior due to memory limitations. In other studies, such as Copeland et al. (2014), participants can switch back and forth between the text and the questions. This creates a complex mix of ordinary reading and information seeking components which are difficult to disentangle. In OneStop, a single question appears immediately after reading a single text page, setting a middle ground between the two primary existing approaches for question presentation, and alleviating their main disadvantages. At the same time, it includes a question preview manipulation which allows to systematically compare reading comprehension in ordinary reading and question guided information seeking.

An additional limitation of prior work is the scope and nature of the evaluations. With the exception of Copeland et al. (2014), both training and evaluation were previously carried out over *aggregated responses* across multiple questions, and in some cases also across multiple texts. These approaches, which focus on measuring overall comprehension, do not enable testing direct links between eye movements and understanding specific aspects of the text. In several studies (Martínez-Gómez and Aizawa, 2014; Makowski et al., 2019; Ahn et al., 2020; Reich et al., 2022), an additional step was taken, binning comprehension scores into two binary categories, high versus low comprehension, thus further simplifying the task.

A second important evaluation limitation in prior work is evaluations in which eyetracking data for both the test participants and items is used in the training set. To our knowledge, except for Makowski et al. (2019), no work has evaluated reading comprehension prediction when neither the participant nor the item appears in the training data. This evaluation regime is needed to fully characterize model generalization ability. Importantly, even in less challenging regimes and with aggre-

gated scores and binning, model performance in prior work is typically only modestly higher than chance level. More stringent evaluations without binning comprehension scores (Martínez-Gómez and Aizawa, 2014), or with held-out participants and/or items (Makowski et al., 2019; Reich et al., 2022) tend to exhibit chance level performance. These results suggest that generalization in reading comprehension prediction is highly challenging.

3 Eyetracking Data

We use OneStop, an extended version of the dataset collected by Malmaud et al. (2020) over the textual materials of OneStopQA (Berzak et al., 2020). OneStop is the largest English L1 eyetracking for reading corpus to date. The data was collected using an Eyelink 1000+ eyetracker at a sampling rate of 1000Hz. In this dataset, 360 adult native English participants read newswire articles from the Guardian, and answer a multiple-choice reading comprehension question about each paragraph. The dataset includes 30 articles divided into 162 paragraphs. The average paragraph length is 109 words. Each paragraph has 3 possible questions, corresponding to a total of 486 questions.

Answer	Category	Degree of Comprehension	Gathering	Hunting
A	Correct	Full comprehension	7,890 (81.2)	8,450 (86.9)
B	Incorrect	Identified question-relevant information	1000 (10.3)	744 (7.7)
C	Incorrect	Some degree of attention to the text	568 (5.8)	374 (3.8)
D	Incorrect	No evidence for comprehension	260 (2.7)	152 (1.6)

Table 1: Summary of the STARC annotation framework for answer types A–D, their corresponding degree of comprehension, and number of trials in which each answer type was chosen in OneStop. Values in parentheses are percentages by reading regime.

The articles are divided into three 10-article batches, where each participant is assigned to one batch. In each trial of the experiment, participants read a paragraph and then proceed to answer one of the three possible questions on a new screen, without the ability to return to the paragraph. 180 participants are in an ordinary reading (Gathering) regime where they do not see the question prior to reading the paragraph. The remaining 180 participants are in an information seeking regime (Hunting) where they are presented with the question (but not the answers) before reading the paragraph. The total number of trials is 19,440, split equally across the two reading regimes. This corresponds to 40 responses per question, 20 for each regime–paragraph combination. The total number of word
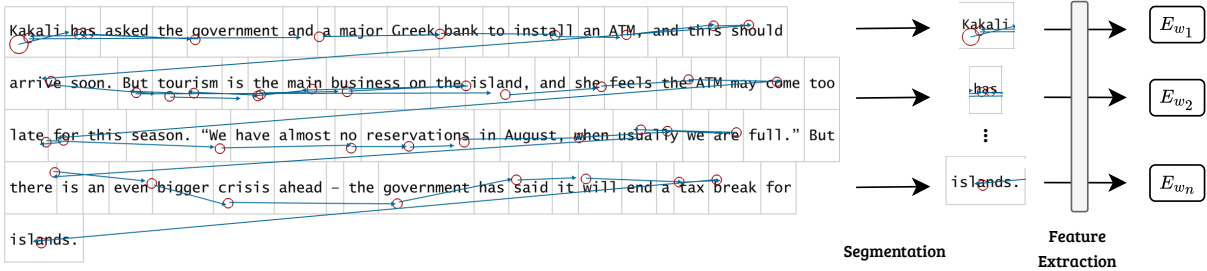
Figure 1: Left: an example of an eye movement trajectory over a paragraph, where red circles represent fixations, and blue arrows represent saccades. Right: a schematic depiction of word-level feature extraction, resulting in a vector E_{w_i}: an eye movements and linguistic word properties feature representation for each word.

tokens over which eyetracking data was collected in OneStop is 3,827,216.

The underlying textual materials and reading comprehension questions follow the STARC annotation framework (Berzak et al., 2020), where answer A is the correct answer, answer B is a miscomprehension of the information required to answer correctly, C refers to another part of the text that does not provide the answer to the question and D has no textual support. These answer types correspond to an ordering of the answers by degree of comprehension. Table 1 presents a summary of the framework along with answer choice statistics in the OneStop eyetracking data.

4 Tasks

4.1 Correct versus Incorrect Comprehension

The primary task we address is item-level prediction of whether a participant will respond correctly to a single question about a paragraph from the participant's eye movements over the paragraph. For each paragraph p and a corresponding question q^p, the possible answers are $Ans^{q^p} = \{a_1^{q^p}, a_2^{q^p}, a_3^{q^p}, a_4^{q^p}\}$. Note that the correct answer A and the three distractors $\{B, C, D\}$ are randomly mapped per trial to a_1 through a_4. The set of p, q^p, and optionally Ans^{q^p}, defines a *textual item* W. Given a participant S tested on item W, where the participant's eye movements over the paragraph are $Eyes_S^p$, the complete trial information is $Trial_S^W := \{W, Eyes_p^S\}$. We make W optional to allow for models that use only eye movements without the text.

The prediction problem can then be formulated as a binary classification task, we predict whether the participant will answer the question correctly. Formally, given a classifier h:

$$h : Trial_S^W \mapsto \{0, 1\} \qquad (1)$$

where 1 indicates a correct answer (A) and 0 indicates an incorrect answer ($B/C/D$).

Note that this task formulation abstracts away from the multiple-choice format. This allows assessing comprehension without depending on the format of the subsequent assessment task (e.g. answer choice, answer production), nor its details such as the number of answer choices and their specific content in the multiple-choice format. The combination of these task characteristics enables applying prior models from the literature, all of which predict a binary outcome without taking into account the answers, and some of which use only eye movements without the text.

4.2 Specific Answer Choice

We further address a task that takes advantage of the multiple-choice assessment format. In this task, given the answers, we predict which *specific answer* the participant will select:

$$h : Trial_S^W \mapsto \{a_1, a_2, a_3, a_4\} \qquad (2)$$

5 Models

We introduce three new models, RoBERTa-QEye, MAG-QEye and PostFusion-QEye, all of which combine text and eye movements information, and rely on the transformer language model encoder. Specifically, we use the RoBERTa$_{\text{LARGE}}$ model (Liu et al., 2019). Each of these models uses a different strategy for combining text with eye movements. RoBERTa-QEye augments the textual input with additional eye movement features. MAG-QEye uses eye movement information to modify contextualized word representations at intermediate layers of the language model. PostFusion-QEye processes text and eye movements separately and then combines them via cross-attention mechanisms. We further adjust a number of prior models
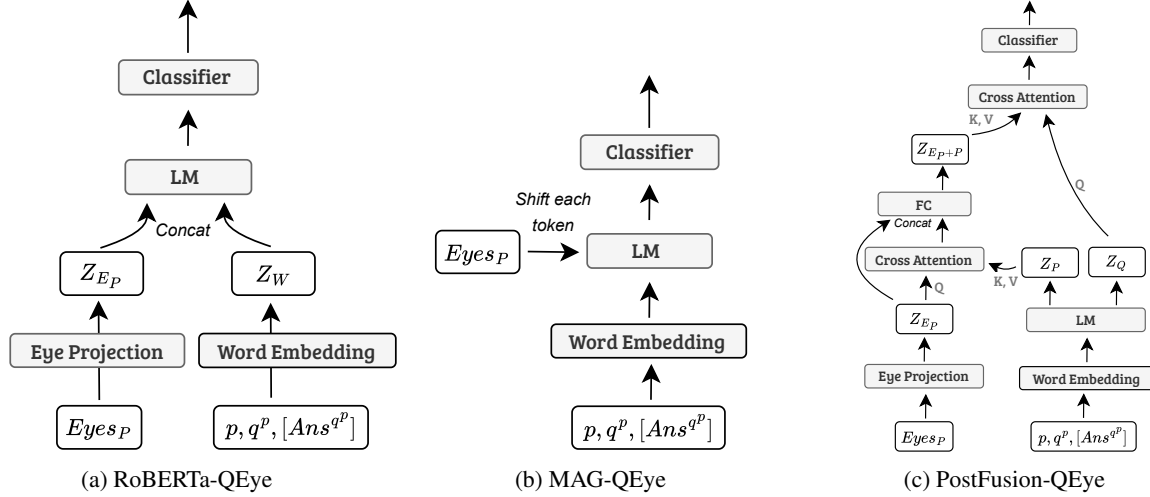
Figure 2: Model architectures. (a) RoBERTa-QEye treats eye movements as additional input features. (b) MAG-QEye uses eye movement information to modify contextualized word representations. (c) PostFusion-QEye processes text and eye movements separately and combines them via cross-attention mechanisms. Model input: $Eyes^P$ represents the participant's eye movements over the paragraph p, q^p is a question and $[Ans^{q^p}]$ are optional answer choices which are provided only in the multiple choice version of the task.

from the literature for the single-item reading comprehension prediction task.

Eye Movement Feature Representations The eyetracking record is commonly represented as a scanpath consisting of fixations (periods in which the gaze position is stable) and saccades (rapid transitions between fixations). The examined models represent this information in three different ways, in increasing level of granularity:

- **Global**: Summarizing fixation and saccade information across all the words in the input.

- **Words**: Summarizing fixation and saccade information for each word.

- **Fixations**: Accounting for each fixation and its preceding and following saccade.

Our new models focus on the word and fixation level approaches, using a variety of eye movement measures from the psycholinguistic literature. As reading times are known to be affected by linguistic word properties such as predictability, frequency, and length (Rayner et al., 2004; Kliegl et al., 2004; Rayner et al., 2011), which are not directly encoded in word embeddings, we further add such properties to the eye movement representations to allow the models to learn eye movements-word property interactions. The strength of such interactions has been shown to be indicative of the readers' linguistic proficiency (Berzak et al., 2018; Berzak and Levy, 2023), which is directly related to reading

comprehension. The eye movement and linguistic word property features used in all the models are listed in Appendix A. Note that two different feature sets are used for representing eye movements at the word and fixation levels. Figure 1 presents an example of an eye movement trajectory over a paragraph and a schematic visualization of the word-level feature extraction approach.

5.1 RoBERTa-QEye

RoBERTa-QEye incorporates eye movements as additional input sequences to RoBERTa by projecting them to the word embedding space. An overview of the architecture is presented in Figure 2a. The model is implemented in two variants, RoBERTa-QEye-Words which has a word-level feature representation and RoBERTa-QEye-Fixations, which uses a fixation-level representation. Both variants combine a textual input Z_W with eye movements input Z_{E_P}.

The textual representation Z_W is the word embedding sequence [CLS; p; SEP; q^p; $[Ans^{q^p}]$; SEP], where p is the paragraph, q^p is the question, $[Ans^{q^p}]$ are optional answers, and SEP is a separator token. The eye movement representation for the paragraph $Z_{E_P} = [Z_{E_{w_1}}, ..., Z_{E_{w_n}}]$ consists of a representation for each fixation or word i as:

$$Z_{E_{w_i}} = \text{FC}(E_{w_i}) + \text{Emb}_{\text{pos}}(i) + \text{Emb}_{\text{eye}} \quad (3)$$

where E_{w_i} are the eye movement and word property features and FC is a fully connected layer

projecting this feature representation to the word embedding space. $\text{Emb}_{\text{pos}}(i)$ is the positional embedding of the i-th word or fixation, initialized to the model's original positional embedding, which ties the eye movement representation to its respective word index. Emb_{eye} is an additional learnable embedding marking the presence of eye movement information. Z_{E_P} is concatenated with the word embedding representation Z_W, separated by a special token SEP_E, initialized as SEP. The combined sequence $[Z_{E_P}; \text{SEP}_E; Z_W]$ is passed through the transformer encoder language model. The resulting CLS token is then provided to a multilayer perceptron for response prediction.

5.2 MAG-QEye

MAG-QEye, shown in Figure 2b, modifies the transformer encoder's hidden word representations based on eye movement information. It is an adaptation of the MAG architecture (Rahman et al., 2020) originally developed for multimodal sentiment analysis. The goal of this model is to emphasize or de-emphasize words based on their respective eye movement features. Formally, for a given model layer k, each hidden token representation in the paragraph $Z_{W_i}^k$ is shifted by H_{W_i}:

$$\bar{Z}^k{}_{W_i} = Z_{W_i}^k + \alpha H_{W_i} \qquad (4)$$

where H_{W_i} is a scaled version of eye movements E_{W_i} transformed into the word embedding space. The final resulting CLS token is passed through a multilayer perceptron classifier. Appendix B.1 provides a detailed description of the architecture.

5.3 PostFusion-QEye

PostFusion-QEye, outlined in Figure 2c, processes text and eye movements separately and combines their representations through two cross-attention mechanisms. The primary objective of these mechanisms is to transform both text and eye movement data into a unified space, which we refer to as the *reading space* while taking into account the reading comprehension prediction task.

The input paragraph is passed through a language model to obtain contextualized embeddings Z_P. The eye movement input features are processed through two 1D convolution layers, resulting in the eye movement representation Z_{E_P}. Cross-attention is then applied between the paragraph embedding Z_P and Z_{E_P}, with eye movements as the query and text embeddings as the key

and the value. This step modifies the paragraph words based on the eye movements. The output is provided along with Z_{E_P} to a fully connected layer, yielding Z_{E_P+P}, a projection of the two into a shared space. Another cross-attention layer is applied between Z_{E_P+P} as key and value and the question embedding Z_Q as query, weighting the shared representation by the relevance to the question. The output of this step is passed to a multilayer perceptron classifier to predict the response.

5.4 Multiple-Choice Variants

For the specific-answer prediction task, we add to the model input the answer choices: $[a_1^{q^p}, a_2^{q^p}, a_3^{q^p}, a_4^{q^p}]$. The answer choices are provided to the model in a randomized order, as presented to the participants.

5.5 Baseline Models

We compare the proposed models to a number of eye movement models from prior work. We focus on models that were either designed for reading comprehension prediction or can be adjusted to the binary task with minimal modifications. As none of the prior models allow encoding of answers, we cannot apply them to the multiple-choice task.

Logistic Regression (Mézière et al., 2023b) Based on Mézière et al. (2023b) who used linear regression for reading comprehension prediction. We use the same feature set which includes reading speed, and global averages of standard eye movement measures.

CNN (Ahn et al., 2020) Similarly to Mézière et al. (2023b), this model is based only on eye movement information, without the underlying text. It uses the fixation sequence, represented by x and y coordinates on the screen, fixation durations, and pupil size, which are passed through a Convolutional Neural Network (CNN) to predict a binary comprehension outcome.

BEyeLSTM (Reich et al., 2022) A model for predicting reading comprehension from eye movements which represents both the fixation sequence and text features, combining LSTMs with affine transformations. BEyeLSTM outperforms the CNN model of Ahn et al. (2020), on the high versus low comprehension task with SB-SAT.

Eyettention (Deng et al., 2023) This model was originally developed for scanpath prediction. Eyettention is a word sequence encoder and a fixation

sequence encoder that uses a pre-trained BERT (Devlin et al., 2019) and an LSTM (Hochreiter and Schmidhuber, 1997), with a cross-attention mechanism for the alignment of the input sequences. We adjust this model for prediction of reading comprehension by using global cross-attention instead of windowed attention, and represent the scanpath using the last hidden representation. Further details on this model are provided in Appendix B.

5.6 No Eye Movements Baselines

We further introduce two baselines with no eye movements. The first is a majority class baseline. The second is **Text-only RoBERTa**. This baseline is of special importance as it is able to take into account item difficulty as reflected in the item textual characteristics and the distribution of item responses in the training data. To our knowledge, no previous reading comprehension prediction method was benchmarked against this kind of baseline.

6 Experimental Setup

We evaluate the models in three evaluation regimes that test different aspects of model generalization.

- **New Participant**: No eyetracking data is available for the given participant, but eyetracking data from other participants is available for the given item (paragraph).

- **New Item**: No eyetracking data is available for the item, but prior eyetracking data is available for the participant on other items.

- **New Item & Participant**: No prior eyetracking data is available for the participant nor for the item.

We further report aggregated results across all three regimes.

We perform model training, hyperparameter tuning, and evaluation separately for the ordinary reading and information seeking parts of the data, with 10-fold cross-validation. Figure 3 presents schematically one of the 10 data splits for a 10-article 60-participant batch. A full data split for a reading regime (ordinary reading or information seeking) is the union of three such splits. In each split, approximately 64% of the data is allocated for training, 17% for validation, and 19% for testing. The test data is further divided into 9% in the New Participant, 9% New Item, and 1% New Item & Participant regimes. In total across the
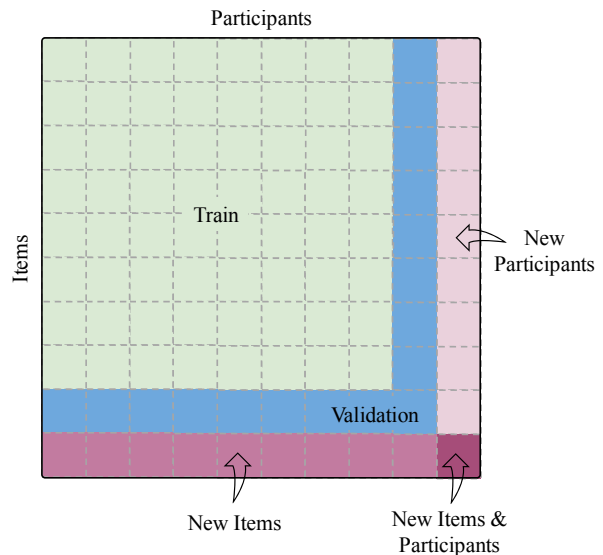

Figure 3: A schematic depiction of a 10-article 60-participant batch split, divided into a train set, a validation set, and the three test sets. A full data split for a reading regime (ordinary reading or information seeking) consists of the union of three batch splits.

10 splits, approximately 90% of the trials in the dataset appear in each of the New Participant and New Item evaluation regimes, and 10% in the New Item & Participant regime. Items are assigned to the train, validation and test portions of each split at the *article level*, such that no article is split across different data portions, ensuring generalization to items whose content is unrelated to items seen in training. See Appendix C for further information on the splits.

Because the data is unbalanced across classes, we use balanced accuracy as the evaluation metric. As prior work has shown considerable differences in reading behavior between the ordinary reading and information seeking reading conditions (Hahn and Keller, 2023; Malmaud et al., 2020; Shubi and Berzak, 2023), we train and evaluate the models on each type of trials separately. We perform hyperparameter tuning for each split, and report balanced accuracy results on the aggregation of the predictions across the 10 test sets. We assume that at test time the evaluation regime of the trial is *unknown*. Model hyperparameter tuning is therefore based on the entire validation set of the split. As prior models from the literature were developed for different tasks and on different datasets, we run a hyperparameter search for each model over a search space that includes the original parameter settings. Hyperparameters are also optimized for the Text-only RoBERTa baseline. To address the

Binary Reading Comprehension			Ordinary Reading (Gathering)				Information Seeking (Hunting)			
Model	Gaze Representation	Text Representation	New Item	New Participant	New Item & Participant	All	New Item	New Participant	New Item & Participant	All
Majority	None	None	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Text-only RoBERTa	None	Emb	54.8	63.1	55.2	58.7	51.8	63.1	50.5	57.1
Log. Reg. (Mézière et al., 2023b)	Global	None	53.3	50.8	53.8	52.2	53.2	52.2	52.3	52.7
CNN (Ahn et al., 2020)	Fixations	None	51.0	51.0	51.9	51.1	51.4	51.3	49.2	51.2
BEyeLSTM (Reich et al., 2022)	Fixations	Ling. Feat.	50.6	55.7	51.1	53.0	50.5	55.1	**55.1**	53.0
Eyettention (Deng et al., 2023)	Fixations	Emb + Word Len.	54.8	60.4	**57.1**	57.6	50.5	56.4	52.3	53.4
RoBERTa-QEye	Words	Emb + Ling. Feat.	**55.5**	63.5	52.1	59.1	50.5	**63.8**	51.0	56.8
RoBERTa-QEye	Fixations	Emb + Ling. Feat.	53.3	61.3	**57.1**	57.3	50.3	60.3	50.8	55.1
MAG-QEye	Words	Emb + Ling. Feat.	54.8	**64.1***	53.8	**59.2**	52.5	62.3	51.3	57.1
PostFusion-QEye	Fixations	Emb + Ling. Feat.	54.8	63.5	55.0	58.9	**53.8***	62.7	53.8	**58.0**

Table 2: Results on balanced accuracy for the main binary reading comprehension prediction task (correct vs incorrect comprehension). 'All' denotes results for the aggregation of all the trials across the three test regimes. 'Emb' stands for word embeddings, 'Ling. Feat.' for linguistic word properties. Statistically significant improvements over the Text-only RoBERTa baseline, using a paired bootstrap test, chosen based on considerations described in (Dror et al., 2018), are marked with '*' at $p < 0.05$.

Multiple-Choice Reading Comprehension			Ordinary Reading (Gathering)				Information Seeking (Hunting)			
Model	Gaze Representation	Text Representation	New Item	New Participant	New Item & Participant	All	New Item	New Participant	New Item & Participant	All
Majority	None	None	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Text-only RoBERTa	None	Emb	25.3	**33.0**	25.2	29.0	25.0	**31.7**	24.8	28.2
RoBERTa-QEye	Words	Emb + Ling. Feat.	28.2*	31.5	32.1**	29.9	28.9***	30.1	27.1	29.3
RoBERTa-QEye	Fixations	Emb + Ling. Feat.	29.2*	32.9	28.1	**30.9**	30.3***	31.0	**29.5**	**30.5***
MAG-QEye	Words	Emb + Ling. Feat.	27.9***	32.5	30.4***	30.2**	26.8	30.0	29.0	28.4
PostFusion-QEye	Fixations	Emb + Ling. Feat.	**29.4**	31.7	**32.9***	30.6*	27.5*	27.9	26.7	27.6

Table 3: Results on balanced accuracy for the multiple-choice specific answer prediction task. Statistically significant improvements over the Text-only RoBERTa baseline, using a paired bootstrap test, are marked with '*' at $p < 0.05$, '**' at $p < 0.01$ and '***' at $p < 0.001$. We note that in some cases, higher balanced accuracy scores correspond to lower p-values due to higher variability in the predictions of the minority classes.

unbalanced nature of the data, shown in Table 1, we sample the same number of trials from each answer class during training. Additional details on feature normalization, model training, hyperparameter search, and number of model parameters are provided in Appendix D.

7 Results

7.1 Correct vs Incorrect Comprehension

In Table 2, we present trial-level reading comprehension prediction results for ordinary reading and information seeking. The best results are achieved by different models under the different evaluation regimes. MAG-QEye achieves the highest overall balanced accuracy in ordinary reading with a score of 59.2, while PostFusion-QEye performs best in information seeking, with a score of 58.0. In all the evaluation regimes, the best performing model outperforms the Text-only RoBERTa baseline. In all but the New Item & Participant evaluation regime, the best performing model is one of our proposed models. Text-only RoBERTa turns out to be a key benchmark, whereby most models are below this baseline especially in the New Participant regime.

We note several key trends in the results. First, results in the New Participant regime tend to be higher than in the New Item regime, highlighting the importance and the challenge of generalization to new items. The strong performance of the RoBERTa text-only baseline in the New Participant regime suggests that much of the gains in this regime do not stem from eye movement information, but rather from item properties and statistics. This highlights the importance of benchmarking against such a baseline for assessing the contribution of eye movement information. It further underscores the importance of explicit representation of the text; the Logistic Regression, CNN and BEyeLSTM models, which do not include such a representation, perform poorly in the New Participant regime. Finally, for any given model, the ordinary reading regime tends to yield higher accuracies compared information seeking. We hypothesize that this difference could be related to higher variability in reading strategies in information seeking across participants (Shubi and Berzak, 2023). We leave a detailed investigation of this hypothesis to future work.

7.2 Multiple-Choice Task

In Table 3 we use our models, MAG-QEye and PostFusion-QEye, and the two RoBERTa-QEye variants to predict participants' specific answer response among the four provided answers. As mentioned above, prior models from the literature are not applicable for this task. We find that all the models outperform the Text-only RoBERTa baseline in the two regimes that involve new items, but not in the New Participant regime. The best performing model in the overall evaluations is RoBERTa-QEye-Fixations. The general trends regarding higher performance in the New Participant regime compared to the New Item regime, as well as the stronger within-model performance in ordinary reading compared to information seeking, extend to this evaluation.

7.3 Additional Experiments

We perform two additional sets of experiments of preliminary nature. In Appendix E we provide ablation experiments on the effect of linguistic word properties on model performance. In Appendix F we further examine different variants of the textual backbone of the models. Finally, we provide validation set results in Appendix G.

8 Summary and Discussion

This paper presents a systematic evaluation of the ability to predict reading comprehension from eye movements in reading at the level of a single question over a single paragraph. We address this task using a range of existing and new models, applied to large scale data across several task variants and evaluation regimes. Our experiments indicate that the task at hand is highly challenging, and further highlight the importance of text-only baselines for assessing the added value of eye movements information. However, we do find that small improvements over a strong text-only baseline are achievable with the proposed and some of the past modeling approaches.

Given the presented results, the extent to which specific aspects of reading comprehension can be reliably decoded from eye movements signal remains an open question. It is possible that eye movements simply do not contain sufficient information for decoding comprehension at high accuracy rates for the examined level of granularity. Alternatively, it may be the case that current modeling techniques do not represent or process eye

movements data effectively enough for this task. Another factor whose role in task difficulty needs to be investigated in more detail is the imbalanced nature of the data, where only a relatively small fraction of the responses are incorrect.

Additional work on eye movement data analysis, new model architectures, feature representations and training regimes is needed for making further progress on this task. Additionally, new datasets with other task variants and other populations such as children and L2 readers are required to study the problem in a more comprehensive manner. We envision that the models, tasks, evaluation protocols, and data presented here will serve as a stepping stone for such work, as well as a broader scientific investigation of the relations between eye movements and reading comprehension.

9 Ethical Considerations

The eyetracking data used in this work 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. Analyses of the relations between eye movements and reading comprehension, and predictive models of comprehension are among the primary use cases for which the data was collected.

Automatic reading comprehension assessments from eye movements can potentially address shortcomings of standard assessment methodologies by reducing test development and test taking costs, and enhancing test availability. However, they also introduce potential risks for biased and inaccurate assessments that may put various populations and individuals at a disadvantage. These include non-native speakers, older participants, participants with cognitive impairments, disabilities, eye conditions and others. Much higher model performance than the current state-of-the-art and a thorough examination of potential biases due to factors unrelated to reading comprehension are needed before considering deploying such assessments.

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. We further emphasize that future reading comprehension assessment systems are to be used only with explicit consent from potential users to have their eye movements collected and analyzed for this purpose.

10 Limitations

Our work has a number of limitations which are related to the experimental design of OneStop. First, the textual data consists of articles with 4-7 paragraphs. Each question is over the content of a single paragraph. Longer and shorter texts, as well as questions that require integration of information from several paragraphs, are not covered. The experimental design does not allow participants to go back and forth between the question and passage, which is common in question answering tasks. Further, participant expectations for upcoming reading comprehension questions, as well as the setting of an in-lab experiment may result in reading patterns that deviate from reading in everyday settings (Huettig and Ferreira, 2022) and could impact the predictive performance of the model.

While our work examines the feasibility of automated assessment of reading comprehension from eye movements, the accuracy of the models presented is still very far from being relevant for deployment in real world scenarios. Our results are further limited to the equipment at hand. Our approach has only been tested using a state-of-the-art eyetracker (Eyelink 1000 Plus) at a sampling rate of 1000Hz. This allows extracting gaze position and duration at a very high temporal resolution and character-level precision. While studies such as Ishimaru et al. (2017) and Chen et al. (2023) have demonstrated predictive modeling capabilities using lower spatial and temporal resolution eye tracking systems, additional work is required to test the feasibility of reading comprehension prediction using such equipment.

Although we use the largest eyetracking for reading comprehension dataset to date, OneStop was collected from adult L1 English speakers, with no cognitive impairments, and in the large majority of cases no eye conditions. We acknowledge that this pool of participants excludes multiple populations, including children, elderly, participants with cognitive and physical impairments and others. Future data collection and analysis work is required to test the generalization capabilities and potential biases of the models in other populations.

In this work we assume the availability of both suitable eyetracking data and a pretrained language model for the language at hand. Although language models for lower-resource languages (e.g. Chriqui and Yahav, 2022; Vamvas et al., 2023) and multilingual models (e.g. Lai et al., 2023) have been made

available, many languages still lack such models. Similarly, to the best of our knowledge, no eye-tracking data with a substantial reading comprehension component is currently available for languages other than English. This limits the generality of the results. More eyetracking data collection and language model development work is required to include additional languages.

Acknowledgments

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 Applications*, ETRA '20 Short Papers, New York, NY, USA. Association for Computing Machinery.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735. Association for Computational Linguistics.

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. CELER: A 365-participant corpus of eye movements in L1 and L2 English reading. *Open Mind*, 6:1–10.

Xiuge Chen, Namrata Srivastava, Rajiv Jain, Jennifer Healey, and Tilman Dingler. 2023. Characteristics of Deep and Skim Reading on Smartphones vs. Desktop: A Comparative Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Hamburg Germany. ACM.

Avihay Chriqui and Inbal Yahav. 2022. HeBERT and HebEMO: A Hebrew BERT Model and a Tool for Polarity Analysis and Emotion Recognition. *INFORMS Journal on Data Science*, 1(1):81–95.

Leana Copeland, Tom Gedeon, and Balapuwaduge Mendis. 2014. Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research*, 3.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Shuwen Deng, David R. Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A. Jäger. 2023. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. In *Proceedings of the ACM on Human-Computer Interaction*, pages 1–24. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1383–1392.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

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

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

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

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. Publisher: SAGE Publications Inc.

Shoya Ishimaru, Kensuke Hoshika, Kai Kunze, Koichi Kise, and Andreas Dengel. 2017. Towards reading trackers in the wild: detecting reading activities by EOG glasses and deep neural networks. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, pages 704–711, New York, NY, USA. Association for Computing Machinery.

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

Marcel Adam Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology - EUR J COGN PSYCHOL*, 16:262–284.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

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

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. ArXiv:1908.03557 [cs].

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*.

Silvia Makowski, Lena A Jäger, Ahmed Abdelwahab, Niels Landwehr, and Tobias Scheffer. 2019. A discriminative model for identifying readers and assessing text comprehension from eye movements. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 209–225. Springer.

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

Pascual Martínez-Gómez and Akiko Aizawa. 2014. Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, page 95–104, New York, NY, USA. Association for Computing Machinery.

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

Diane C. Mézière, Lili Yu, Erik D. Reichle, Genevieve McArthur, and Titus von der Malsburg. 2023a. Scanpath regularity as an index of reading comprehension. *Scientific Studies of Reading*.

Diane C. Mézière, Lili Yu, Erik D. Reichle, Titus von der Malsburg, and Genevieve McArthur. 2023b. 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.

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

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

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

Keith Rayner, Elizabeth R Schotter, Michael EJ Masson, Mary C Potter, and Rebecca Treiman. 2016. So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest*, 17(1):4–34.

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

David R. 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 *Symposium on Eye Tracking Research and Applications*, ETRA '22. Association for Computing Machinery.

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

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

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Rosy Southwell, Julie Gregg, Robert Bixler, and Sidney K D'Mello. 2020. What eye movements reveal about later comprehension of long connected texts. *Cognitive Science*, 44(10):e12905.

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

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. Swissbert: The multilingual language model for switzerland. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 54–69.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal Learning With Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Duo Yang and Nora Hollenstein. 2023. PLM-AS: Pretrained Language Models Augmented with Scanpaths for Sentiment Classification. *Proceedings of the Northern Lights Deep Learning Workshop*, 4.

Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. 2023. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325.

A   Features

Feature Name	Description
Word-Level Eye Movement Features	
IA_DWELL_TIME	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 / TRIAL_DWELL_TIME).
IA_FIXATION_%	Percentage of all fixations in a trial falling in the current interest area.
IA_FIXATION_COUNT	Total number of fixations falling in the interest area.
IA_REGRESSION_IN_COUNT	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_RUN_COUNT	Number of times the Interest Area was entered and left (runs).
IA_FIRST_FIX_PROGRESSIVE	Checks whether the first fixation in the interest area is a first-pass fixation.
IA_FIRST_FIXATION_DURATION	Duration of the first fixation event that was within the current interest area
IA_FIRST_FIXATION_VISITED_IA_COUNT	This reports the number of different interest areas visited so far before the first fixation is made to the current interest area.
IA_FIRST_RUN_DWELL_TIME	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_FIRST_RUN_FIXATION_COUNT	Number of all fixations in a trial falling in the first run of the current interest area.
IA_SKIP	An interest area is considered skipped (i.e., IA_SKIP = 1) if no fixation occurred in first-pass reading.
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_PATH_DURATION	The summed fixation duration from when the current interest area is first fixated until the eyes enter an interest area with a higher IA_ID.
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.
IA_SELECTIVE_REGRESSION_PATH_DURATION	Duration of fixations and refixations of the current interest area before the eyes enter an interest area with a higher ID.
IA_LAST_FIXATION_DURATION	Duration of the last fixation event that was within the current interest area.
IA_LAST_RUN_DWELL_TIME	Dwell time of the last run (i.e., the sum of the duration of all fixations in the last run of fixations within the current interest area).
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_PUPIL	Average pupil size during the current fixation.
CURRENT_FIX_X	X coordinate of the current fixation.
CURRENT_FIX_Y	Y coordinate of the current fixation.
NEXT_FIX_ANGLE, PREVIOUS_FIX_ANGLE	Angle between the horizontal plane and the line connecting the current fixation and the next/previous fixation.
NEXT_FIX_DISTANCE, PREVIOUS_FIX_DISTANCE	Distance between the current fixation and the next/previous fixation in degrees of visual angle.
NEXT_SAC_AMPLITUDE	Amplitude of the following saccade in degrees of visual angle.
NEXT_SAC_ANGLE	Angle between the horizontal plane and the direction of the next saccade.
NEXT_SAC_AVG_VELOCITY	Average velocity of the next saccade.
NEXT_SAC_DURATION	Duration of the next saccade in milliseconds.
NEXT_SAC_PEAK_VELOCITY	Peak values of gaze velocity (in visual degrees per second) of the next saccade.

Table 4: 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 5: Linguistic word properties and their descriptions. POS tags and parse trees were obtained using SpaCy (Honnibal et al., 2020).

B Adaptations of Prior Models

B.1 MAG

We replace the vision and acoustic input with word-level eye movement features. To align them with the tokenized text, we duplicate the word-level features for each subword token. Additionally, for a fair comparison with other models, we replace BERT with RoBERTa$_{\text{LARGE}}$ as the textual backbone model.

Formally, each token embedding Z_i is *displaced* by H_i.

$$\bar{Z}_i = Z_i + \alpha H_i \tag{5}$$

H_i is a scaled and transformed version of the eye movements E_i,

$$H_i = g_i \cdot (W_e E_i) + b_H \tag{6}$$

where the scaling is defined by,

$$g_i = ReLU(W_g[Z_i; A_i] + b_g) \tag{7}$$

The amount of displacement is defined by

$$\alpha = min(\frac{||Z_i||_2}{||H_i||_2}\beta, 1) \tag{8}$$

where β is a hyper-parameter, and W_e, W_g, b_H, b_g are learned.

Finally, the contextualized CLS token is used for classification.

B.2 Eyettention

We adjust the prediction objective of the model from next fixation to trial-level classification. To this end, we use global cross attention between the word sequence and the scanpath sequence instead of fixed window cross attention, as suggested in Deng et al. (2023). We then represent the whole scanpath using the last hidden representation of the scanpath LSTM. We further replace BERT, with RoBERTa$_{\text{LARGE}}$ for consistency with the other models.

B.3 BEyeLSTM

First, we employ SpaCy tokenization based on paragraph-level input rather than word-level input, resulting in a more precise tokenization. Second, the textual materials used here include a more fine-grained set of part-of-speech tags and named entities, which results in a larger final feature set. Lastly, we omit the "words in fixed context on unigrams" feature, as it presupposes that all the participants read the same texts, which is not the case in OneStop.

B.4 CNN

Ahn et al. (2020) resort to artificially subdividing SB-SAT texts into smaller segments in order generate a sufficient number of training examples to make the dataset usable for their task of predicting low versus high comprehension over multiple items. This heuristic is problematic in general, and not applicable to the single item task addressed here. In the current work we use the entire fixation sequence as the input to the model.

C   Cross Validation Splits

Each split guarantees an equal number of participants from each OneStopQA batch in each portion of the split, and is approximately stratified by answer type. Recall that each participant is presented with a specific combination of a paragraph and one of its three associated questions. Due to the stratification by answer type, it is not guaranteed that the appearances of any given paragraph will be balanced across the three possible questions in any of the split portions. Note that across the 10 test sets, not all participant – item combinations are covered in the test sets, as this would require 100 data splits.

D   Feature Standardization and Hyperparameter Tuning

We apply standardization for each feature in E_P, where the statistics 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). The search space for learning rates is $\{0.00001, 0.00003, 0.0001\}$ and for dropout $\{0.1, 0.3, 0.5\}$.

- For **Logistic Regression**, we search over regularization parameter C values of $\{0.1, 5, 10, 50, 100\}$, with and without an L2 penalty.

- For the **CNN** we include a learning rate of 0.001 as in Ahn et al. (2020).

- Following (Reich et al., 2022), for **BEyeLSTM** the search space for learning rates is $\{0.001, 0.003, 0.01\}$, embedding dimensions of $\{4, 8\}$ and hidden dimension of $\{64, 128\}$.

- For **Eyettention** we also include a learning rate of 0.001 and dropout of 0.2, as in Deng et al. (2023).

- For **MAG-QEye**, the search space for the injection layer index is $\{0, 11, 23\}$. We set the MAG-internal dropout to 0.5, and the scaler parameter to 1e-3, as suggested by (Rahman et al., 2020).

- In **PostFusion-QEye**, the 1D convolution layers have a kernel size of three, stride 1, and padding 1.

All neural networks are trained using the Pytorch Lighting library (Falcon and The PyTorch Lightning team, 2019; Paszke et al., 2019) and evaluated using torch-metrics (Nicki Skafte Detlefsen et al., 2022) on a NVIDIA A100-40GB and A40-48GB GPUs. We adapt Huggingface's RoBERTa implementation (Wolf et al., 2020). The baselines described in Section 5.5 are reimplemented in this framework as well. A single training epoch took approximately 5 minutes. We train for a maximum of ten epochs, stopping after three epochs without improvement on the validation set.

The number of model parameters is 355M for the RoBERTa$_{\text{LARGE}}$ backbone, and an additional 1.1M for MAG-QEye and RoBERTa-QEye, and 9M for PostFusion-QEye.

E   The Role of Linguistic Word Property Features

Our proposed models tend to outperform the Text-only RoBERTa baseline, especially in the two evaluation regimes that involve new items. Note however, that in addition to eye movements, these models also include linguistic word properties, which may provide information on the textual item that is not fully encoded in word embeddings. Some of them (e.g. word length, frequency and surprisal) are also known to be predictive of reading times.

What is the effect of these features on model performance? To examine this question, we carry out two ablation experiments. In the first experiment, we ablate the linguistic word property features. In the second experiment we ablate the eye movement features. The latter ablation is not possible with fixation based models, because even with the eye movement features removed, these models still have information about the gaze trajectory through the order and word identity of the fixations. We therefore perform these experiments only with the word based models RoBERTa-QEye-Words and MAG-QEye.

Table 6 in Appendix E presents the ablation results for the binary task. In the first experiment, removal of linguistic word properties does not substantially affect model performance. This outcome does not match our expectation regarding the potential benefits of allowing models to learn eye movement – linguistic word property interactions. In the second experiment, overall, we again do not observe performance degradation when ablating the eye movement features. While this experiment is not sufficient for drawing general conclusions regarding the value of eye movement information for our task, it suggests that in our two instances of word-based models, eye movements do not seem to provide substantial performance gains above and beyond features that can be readily extracted from the text. We leave a more extensive investigation regarding the impact of linguistic features on model performance to future work.

Binary Reading Comprehension	Gathering Trials				Hunting Trials			
Model	New Item	New Participant	New Item & Participant	All	New Item	New Participant	New Item & Participant	All
Text-only RoBERTa	54.8	63.1	55.2	58.7	51.8	63.1	50.5	57.1
MAG-QEye	54.8	**64.1***	53.8	59.2	**52.5**	62.3	51.3	57.1
MAG-QEye w/o Ling. Feat	55.9	63.8	55.5	59.6	52.3	63.3	**54.8**	**57.7**
MAG-QEye w/o Eyes	54.2	63.7	56.7	58.8	51.9	63.3	53.8	57.4
RoBERTa-QEye-Words	55.5	63.5	52.1	59.1	50.5	**63.8**	51.0	56.8
RoBERTa-QEye-Words w/o Ling. Feat	55.4	63.3	56.3	59.2	51.1	62.7	50.7	56.6
RoBERTa-QEye-Words w/o Eyes	**56.7***	63.7	**57.5**	**60.0****	49.3	63.2	51.2	56.0

Table 6: The effect of ablating word-level eye movement features (Table 4) and linguistic word properties (Table 5) on balanced accuracy for binary classification of the word based models MAG-QEye and RoBERTa-QEye-Words. Statistically significant improvements over Text-only RoBERTa, using a paired bootstrap test, are marked with '*' at $p < 0.05$, '**' at $p < 0.01$ and '***' at $p < 0.001$.

F   Textual Backbone Variants

Our models use RoBERTa as a textual backbone model, and the parameters of this backbone are subjected to change during model training. Other choices for this model component are possible. For example, one can pre-train the model on multiple choice question answering, freeze the textual backbone parameters during model training, or choose a different textual backbone model altogether. Preliminary experiments with MAG-QEye in Appendix F Table 7 do not show a consistent effect of these choices on model performance in the main prediction task. We leave a comprehensive investigation of textual backbone model choice and training to future work.

Binary Reading Comprehension	Gathering Trials				Hunting Trials			
MAG-QEye Backbone	New Item	New Participant	New Item & Participant	All	New Item	New Participant	New Item & Participant	All
RoBERTa Large	**54.8**	64.1	53.8	59.2	**52.5**	62.3	51.3	**57.1**
RoBERTa Large Frozen	54.3	61.4	51.4	57.5	51.9	60.0	**53.3**	55.8
RoBERTa Large Trained for QA on RACE	**54.8**	**64.6**	52.7	**59.3**	48.3	62.7	44.9	54.9
RoBERTa Base	52.8	64.0	**56.9**	58.3	50.8	**63.5**[*]	51.6	56.9

Table 7: Balanced accuracy performance comparison of different backbone architectures and training strategies for MAG-QEye. Statistically significant improvements compared to an unfrozen RoBERTa Large backbone are marked with '*' at $p < 0.05$, '**' at $p < 0.01$ and '***' at $p < 0.001$ using a paired bootstrap test.

Binary Reading Comprehension			Ordinary Reading (Gathering)				Information Seeking (Hunting)			
Model	Gaze Representation	Text Representation	New Item	New Participant	New Item & Participant	All	New Item	New Participant	New Item & Participant	All
Majority	None	None	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Text-only RoBERTa	None	Emb	59.8	**65.8**	57.9	62.5	57.1	65.1	56.8	60.8
Log. Reg. (Mézière et al., 2023b)	Global	None	53.4	51.1	53.9	52.3	51.8	53.0	51.9	52.4
CNN (Ahn et al., 2020)	Fixations	None	53.3	53.7	53.4	53.5	55.1	54.5	55.0	54.8
BEyeLSTM (Reich et al., 2022)	Fixations	Ling. Feat.	55.0	58.5	55.7	56.7	57.3	58.6	58.3	58.0
Eyettention (Deng et al., 2023)	Fixations	Emb + Word Len.	58.5	62.4	57.9	60.3	57.0	59.5	56.9	58.2
RoBERTa-QEye	Words	Emb + Ling. Feat.	57.0	65.5	**60.5**	61.2	55.3	64.7	52.2	59.6
RoBERTa-QEye	Fixations	Emb + Ling. Feat.	57.0	63.5	60.4	60.3	54.6	62.4	56.5	58.4
MAG-QEye	Words	Emb + Ling. Feat.	**60.4**	**65.8**	58.9	**62.9**	57.3	**66.0**	**59.5**	61.6
PostFusion-QEye	Fixations	Emb + Ling. Feat.	60.1	65.2	60.4	62.5	**58.3**	65.8	59.3	**61.9**[*]

Table 8: Balanced accuracy for the binary reading comprehension prediction task (correct vs incorrect comprehension).

Multiple-Choice Reading Comprehension			Ordinary Reading (Gathering)				Information Seeking (Hunting)			
Model	Gaze Representation	Text Representation	New Item	New Participant	New Item & Participant	All	New Item	New Participant	New Item & Participant	All
Majority	None	None	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Text-only RoBERTa	None	Emb	25.7	35.7	25.6	30.4	25.0	**34.4**	25.5	29.5
RoBERTa-QEye	Words	Emb + Ling. Feat.	34.0[***]	34.4	37.4[**]	34.3[***]	33.3[***]	34.3	32.9	33.7[*]
RoBERTa-QEye	Fixations	Emb + Ling. Feat.	33.6[***]	34.7	37.9[***]	34.3[***]	34.0[***]	34.4	37.4	34.3[***]
MAG-QEye	Words	Emb + Ling. Feat.	33.8[***]	36.1	34.3[**]	34.9[**]	34.8[***]	33.6	32.9	34.1[***]
PostFusion-QEye	Fixations	Emb + Ling. Feat.	33.2[***]	35.1	33.5[*]	34.1[**]	34.0[**]	31.8	35.4	33.0

Table 9: Balanced accuracy for the multiple-choice specific answer prediction task.