DECODING READING GOALS FROM EYE MOVEMENTS

Anonymous authors

000

001 002 003

004

005 006 007

008 009

010

011

012

013

014

015

016

017

018

019

021

023 024

025

Paper under double-blind review

Abstract

Readers can have different goals with respect to the text they are reading. Can these goals be decoded from the pattern of their eye movements over the text? In this work, we examine for the first time whether it is possible to decode two types of reading goals that are common in daily life: *information seeking* and *ordinary reading*. Using large scale eye-tracking data, we apply to this task a wide range of state-of-the-art models for eye movements and text that cover different architectural and data representation strategies, and further introduce a new model ensemble. We systematically evaluate these models at three levels of generalization: new textual item, new participant, and the combination of both. We find that eye movements contain highly valuable signals for this task. We further perform an error analysis which builds on prior empirical findings on differences between ordinary reading and information seeking and leverages rich textual annotations. This analysis reveals key properties of textual items and participant eye movements that contribute to the difficulty of the task.¹

1 INTRODUCTION

Reading is a ubiquitously practiced skill that is indispensable for successful participation in modern society. When reading, our eyes move over the text in a saccadic fashion, where there are periods of time in which the gaze is stable at a specific location, called *fixations*, and rapid transitions between fixations called *saccades* (Rayner, 1998). This sequence of fixations and saccades is generally hypothesized to contain rich information about how readers interact with text. Automatic decoding of such information is currently a growing area of research (e.g. Reich et al., 2022a; Shubi et al., 2024).

032 In daily life, a reader may have one or several *goals* that they pursue with respect to the text. For 033 example, they may read the text closely or skim it to obtain the gist of the text's content, they may proofread it, or they may be seeking specific information of interest. Each such goal can have a 035 profound impact on online linguistic processing and on the corresponding eye movement behavior while reading. Despite the many reading goals readers pursue in everyday life, research on eye 037 movements in cognitive science as well as work that integrated eye movements data in Natural Language Processing (NLP) and machine learning (ML) have primarily focused on one reading 038 regime, which can be referred to as *ordinary reading*. In this regime, the reader's goal is typically general comprehension of the text. Nearly all broad coverage datasets used in research on eye 040 movements in reading, such as Dundee (Kennedy et al., 2003), MECO (Siegelman et al., 2022) and 041 CELER (Berzak et al., 2022) were collected in this regime. Other forms of reading, although widely 042 acknowledged (Radach & Kennedy, 2004), received much less attention and remain understudied. 043

In this work, we go beyond ordinary reading and ask whether broad reading goals can be reliably decoded from the pattern of the reader's eye movements over the text. We focus on the distinction between ordinary reading and *information seeking*, a highly common reading regime in everyday life, where the reader is interested in obtaining specific information from the text. Prior work suggests that on average across participants and texts, there are substantial differences in eye movement patterns between these two reading regimes (Hahn & Keller, 2023; Shubi & Berzak, 2023). However, it is currently unknown whether there is sufficient signal in eye movements for automatic decoding of the reading goal given eye movements of a single participant over a single textual item. Furthermore, little is known about the factors that contribute to the difficulty of this task.

¹Code is available at the following anonymous link: https://anonymous.4open.science/r/ Decoding-Reading-Goals-from-Eye-Movements/. Data will be made publicly available.



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

In this work, we address this gap by conducting a series of experiments on reading goal decoding. We use a range of recent models that implement different strategies for representing eye movements 068 and combining them with text, and perform systematic evaluations of their individual and ensemble 069 generalization abilities. We demonstrate that it is indeed possible to perform this task with relatively high accuracy rates. We further perform a systematic error analysis which leverages rich data annotations to reveal key axes of variation that contribute to task difficulty.

- 073 The main contributions of this work are the following:
 - **Task**: We introduce a new decoding task: given eye movements from a single participant over a passage, predict whether they engaged in ordinary reading or in information seeking.
 - Modeling and Evaluation: We adapt and apply to this task 10 different state-of-the-art predictive models for eye movements in reading. We further introduce an ensemble model which outperforms the best single model. We characterize the generalization ability of all the models across textual items and participants.
 - Error Analysis: We use statistical modeling and detailed textual annotations to perform an error analysis which reveals interpretable axes of variation in item and participant-related properties of trials that are indicative of classification difficulty. This analysis provides new insights on the nature of the task.

TASK 2

We address the task of predicting whether a reader is engaged in ordinary reading for comprehension or in seeking specific information, based on their eye movements over the text. Let S be a participant, P a textual passage, and E_P^S the recording of the participant's eye movements over the passage. Given a ground-truth mapping $\mathcal{C}(S, P) \to \{\text{Information Seeking}, \text{Ordinary Reading}\}, we aim to$ approximate C with a classifier h: 092

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

094 Where the passage P is an optional input, such that the classifier can be provided only with the eye 095 movement data E_S^P or with both the eye movements and the underlying text P. We assume that the 096 participant has not read the paragraph previously.

The information seeking regime is operationalized by presenting the participant with a question Q098 prior to reading the passage. This question prompts the participant to seek specific information in the text. We assume that the classifier does not receive the question nor any information on the 100 participant. Figure 1 presents the task schematically. 101

102

061

062

063

064

065 066 067

071

074

075

076

077

078

079

081

082

084 085

087

090

091

093

3 MODELING

103 104

Eye movements during reading present a highly challenging case of temporally and spatially aligned 105 multimodal data, where fixations are temporal and correspond to specific words in the text. We use 106 a wide range of state-of-the-art architectures for processing this data. While the models were origi-107 nally applied to other prediction tasks, such as predicting participant's reading comprehension and 108 various participant properties, they offer general purpose mechanisms for processing eye movement 109 data during reading that can be applied to other tasks, and cover different architectural and data rep-110 resentation strategies. The models can be broadly divided along three primary axes, the modalities 111 used (eye movements-only, or eye movements and text), the granularity at which eye movement 112 information is represented (global averages across the text, single word, or a single fixation), and for the multimodal approaches, the nature of the text representations and strategy for combining them 113 with eye movements. These models are directly applicable, in some cases with small adaptations, to 114 the task of reading goal prediction. See Figures 5 and 6 in Appendix A for model diagrams. 115

116 117

3.1 EYE MOVEMENTS-ONLY MODELS

These models use only eye movement information, without taking into account the text. Such models are valuable in common scenarios where the underlying text for the eye movement recording is not available. It is also the go-to approach when the eye-tracking calibration is of low quality, leading to imprecise information on the location of fixations with respect to the text. This is a highly common situation, especially with web-based eye-tracking and lower grade eye-tracking devices. Beyond practical considerations, the eye movements-only approach allows assessing the added value of textual information for our task.

125

126 GLOBAL REPRESENTATION

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

132

134

133 FIXATION-BASED REPRESENTATION

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

140 141

142

3.2 EYE MOVEMENTS AND TEXT MODELS

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

149 The models implement three primary strategies for combining the two modalities at progressively later stages of processing: (i) in the model input, (ii) merging them within intermediate model repre-150 sentations, or (iii) with architectures that fuse the modalities using cross attention mechanisms after 151 each modality has been processed separately. Furthermore, since eye movements in reading are both 152 temporally and spatially aligned with the underlying text, the models can be categorized based on 153 how they capture this alignment: (i) by aggregating eye movements for each word, thereby focusing 154 on spatial alignment; or (ii) by aggregating eye movement information for each individual *fixation*, 155 which explicitly encodes both spatial and temporal correspondences between eye movements and 156 text. 157

158 WORD-BASED REPRESENTATIONS 159

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

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

166

167 FIXATION-BASED REPRESENTATIONS

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

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

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

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

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

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

190 191 3.3 BASELINES

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

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

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

203 204

4 EXPERIMENTAL SETUP

- 205 206 207
- 4.1 Data

208 Addressing the proposed task is made possible by OneStop Eye Movements (Malmaud et al., 2020), 209 the first dataset that contains broad coverage eye-tracking data in both ordinary reading and infor-210 mation seeking regimes. The textual materials of OneStop are taken from OneStopQA (Berzak 211 et al., 2020), a multiple-choice reading comprehension dataset that comprises 30 Guardian articles 212 from the OneStopEnglish corpus (Vajjala & Lučić, 2018). Each article is available in the original 213 (Advanced) and simplified (Elementary) versions. Each paragraph has three multiple choice reading comprehension questions that can be answered based on any of the two paragraph difficulty 214 level versions. Each question is paired with a manually annotated textual span, called the *critical* 215 span, which contains the vital information for answering the question. The mean paragraph length

is 109 words (min: 50; max: 165; std: 28). An example of a OneStopQA paragraph along with one question and its critical span annotation is provided in Table 2 in Appendix B.

Eye movements data for OneStopQA were collected in-lab from 360 adult native English speakers 219 using an EyeLink 1000+ eye tracker (SR Research), a state-of-the-art eyetracking device for reading 220 research. Each participant read a batch of 10 articles (54 paragraphs) paragraph by paragraph and answered a comprehension question about each paragraph. Each paragraph was shown to a participant 222 either in the Advanced or Elementary version, along with one of the three possible questions. The experiment has two between-subjects reading goal tasks: information seeking and ordinary reading. 224 In the information seeking task, participants were presented with the question (without the answers) 225 prior to reading the paragraph. In the ordinary reading task, they did not receive the question prior 226 to reading the paragraph. In both tasks, after having read the paragraph, participants proceeded to answer the question on a new screen, without the ability to return to the paragraph. 227

In sum, each participant read 54 paragraphs, all of which are either in information seeking or in ordinary reading. Each paragraph was read by 120 participants: 60 in ordinary reading and 60 in information seeking (split equally between the Advanced and Elementary versions of the paragraph).
Overall, the data consists of 19,438 trials, where a trial is a recording of eye movements from a single participant over a single paragraph. The data is balanced, with 9,718 trials in ordinary reading and 9,720 in information seeking. Figure 7 in Appendix B shows example trials for each of the two reading regimes. Additional data statistics are described in Appendix B.

236 4.2 MODEL TRAINING AND EVALUATION PROTOCOL

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

- New Item (paragraph): eye-tracking data is available during training for the participant but not for the paragraph.
- New Participant: eye-tracking data is available during training for the paragraph, but not for the participant.
- New Item & Participant: Neither the participant nor the paragraph are in the training data.

We further report aggregated results across all three regimes.



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

268

235

237

241

242

243

244

245

246 247

248 249

250 251

253

254

256

257

258 259

260

261

262

The data is split separately for each batch of 10 articles and the 120 participants who read the batch. The three batch splits are then combined to form the full split of the dataset. Paragraphs are

allocated to the train, validation and test portions of each batch split at the *article level*, such that all
the paragraphs of each article appear in the same portion of the split. This ensures that items in the
test set are unrelated in content to items in training and validation. Each split further guarantees an
equal number of participants from each OneStopQA batch in each data portion.

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

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

We perform hyperparameter optimization and model selection separately for each split. We assume that at test time, the evaluation regime of the trial is *unknown*. Model selection is therefore based on the entire validation set of the split. All neural network-based models were trained using the PyTorch Lighting (Falcon & team, 2024) library on NVIDIA A100-40GB and L40S-48GB GPUs. Further details regarding the training procedure, including the full hyperparameter search space for all the models are provided in Appendix C.

291 292 293

294 295

296

297

4.3 STATISTICAL TESTS FOR MODEL PERFORMANCE COMPARISONS

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

5 RESULTS

302 303

304 Test set accuracy results are presented in Table 1. In line with prior findings regarding faster reading in information seeking compared to ordinary reading (Malmaud et al., 2020; Hahn & Keller, 2023; 305 Shubi & Berzak, 2023), the Reading Time baseline yields above chance accuracies (p < 0.01 in all 306 evaluation regimes), thus providing a strong benchmark for the eyetracking-based models. Among 307 the 10 examined models, RoBERTa-Eye-F achieves the highest accuracy in the New Item, New 308 Item & Participant and All evaluations, while RoBERTa-Eye-W slightly outperforms it in the New 309 Participant regime. Differently from the New Item and the New Participant regimes, where multiple 310 models outperform the Reading Time baseline, in the most challenging New Item & Participant 311 regime, the improvements over this baseline are not statistically significant.

312 In addition to RoBERTa-Eye-F, three other fixation based models, PostFusion-Eye and the two 313 BEyeLSTM models, excel in the New Item regime. One possible explanation for this outcome 314 is better generalization of fixation based models to new items compared to new participants. Note 315 however, that due to the between-subjects design where all the training and test examples for a given 316 participant have the same label, it could alternatively reflect, at least in part, an ability of fixation 317 based models to learn participant specific reading behavior without explicit information on the par-318 ticipant (i.e. identify the participant), an ability that is not directly pertinent to the task at hand. The 319 current experimental setup makes it challenging to adjudicate between these two possibilities. In 320 either case, it is highly non-trivial that models are able to generalize from prior participant data to 321 new items. In Appendix D, we report the Receiver Operating Characteristic (ROC) curves across the ten cross-validation splits and their corresponding Area Under the ROC Curve (AUROC) scores 322 (mean and standard deviation) (Bradley, 1997) in Figure 8. Validation set accuracies are reported in 323 Table 3. The outcomes of these evaluations are consistent with the test results reported in Table 1.

Table 1: Test accuracy results aggregated across 10 cross-validation splits. 'E' stands for word embeddings, 'LF' for linguistic word properties such as word length, frequency and surprisal, and 'Fix' for fixations. Model performance is compared to the Reading Time baseline using a linear mixed effects model. In R notation: $is_correct \sim model + (model | participant) + (model |$ paragraph). Significant gains over this baseline are marked with '*' p < 0.05, '**' p < 0.01 and '***' p < 0.001 in superscript, and significant drops compared to the best model in each regime are marked in subscript with '+'.

Model	Gaze Rep.	Text Rep.	New Item	New Participant	New Item & Participant	All
Majority Class / Chance	-	-	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0
Reading Time	-	-	$59.0 \pm 0.4_{+++}$	$58.9 \pm 1.0_{++}$	60.4 ± 1.2	$59.0 \pm 0.5_{+++}$
Log. Regression (Mézière et al., 2023)	Global	_	$62.4 \pm 0.3^{**}_{+++}$	$60.6 \pm 1.4_{++}$	60.8 ± 1.6	$61.5 \pm 0.8^{*}_{+++}$
BEyeLSTM - No Text	Fix.	-	$71.5 \pm 0.6^{***}_{+++}$	61.0 ± 1.1	61.5 ± 1.5	$65.9 \pm 0.4^{***}_{++}$
RoBERTa-Eve-W (Shubi et al., 2024)	Words	E+LF	$65.9 \pm 0.6^{***}_{+++}$	$63.2 \pm 1.1^{**}$	62.3 ± 1.3	$64.4 \pm 0.6^{***}_{++}$
MAG-Eye (Shubi et al., 2024)	Words	E+LF	$52.1\pm0.3_{\texttt{+++}}$	$52.3\pm0.4_{\text{+++}}$	$51.5 \pm 0.4_{+++}$	$52.1 \pm 0.2_{++}$
PLM-AS (Yang & Hollenstein, 2023)	Fix. Order	Е	$58.6 \pm 0.4_{+++}$	$59.5 \pm 0.5_{+++}$	$57.5 \pm 0.9_{++}$	$59.0 \pm 0.4_{++}$
Haller RNN (Haller et al., 2022)	Fix.	Е	$61.3 \pm 0.8_{+++}$	61.5 ± 0.8	60.3 ± 1.6	$61.3 \pm 0.4_{++-}$
BEyeLSTM (Reich et al., 2022a)	Fix.	LF	$71.4 \pm 0.9^{***}_{+++}$	61.6 ± 1.1	62.2 ± 1.3	$66.2 \pm 0.7^{***}_{+}$
Eyettention (Deng et al., 2023)	Fix.	E+LF	$55.8 \pm 0.8_{+++}$	$55.7 \pm 1.1_{+++}$	$55.4 \pm 1.8_{+++}$	$55.8 \pm 0.9_{+++}$
RoBERTa-Eye-F (Shubi et al., 2024)	Fix.	E+LF	${\bf 74.7 \pm 1.2}^{***}$	$63.1 \pm 1.3^{*}$	62.6 ± 1.6	$68.5 \pm 0.9^{**}$
PostFusion-Eye (Shubi et al., 2024)	Fix.	E+LF	$70.7 \pm 1.1^{***}_{+++}$	$62.2 \pm 1.1^*$	61.4 ± 1.0	$66.1 \pm 0.8^{***}_{++}$

343 344

345

5.1 LOGISTIC ENSEMBLE

346 Ensemble models combine predictions from multiple individual models, and often yield better re-347 sults than single models, especially when models have diverse predictive behaviors (Sagi & Rokach, 348 2018). To examine the appropriateness of this approach for our case, we compute the agreement 349 between models' predictions, quantified with Cohen's Kappa (Cohen, 1960). Figure 3a depicts the 350 pairwise agreement rates across models in the validation data, where we observe mostly moderate 351 agreement rates. Figure 9 in Appendix E further suggests consistency across the three evaluation 352 regimes. These outcomes indicate that while the examined models do not always differ substantially 353 in accuracy, they likely capture different aspects of the eye movement data and the task.

354 Motivated by the results of this analysis, we introduce a simple **Logistic Ensemble**: a 10-feature lo-355 gistic regression model that predicts the reading goal from the probability outputs of our 10 models. 356 The model is trained for each fold on the validation set and tested on the corresponding three test 357 sets. In Figure 3b we present the performance of the Logistic Ensemble, against the best perform-358 ing model, RoBERTa-Eye-F, and the Reading Time baseline. We find that the Logistic Ensemble 359 improves over the accuracy of RoBERTa-Eye-F in all the regimes, with an accuracy of 77.3 for 360 New Items, 64.6 for New Participants, 64.3 for New Items & Participants, and 70.5 overall. The 361 improvements over RoBERTa-Eye-F are statistically significant in all but the New Item & Partici-362 pant regime, in which it is the only model that is statistically better than the Reading Time baseline. These performance improvements suggest that the information encoded by the different models is 363 to some extent complementary. 364

365 366

367

6 WHAT MAKES THE TASK EASY OR HARD?

Having established that the prediction task at hand can be performed with a considerable degree of
success, we now leverage the best performing model to obtain insights about the task itself. We
do so by examining which trial features, which were not given to the model explicitly, affect the
ability of the model to classify trials correctly. This analysis takes advantage of the rich structure
and auxiliary annotations of the OneStop dataset.

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

- 376 377
- 1-3. **Reading time before, within, and after critical span**: These features are motivated by the findings of Malmaud et al. (2020) and Shubi & Berzak (2023) regarding faster reading

403

404

405

406

407

408

409

410

411 412

413

414

415 416

417

418 419

420

421

422

423

424

425

430



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

times in information seeking compared to ordinary reading, primarily before and after the critical span, as well as the reported classification results of the Reading Time baseline.

- 4. Paragraph position (1-54): Each participant reads 54 paragraphs, in a random article order. This feature captures the position of the paragraph in the experiment's presentation sequence. It is included as reading strategies can change as the experiment progresses (e.g. Meiri & Berzak (2024) show that readers become faster as the experiment progresses).
- 5. Answered correctly: this feature encodes whether after having read the passage, the participant answered the given reading comprehension question correctly. It captures participant-specific task difficulty and the extent to which the participant read the passage attentively.
- We further include the following item (paragraph and question), reader-independent features:
 - 6. **Paragraph length** (in words): this feature is chosen as we hypothesize that more data could lead to more accurate predictions for the item.
 - 7. **Paragraph level** (Advanced / Elementary): is chosen as eye movements could be influenced by the difficulty level of the text, for example through differences in word frequency and surprisal (Singh et al., 2016; Hollenstein et al., 2022).
 - 8. **Critical span start location** (relative position, normalized by paragraph length): Shubi & Berzak (2023) showed skimming-like reading patterns after processing task critical information in information seeking. We thus hypothesize that earlier appearance of task critical information could facilitate the ability to correctly identify information seeking reading.
- Critical span length (normalized by paragraph length): we also hypothesize that less task critical information during information seeking could further aid distinguishing it from ordinary reading.
- 426
 427
 428
 428
 429
 10. Question difficulty (percentage of participants who answered the question incorrectly): based on data from the Prolific QA experiment in Berzak et al. (2020). We include this feature as it can influence eye movements in information seeking, with harder questions potentially obscuring patterns of goal oriented reading.
- In order to establish the relation of these features to task difficulty, we use a linear mixed effect model that uses these features to predict whether the trial prediction of the best performing model,



Figure 4: Coefficients from a mixed-effects model that predicts whether RoBERTa-Eye-F's prediction for a given trial is correct from properties of the trial. CS stands for the critical span, the portion of the paragraph that contains the information that is essential for answering the question correctly. Two models are fitted separately for ordinary reading and information seeking trials. Predictors are z-normalized. Depicted are the coefficients of the fitted models after a 10x Bonferroni correction, to mitigate the risk of false positives when testing multiple hypotheses simultaneously. '*' p < 0.05, '**' p < 0.01, '**' p < 0.001.

457

458

448

449

450

451

452

RoBERTa-Eye-F, was correct. In R notation:

is_correct ~ 1 + feat₁ + feat₂ + \cdots + feat₁₀ + (1 | participant) + (1 | item) + (1 | evaluation regime)

where the random effects account for correlations in predictions within participants, items and eval-459 uation regimes². The advantage of a multiple regression model over separate univariate analyses 460 of each feature, is the ability to tease apart the contribution of each feature above and beyond all 461 other features. We fit this model separately on the information seeking and ordinary reading trials. 462 To make the contributions of the features to prediction accuracy comparable, we normalize each 463 feature to be a z-score (zero mean and unit variance). We then examine feature contribution via 464 statistical significance, magnitude and sign of the corresponding coefficient. A significant coeffi-465 cient for a feature indicates that it correlates with task difficulty, the absolute value determines its 466 importance rank relative to other features, and the sign indicates the direction of the association. 467

The resulting feature coefficients are presented in Figure 4. In line with the findings of Shubi & 468 Berzak (2023), we observe that the most prominent features for correctly classifying both informa-469 tion seeking and ordinary reading trials are reading time before and after the critical span. Slower 470 readers in these regions are easier to correctly classify as ordinary reading and harder to correctly 471 classify as information seeking. We also find, as hypothesized, that shorter critical spans facilitate 472 correct classification of information seeking trials, presumably by making information seeking more 473 targeted. Finally, paragraph position is significant in information seeking, suggesting that readers 474 develop more efficient goal oriented reading strategies as they progress through the experiment. 475 Overall, this analysis provides a highly interpretable characterization of task difficulty, with insights on reading behavior in both reading regimes. 476

477 478

479

485

7 RELATED WORK

The vast majority of the literature on the psychology of reading and its interfaces with computational modeling is concerned with ordinary reading. However, several studies did address goal-oriented (also referred to as task-based) reading. Most prior work focused on a small number of canonical tasks: skimming, speed reading and proofreading. Several studies found different eye movement patterns in these tasks as compared to ordinary reading (Just et al., 1982; Kaakinen & Hyönä, 2010;

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

486 Schotter et al., 2014; Strukelj & Niehorster, 2018; Chen et al., 2023). Rayner & Raney (1996) ex-487 amined differences between ordinary reading and searching through the text for a target word. Prior 488 work also analyzed eye movements during human linguistic annotation, often used for generating 489 training data for NLP tools, such as annotation of named entities (Tomanek et al., 2010; Tokunaga 490 et al., 2017). Differences in reading patterns were also found when readers were asked to take different perspectives on a given text (Kaakinen et al., 2002) or given different sets of learning goals 491 (Rothkopf & Billington, 1979). We further note that information seeking has also been previously 492 examined in visual search (Jang et al., 2014a;b; Sharma et al., 2023). 493

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

501 While the above studies focus primarily on descriptive data analysis, Hollenstein et al. (2023) took 502 a predictive approach and attempted to automatically classify the reading task from eye movement 503 features. In this study, participants read single sentences from the ZuCo corpus (Hollenstein et al., 2020), and engaged either in ordinary reading or in an annotation task where they had to determine 504 the presence of one of seven semantic relations in the sentence. While this benchmark is conceptu-505 ally related to the current work, it is restricted in scope and ecological validity. First, the benchmark 506 contains only 18 participants, which makes generalization and reliable evaluations highly challeng-507 ing. More fundamentally, the framework is limited by the nature of the tasks, which focus on highly 508 specialized linguistic annotations that are not performed by readers in everyday life. If viewed 509 through the lens of question answering, the setup comprises only seven possible questions that con-510 cern the presence of a semantic relation in the sentence. In the current study we take a different 511 and more general stance on task based reading, with unrestricted questions that are more represen-512 tative of the tasks commonly pursued by readers. Finally, additional related lines of work used eye 513 movements in reading for prediction of other types of reader cognitive state, such as reading comprehension (Reich et al., 2022b; Shubi et al., 2024), as well as inferring properties of the text such 514 document type (Kunze et al., 2013) and readability level (González-Garduño & Søgaard, 2017). 515

516 517

518

8 SUMMARY AND DISCUSSION

519 Is it possible to decode reader goals from eye movements? We address this question by examining 520 the possibility to automatically discriminate between ordinary reading and information seeking at 521 the challenging granularity level of a single paragraph. We find that it is indeed possible to perform 522 this task with considerable success, and show that eye movements provide valuable information for 523 this task, above and beyond reading time. We further find that different models capture different 524 aspects of the signal and the task, and can be successfully combined with ensembling, which yields 525 the strongest results for our task. Our error analysis leverages the models to further reveal new insights on the factors that determine task difficulty. Our evaluation setup which addresses different 526 aspects of model generalization provides an infrastructure for future work on the proposed task. 527

Our study has a number of important limitations. The information seeking tasks are over individual paragraphs that span 3-10 lines of text. This leaves out shorter texts (e.g. single sentences) as well as longer texts. It is also restricted to newswire texts, and does not include texts from other genres.
Finally, new datasets for the information seeking task, other types of tasks, additional populations (e.g. second language readers, younger and older participants), and datasets in languages other than English are all needed in order to study goal decoding more broadly.

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

540 9 **ETHICS STATEMENT**

541 542 543

544

545

546

548

549

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

It has previously been shown that eye movements can be used for user identification (e.g. Bednarik et al., 2005; Jäger et al., 2020). We do not perform user identification in this study, and emphasize 547 the importance of not storing information that could enable participant identification in future applications of goal decoding. We further stress that future systems that automatically infer reader goals are to be used only with explicit consent from potential users to have their eye movements collected and analyzed for this purpose.

554

555

556

558

559

560

561 562

563

571

REPRODUCIBILITY STATEMENT 10

We describe the model training and selection procedure, the evaluation protocol, and the hardware and software specifications in Section 4.2. The hyperparameter search space for each model is described in Appendix C. The dataset used in the experiments is described in Section 4.1 and appendix B. The source code, which implements the models, experimental procedures, and analysis, is available at: https://anonymous.4open.science/r/ Decoding-Reading-Goals-from-Eye-Movements/. Data will be made publicly available

REFERENCES

- 564 Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for 565 confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3):255-566 278, April 2013. ISSN 0749-596X. doi: 10.1016/j.jml.2012.11.001. URL https://www. 567 sciencedirect.com/science/article/pii/S0749596X12001180. 568
- 569 Douglas Bates, Phillip Alday, Dave Kleinschmidt, José Bayoán Santiago Calderón, Likan Zhan, Andreas Noack, Milan Bouchet-Valat, Alex Arslan, Tony Kelman, Antoine Baldassari, 570 Benedikt Ehinger, Daniel Karrasch, Elliot Saba, Jacob Quinn, Michael Hatherly, Morten Piibeleht, Patrick Kofod Mogensen, Simon Babayan, Tim Holy, Yakir Luc Gagnon, and Yoni 572 Nazarathy. JuliaStats/MixedModels.jl: v4.26.0, September 2024. URL https://zenodo. 573 org/records/13754074. 574
- 575 Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and Pasi Fränti. Eye-movements as a biometric. 576 In Image Analysis: 14th Scandinavian Conference, SCIA 2005, Joensuu, Finland, June 19-22, 577 2005. Proceedings 14, pp. 780-789. Springer, 2005. 578
- 579 Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. STARC: Structured Annotations for Reading 580 Comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics., 2020. URL https://github.com/. 581
 - Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. CELER: A 365-participant corpus of eye movements in 11 and 12 english reading. *Open Mind*, pp. 1–10, apr 2022.
- 586 Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7):1145–1159, 1997. 588
- 589 Xiuge Chen, Namrata Srivastava, Rajiv Jain, Jennifer Healey, and Tilman Dingler. Characteristics 590 of deep and skim reading on smartphones vs. desktop: A comparative study. In *Proceedings of* 591 the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–14, 2023.
- 592

582

583

584

585

587

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

594 595 596 597 598	Shuwen Deng, David R. Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A. Jäger. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. In <i>Proceedings of the ACM on Human-Computer Interaction</i> , pp. 1–24. Association for Computing Machinery, may 2023. doi: 10.1145/3591131. URL https://doi-org.ezproxy.uzh.ch/10.1145/3591131.
599 600 601	William Falcon and The PyTorch Lightning team. Pytorch lightning, August 2024. URL https://doi.org/10.5281/zenodo.13254264.
602 603 604 605 606	Ana Valeria González-Garduño and Anders Søgaard. Using gaze to predict text readability. In Joel Tetreault, Jill Burstein, Claudia Leacock, and Helen Yannakoudakis (eds.), <i>Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications</i> , pp. 438–443, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5050. URL https://aclanthology.org/W17-5050.
607 608	Michael Hahn and Frank Keller. Modeling task effects in human reading with neural network-based attention. <i>Cognition</i> , 230:105289, 2023.
609 610 611	John Hale. A probabilistic earley parser as a psycholinguistic model. In Second meeting of the north american chapter of the association for computational linguistics, 2001.
612 613 614	Patrick Haller, Andreas Säuberli, Sarah Elisabeth Kiener, Jinger Pan, Ming Yan, and Lena Jäger. Eye-tracking based classification of mandarin chinese readers with and without dyslexia using neural sequence models. <i>arXiv preprint arXiv:2210.09819</i> , 2022.
615 616 617 618 619	Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. <i>Neural Computation</i> , 9(8): 1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://ieeexplore.ieee.org/abstract/document/6795963. Conference Name: Neural Computation.
620 621 622 623 624 625 626	Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. ZuCo 2.0: A dataset of physio- logical recordings during natural reading and annotation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isa- hara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Ste- lios Piperidis (eds.), <i>Proceedings of the Twelfth Language Resources and Evaluation Conference</i> , pp. 138–146, Marseille, France, May 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.18.
627 628 629 630 631	Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. Patterns of text readability in human and predicted eye movements. In Michael Zock, Emmanuele Chersoni, Yu-Yin Hsu, and Enrico Santus (eds.), <i>Proceedings of the Workshop on Cognitive Aspects of the Lexicon</i> , pp. 1–15, Taipei, Taiwan, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.cogalex-1.1.
632 633 634	Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A Jäger, and Nicolas Langer. The zuco benchmark on cross-subject reading task classification with eeg and eye-tracking data. <i>Frontiers in Psychology</i> , 13:1028824, 2023.
635 636 637 638 639	Lena A Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler, and Tobias Scheffer. Deep eyedentification: Biometric identification using micro-movements of the eye. In <i>Machine</i> <i>Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019,</i> <i>Würzburg, Germany, September 16–20, 2019, Proceedings, Part II</i> , pp. 299–314. Springer, 2020.
640 641 642	Young-Min Jang, Rammohan Mallipeddi, and Minho Lee. Identification of human implicit visual search intention based on eye movement and pupillary analysis. <i>User Modeling and User-Adapted Interaction</i> , 24:315–344, 2014a.
643 644 645	Young-Min Jang, Rammohan Mallipeddi, Sangil Lee, Ho-Wan Kwak, and Minho Lee. Human in- tention recognition based on eyeball movement pattern and pupil size variation. <i>Neurocomputing</i> , 128:421–432, 2014b.
646 647	Marcel Adam Just, Patricia A Carpenter, and MEJ Masson. What eye fixations tell us about speed reading and skimming. <i>Eye-lab Technical Report Carnegie-Mellon University, Pittsburgh</i> , 1982.

- 648 Johanna Kaakinen and Jukka Hyönä. Task Effects on Eye Movements During Reading. Journal 649 of experimental psychology. Learning, memory, and cognition, 36:1561–6, November 2010. doi: 650 10.1037/a0020693. 651 Johanna K Kaakinen, Jukka Hyönä, and Janice M Keenan. Perspective effects on online text pro-652 cessing. Discourse processes, 33(2):159-173, 2002. 653 654 Alan Kennedy, Robin Hill, and Joël Pynte. The dundee corpus. In European conference on eye 655 movement, 2003. 656 Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. Length, frequency, and predictabil-657 ity effects of words on eye movements in reading. European journal of cognitive psychology, 16 658 (1-2):262-284, 2004. 659 660 Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. I know what you are reading: recognition of document types using mobile eye tracking. In Proceedings of the 2013 661 international symposium on wearable computers, pp. 113–116, 2013. 662 663 Roger Levy. Expectation-based syntactic comprehension. Cognition, 106(3):1126–1177, 2008. 664 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike 665 Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pre-666 training Approach. arXiv:1907.11692 [cs], July 2019. URL http://arxiv.org/abs/ 667 1907.11692. arXiv: 1907.11692. 668 669 Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In International Con-670 ference on Learning Representations, September 2018. URL https://openreview.net/ 671 forum?id=Bkg6RiCqY7. 672 Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. Bridging Information-Seeking Human Gaze 673 and Machine Reading Comprehension. In Proceedings of the 24th Conference on Computational 674 Natural Language Learning, pp. 142–152, Stroudsburg, PA, USA, 2020. Association for Compu-675 tational Linguistics. doi: 10.18653/v1/2020.conll-1.11. URL https://www.aclweb.org/ 676 anthology/2020.conll-1.11. 677 Yoav Meiri and Yevgeni Berzak. Déjà vu: Eye movements in repeated reading. In Proceedings of 678 the Annual Meeting of the Cognitive Science Society, volume 46, 2024. 679 680 Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-681 tuning {bert}: Misconceptions, explanations, and strong baselines. In International Confer-682 ence on Learning Representations, 2021. URL https://openreview.net/forum?id= 683 nzpLWnVAyah. 684 Diane C. Mézière, Lili Yu, Erik D. Reichle, Titus von der Malsburg, and Genevieve McArthur. 685 Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 58 686 (3):425-449, 2023. doi: 10.1002/rrq.498. URL https://ila.onlinelibrary.wiley. 687 com/doi/abs/10.1002/rrq.498. 688 Ralph Radach and Alan Kennedy. Theoretical perspectives on eye movements in reading: Past 689 controversies, current issues, and an agenda for future research. European journal of cognitive 690 psychology, 16(1-2):3-26, 2004. 691 692 Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-693 Philippe Morency, and Ehsan Hoque. Integrating Multimodal Information in Large Pretrained 694 Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computa-695 tional Linguistics, pp. 2359–2369, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.214. URL https://aclanthology.org/2020. 696 acl-main.214. 697 Keith Rayner. Eye movements in reading and information processing: 20 years of research. Psy-699 chological bulletin, 124(3):372, 1998. 700
- 701 Keith Rayner and Gary E Raney. Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review*, 3(2):245–248, 1996.

702 703 704	Keith Rayner, Jane Ashby, Alexander Pollatsek, and Erik D Reichle. The effects of frequency and predictability on eye fixations in reading: Implications for the ez reader model. <i>Journal of</i> <i>Exparimental Psychology: Human Parameter and Parformance</i> 30(4):720–2004
705	Experimental 1 Sychology. Human 1 erception and 1 erjormance, 50(4).720, 2004.
706	Keith Rayner, Timothy J Slattery, Denis Drieghe, and Simon P Liversedge. Eye movements and
707	word skipping during reading: Effects of word length and predictability. Journal of Experimental
708	Psychology: Human Perception and Performance, 37(2):514, 2011.
709	David Dahart Daiah Daul Dragga Chiara Tashirpar Datriak Hallar Frank Caldhammar and Lana A
710	Läger Inferring native and non-native human reading comprehension and subjective text difficulty
711	from scannaths in reading. In 2022 Symposium on Eve Tracking Research and Applications, pp.
712	1–8, 2022a.
713	
714	David Robert Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A.
715	Difficulty from Scappaths in Reading. In 2022 Symposium on Eva Tracking Research and Appli
716	cations pp 1–8 Seattle WA USA June 2022 h ACM JSBN 978-1-4503-9252-5 doi: 10.1145/
717	3517031.3529639. URL https://dl.acm.org/doi/10.1145/3517031.3529639.
718	
719	Ernst Z Rothkopf and MJ Billington. Goal-guided learning from text: inferring a descriptive pro-
720	cessing model from inspection times and eye movements. <i>Journal of educational psychology</i> , 71 (2):210–1070
721	(3):310, 1979.
722	Omer Sagi and Lior Rokach. Ensemble learning: A survey. Wiley interdisciplinary reviews: data
723	mining and knowledge discovery, 8(4):e1249, 2018.
724	Elizabeth D. Sebetter, Klinten Dieknell, Jen Heurend, Degen Leur, and Keith Derman, Teels offeete
720	reveal cognitive flexibility responding to frequency and predictability: Evidence from eve move-
720	ments in reading and proofreading. <i>Cognition</i> , 131(1):1–27, 2014.
728	
729	Mansi Sharma, Shuang Chen, Philipp Müller, Maurice Rekrut, and Antonio Krüger. Implicit search
730	ceedings of the 25th International Conference on Multimodal Interaction, pp. 345–354, 2023.
732	Omer Shubi and Yevgeni Berzak. Eye movements in information-seeking reading. In <i>Proceedings</i>
733	of the Annual Meeting of the Cognitive Science Society, 2025.
734	Omer Shubi, Yoav Meiri, Cfir A. Hadar, and Yevgeni Berzak. Fine-grained prediction of reading
736	comprehension from eye movements. In <i>The 2024 Conference on Empirical Methods in Natural Language Processing</i> , 2024.
737	
738	Noam Siegelman, Sascha Schroeder, Cengiz Acarturk, Hee-Don Ahn, Svetlana Alexeeva, Simona
739	panding horizons of cross-linguistic research on reading. The multilingual eve-movement corrus
740	(meco). Behavior research methods, 54(6):2843–2863. 2022.
741	
743	Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. Quantifying sen-
744	tence complexity based on eye-tracking measures. In Dominique Brunato, Felice Dell'Orletta,
745	Computational Linguistics for Linguistic Complexity (CLALC) pp. 202–212 Osaka Japan De
746	cember 2016. The COLING 2016 Organizing Committee, URL https://aclanthology.
747	org/W16-4123.
748	
749	Alexander Strukelj and Diederick C Niehorster. One page of text: Eye movements during regular
750	and thorough reading, skimming, and spell checking. <i>Journal of Eye Movement Research</i> , 11(1), 2018
751	2010.
752	Takenobu Tokunaga, Hitoshi Nishikawa, and Tomoya Iwakura. An eye-tracking study of named
753	entity annotation. In Proceedings of the International Conference Recent Advances in Natu-
754	ral Language Processing, RANLP 2017, pp. 758–764, Varna, Bulgaria, September 2017. IN-
755	COMA Ltd. doi: 10.26615/978-954-452-049-6_097. URL https://doi.org/10.26615/ 978-954-452-049-6_097.

756 757 758 759	Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jürgen Ziegler. A cognitive cost model of annotations based on eye-tracking data. In <i>Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics</i> , pp. 1158–1167, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1118.
760 761 762 763 764 765	Sowmya Vajjala and Ivana Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In <i>Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications</i> , pp. 297–304, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0535. URL https://aclanthology.org/W18-0535.
766 767 768	Duo Yang and Nora Hollenstein. Plm-as: Pre-trained language models augmented with scanpaths for sentiment classification. <i>Proceedings of the Northern Lights Deep Learning Workshop</i> , 2023. URL https://api.semanticscholar.org/CorpusID:256262367.
769	
770	
771	
772	
773	
774	
775	
776	
777	
778	
779	
780	
782	
783	
784	
785	
786	
787	
788	
789	
790	
791	
792	
793	
794	
795	
796	
797	
798	
799	
801	
802	
803	
804	
805	
806	
807	
808	
809	



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



910 Figure 6: Visualization of the different model architectures (Part 2). P represents the paragraph, 911 E_S^P the eye movements of participant S on P. LM stands for a language model, and FC for fully 912 connected layers. FF_{f_i} stands for the fixation features and w_{f_i} for the word corresponding to the 913 *i*-th fixation respectively.

- 914
- 915
- 916 917

918 B ONESTOP EYE MOVEMENTS DATASET - ADDITIONAL DETAILS 919

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

Table 2: An example of a OneStopQA paragraph (Advanced and Elementary version) along with one of its three questions. The critical span is marked in bold red. Adapted from Berzak et al. (2020).

931 932	Advanced A major international disagreement with wide-ranging implications for globa				
933		principal ingredient in cocaine. Boliv	via has obtained a special exemption from the 1961		
934		Single Convention on Narcotic Dru	gs, the framework that governs international drugs		
935		policy, allowing its indigenous peo	ple to chew the leaves. Bolivia had argued that the		
936		"convention was in opposition to its n "protect native and ancestral coca as	ew constitution, adopted in 2009, which obliges it to s cultural patrimony" and maintains that coca "in its		
937		natural state is not a dangerous nar	cotic."		
938	Elementary	A big international disagreement has	started over the right of Bolivia's indigenous Indian		
939		tribes to chew coca leaves, the main	ingredient in cocaine. This could have a significant		
940		effect on global drugs policy. Bolivi	ia has received a special exemption from the 1961		
941		Convention on Drugs, the agreem	ent that controls international drugs policy. The		
942		convention was against its new const	titution, adopted in 2009, which says it must "protect		
943		native and ancestral coca" as part of	its cultural heritage and says that coca "in its natural		
944		state is not a dangerous drug."			
945	Question	What was the purpose of the 1961	Convention on Drugs?		
946	Answers	A Regulating international policy on	drugs		
947		C Discussing the legal status of Boliv	via's constitution		
948		D Negotiating extradition agreements	s for drug traffickers		
949			6		
0 = 0					
950		()rdinary Reading	Information Socking		
950 951	Ot mainternational disage	Ordinary Reading	Information Seeking		
950 951 952	A-major Gintergational disage	Ordinary Reading eenent <u>eistewide²⁴anging inpitcitions fo²² algoal arugs potty</u> has ² Huis's Indiamous Indian remos to cheatures lokues, the aritminal	Information Seeking		
950 951 952 953	A-major:Antergational-disagr erupted over the Pight of Bo ingr@dient.in co2aine_BofTo	Ordinary Reading eeent estemated and a second state of the second	Information Seeking		
950 951 952 953 954	CA-major Cintergational disage erupted over the Pight of Be ingredient of colaine. Boffy on Narcobic Drugs, the frame	Ordinary Reading essent sittewiseffanging implications for global drugs polity has Maria's indigenous Indian trabes to cheefcoca isous, the principal is has obtained a special exemption from the 1961 SingleConvertion more that governs international drugs.policy, alloging its	Information Seeking A server property of the server of th		
950 951 952 953 954 955	CA major Cintergational disage erupted over the Pight of Be ingredient of coloring the second on Natoric Drugs, the Frame indigenous people to chew th	Urdinary Reading essent sitewitefrenging implications for global drugs polify has bevia's indigénous Indian trobes to cheolocca léaves, the point/bal is has obtained a special exemption from the 1961 Single-Konvention with that governs international drugs,policy, alloging its e Taues oblidanded-araged that the convention was in apposition	Information Seeking Association of the second secon		
950 951 952 953 954 955 956	CA major:Ginternational-disage erupted over the Pight-of-Be ingredent.gn_codaine: Boff on NaTcoff: Drugs, the frame indigenous people to chew th to its new constitution, add	Urdinary Reading essent sitemated anging tepticitions for global drugs polify has bevia's indigenous indian reflect to checkocka i daves, the perint pal to has obtained a special exemption from the 1961 Single convertion with that governs international drugs, policy, alloging its a flavor, doited abased argued that the convegtion was in apposition pred fn 2009, which obliggs it to product native, and ancestral	Information Seeking Access provide the state of a service of the state of the service of the se		
950 951 952 953 954 955 956 957	C4 major@ntergational-disage erubted offer the Pight of Be ingredient. <u>In codate</u> .BOTT on Nafcoffc Drugs, the Fram indigenous Despite to chew th to its new constitution, add coca as cultural-patrimony	Ordinary Keading eeenet eithewide/hanging impirchtions for glabal drugs polity haf iskra's indigenous indfan trebes to cheekcoca Teares, iba-princepal in has obtained a spectal exemption from the 1991 Singlesconvertion more that the governs international drugs, policy, allowing its reflaves: solida, bad argued that the convection was is a spossition pred th 2009, which obliggs it to protect native, and ancestral and minimation-that cogn in itsconstant astes, is not, a	Information Seeking A serie provide an adjust with right canging implications to allow drops bolicy has erupted over the series an Ballyry's indigenous Indian tribes to chew gots leaves, the principal ingredient in contra Ballyry's indigenous Indian tribes to chew gots leaves, the principal ingredient in contra Ballyry's that begins indigenous Indian tribes to chew gots leaves, the principal ingredient in contra Ballyry's the fragment that over intercentional drugs malicy, allowing its indigenous propher to chewethe leaves. Ballyria has goged that the convention me to copposition to regenerize the fragment in 2009, within golinges it to "gotekt native with antestral core agenerized any and mathement with the 'ne review understate. (Figt a		
950 951 952 953 954 955 956 957 958	CA major(Intergational-disage erupted over the Ptylet of the ingreetent of collater soft on NaPcoPic Drugs, the Frame indigenous people to chew th to its new constitution, add coca as cultured-petringop?" dangegous maRcofic."	Ordinary Keading eeeent statewidefranging impircitions for global drugs polify haf idvia's indigenous indian trades to cheef coca idves, the principal is has obtained a spectal exemption from the 1961 Single Convertion work tidf governs internacional drugs_policy, alloging its a flaves: dolidathed argsget that the corrugtion was in agnospition bred fn 2009, which obliggs it to group and ancestral and winterfine that coga "in its onsural state	Information Seeking A serie professional disagreegent with right compiles inplications root plots drops boilty has experied over the experient statistics's indigenous Indian tribes to thew gota leaves, the principal ingredient in costing ability is indigenous Indian tribes to thew gota leaves, the principal ingredient in costing ability is indigenous Indian tribes to the gota leaves, the principal ingredient in costing ability is indigenous Indian tribes to the gota leaves, the principal ingredient in costing ability is indigenous international drops policy, allowing its indigenous people to cheege leaves, soluble has appeared that the company of a significant of the set of protect native full and esteral cost a second set of the		
950 951 952 953 954 955 956 957 958 959	A major Antergational disage erupted over the Hight of the ingredient of Colates Bolto on Narcofic Drugs, the frame indigénous people to the th to its new constitution, add coca as cultured-patrimony. dangenous narkotic.	Ordinary Reading essent sitewiseffanging imprications for global drugs polify has hisra's indigenous inflan trobes to cheefcoca lidees, the principal is has obtained a special exemption from the 1961 Singledoweirfon work that governs international drugs_policy, alloging its a flaves: holiday, had arged that the convection mas in appaasition pred fi 2009, which obliggs it to "product native, and ancestral and mirratins-that-ogg "in isconstural-state	Information Seeking A sector professional disagregment with right canging implications too placed drops for try has erupted over the want caling is indigenous Indian tiples to chew goca leaves, the principal ingredient in coratine Boligie flat thating a special exemption from the 1961 Single Convention on Naroffic (Drops, the fragment that devent interactional drops molicy, allowing its indigenous people to chemethe leaves. Boligie had argued that the convention are in opposition to its and construction, adopted in 2009, with allowing its of the the convention are in opposition to its and construction, adopted in 2009, with allows it to "gradiet native indianets and coca agenting and adopted in 2009, with allows in its Sectoral tractions in antestral coca agenting agents and administry marking its in its Sectoral tractions in its Sectoral traction. Its mark a dangerous narcotic."		
950 951 952 953 954 955 956 957 958 959 960	A major international disage erubted over the Pight of Be ingredient of Catanese Boff on Narcofic Drugs, the frame indigenous people to chew th to its new constitution, add coca as culturel patrimony dangegous narkotte. A major international disagen	Urdinary Reading essent sitewide/enging imprications for global drugs porfly haf idvia's indigenous indian trabes to chearcera leaves, the participal is has obtained a spectal ecception from the 1961 SingleGoweinflon work tild? governs international drugs_policy, allowing its e Teaves: Bolidaghed_argogithat the convegtion was in apposition pred fr 2009, which obliggs it to "product native and ancestral and maintains that coga in inconstinual state	Information Seeking A sector proprietional disagreegent with right canging implications for planar drops for try has erupted over the wright of Ballyin's indigenous Indian trybes to thew goed leaves, the grincipal ingredient in coating. Boligher flat they indigenous Indian trybes to thew goed leaves, the grincipal ingredient in coating. Boligher flat they wright a special exemption from the 1961 Single Convention on Narotic Drops, the fragement shat severet intercartonal drugs molicy, allowing its indigenous people to thewette leaves. Boligha that the good find antestral coat agentization, molopeds in 2000, which gotings in to "good it networks and antestral coate agentization and wright more than the try setural catate. If the a dangerbor marcotic."		
950 951 952 953 954 955 956 957 958 959 960 961	On major Cinternational disage erubled over the Ptylet of Be ingreeient on Calatest soft on Narcofic Drugs, the Fram indigenous people to chew th to its new constitution, and cocca as cultural partmaps dangegous markottc. Apage_international disagen erupted-over the right of Bo	Urdinary Reading essent sitewisefearing implications for global drugs polify has being a special exemption from the 1961 Single Convertibut which that governs international drugs polify, allowing its e Tearer, which abliggs its convertibut native, and ancestral and white that convertibut is to opposite native, and ancestral and white that convertibut is to oppose the solution when which obliggs its to oppose the solution of the solution when which obliggs its to oppose the solution rement with bide-spanging-implications for global drugs-policy has gover's indegenous indian trible to oppose the solutions	Information Seeking A sector proprietional disagreegent with ride complex implications too plana drops bottly has erupted over the engle of Bally(is's indigenous Indian tribes to the goed lawes, the principal ingredient in constant. Horizon that deverse international drugs policy, allowing its indigenous account of the engles. Bally(is's indigenous international drugs policy, allowing its indigenous account of the engles. Bally(is's indigenous international drugs policy, allowing its indigenous account of the engles. Bally(is's indigenous international drugs policy, allowing its indigenous account of the engles. Bally(is's indigenous international drugs policy, allowing its indigenous account of the engles in 2000, within golieges it to "gratelt native with antestral core against a margine and within me the the engles indigenous indigenous for the sector of the engle a danger bally and of the other of the engles ingligations for abball drugs nglicy, has endper to the train of Borbivia's indigenous Indian traines for abball drugs nglicy has endper over the train of Borbivia's indigenous Indian traines for allowing the principal endperiod over the train of Borbivia's indigenous Indian traines to the train traines. The principal endperiod over the train of Borbivia's indigenous Indian traines to the traines the principal endperiod over the traine of Borbivia's indigenous Indian traines to the traines the principal endperiod over the traines of Borbivia's indigenous Indian traines to the traines the principal endperiod over the traines of Borbivia's indigenous Indian traines to the traines of the principal endperiod over the principal in the Borbivia's indigenous Indian traines to the traines of the principal endperiod over the principal and the pri		
950 951 952 953 954 955 956 957 958 959 960 961 962	CA majorCintergational-disage erubited over the Ptylet of Be ingredient. 50, Colather-Boff on Narcofic Drugs, the Frame indighous DeoDe to chew th to its new constitution, add coca as cultural-partings/ dangegoue naRotic. ⁴ Apraige_interNational disagin erupted-over.the right.of Boi ingredient_in cocaster. Boffy	Urdinary Reading essent sidewideFenging imprications for global drugs polify has ldvia's indigenous Indfan trebes to checkocci leaves, the permitipal is has obtained a special exemption from the 1961 Single convention with the governs international drugs, policy, allowing its is faster, solida, bad argued that the convertion was in agnosistion pred fr 2009, which abliggs its opposed in argue and ancestral and winnesfins that coga is its convertion for global drugs-policy has updates individual and tribles to choose drugs drugs-policy has updates individual and tribles to choose drugs drugs-policy has update individual tribles to choose drugs drugs-policy has update individual and tribles to choose drugs drugs-policy has update individual and tribles to choose drugs drugs-policy has update individual tribles to choose drugs drugs-policy has update in the oppeding exemption from the 1961 Single Convention	Information Seeking A sector provide and disagreement with ride complex implications too plotted drops bolicy has erupted over the steps of bolicit's indigenous Indian types to the geca leaves, the principal ingredient in cocatme. Bolicit's indigenous Indian types to the geca leaves, the principal ingredient in cocatme. Bolicit's indigenous Indian types to the geca leaves, the principal indigenous for the fragment that deverse international drugs policy, allowing its indigenous feedbe to cheeve the leaves. Bolicit's indigenous international drugs collect active for antestral coca asserting and antiference and allowing that the convention for allowed antestral coca asserting and antiference and allowing indigenous indigenous indigenous indigenous indigenous a specific international disagreement of whose doming indigenous for allowing the hose explored the whom of Bolivia's-indigenous Indian toppes to coca leaves, the anticipal indigenous to colored a special exemption from the 1981 Shelle Convention indigenous in colored a special exemption from the 1981 Shelle Convention		
950 951 952 953 954 955 956 957 958 959 960 960 961 962 963	A majorCintergational-disage erupted over the Hight of Bo ingredient. Br. Colather: Boft on NaFCoFt: Drugs, the Frame indighous Deople to chew th to its new constitution, and coca as collural-partingay: dangegous insPort: A major_internactional disagfine erupted-over, the right-of-boi ingrediented a cogainer. Bhft on Narcogic Drugs, the Frame	Urdinary Reading essent sidewidefenging implications for global drugs polify has being's indigenous indian trades to checkocca leaves, the paintspal is has obtained a special exemption from the 1961 Single Convertion work that governs international drugs, policy, allowing its a favore, dollida, bad argued that the convertion was to apposition bred th 2000, which obliggs it to protect native and ancestral and entresting the coga "in its network to global drugs-policy has issues which rold-equaging implications for global drugs-policy has issues the traditions that coga is its other code to global drugs-policy has issues to the coga which the traditions for global drugs-policy has issues to the coga which the tradition from the 1962-Single Convention or that governo international drugs policy, at Dorige to the governo international drugs policy, at Dorige to	Information Seeking A server provide and disagreement with ride complex implications row plantal drops bilty has eruption over the end bally is indigenous Indian types to the goal laws, the principal ingreading in coastant ability is indigenous Indian types to the goal laws, the principal ingreading poper to the end bally is indigenous Indian types to the goal laws, the principal indigenous provide to the end bally in 2009, with different to ingread in an end of an end of an end to ingread an end of the end of an end of an end indigenous provide to the end of an end of the to ingread an end of the end		
950 951 952 953 954 955 956 957 958 959 960 961 962 963 964	A sujorCintergational-disage erupted over the Hight of Bo ingredient. BJ. Collates - Boft on NaTcolic Drugs, the Frame indigenous second to the the trans- indigenous resolution, and coca as culturel-petringor dangegous material. A maige-international disagin erupted-over the right of Bos ingrediented accelere boff ingrediented accelere boff	Urdinary Reading essent sidewidefenging implications for global drugs polify has being's indigenous infran trebes to checkock a lakes, the periods of the has obtained a special exemption from the 1961 Single Convertion whit that governs international drugs, policy, alloging its a Bavar, doited a special exemption from the 1961 Single Convertion when that doited a special and the convergion was to apposition pred in 2009, which obliggs its the convergion was to apposition pred in 2009, which obliggs its to converge the second and minimating that the converge of a second and accepted and minimating that the converge of the second and accepted and minimating that the converge of the second accepted and minimating that the converge of the second accepted and minimating that the converge of the second accepted and the second accepted of the second accepted of the second accepted a specific converge of the second accepted of the second accepted a specific converge of the second accepted of the of the second accepted of the second of the second accepted of the of the second accepted accepted of the second of the second of the of the second accepted accepted of the second of the second of the of the second of the second of the second of the second of the of the second of the second of the second of the second of the of the second of the second of the second of the second of the of the second of the second of the second of the second of the of the second of the of the second of the of the second of the of the second of the of the second of the of the second of the of the second of the second of the second of	Information Seeking A set performant of ingregerent with right complex implications not plant drops for the per- reputed over the weather saling is indigenous Indian tribes to the gran plant drops for the period ingredient in cratter. Bolight fit indigenous Indian tribes to the gran plant drops for the period ingredient in cratter. Bolight fit indigenous Indian tribes to the gran blant drops for the period indigenous people to chearge leaves, ball what drops it to "protect networks are fit indigenous people to chearge leaves, and and aggred that the convention are for popped to the second and the second and the second of the second of the second of the second and the second cost as second and the second of the second and international disgredeent are black down in the second of th		
950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965	A major international disage erubted over the Pipht of the ingredient of calane. Boff on Narcolic Drugs, the frame indigénous people to chee th to its new constitution, add coca as culturel perflexor dangeous narketic. Analge international disage erupted over the right of ess ingredient la cocaine. Bhit on Narcolic Drugs, the right on Narcolic Drugs, the right on Narcolic Drugs, the right	Urdinary Keading essent situation of a provide any provide and an experimental the has obtained a spectal exception from the 1961 Single Convertion with this governs international drugs, policy, allowing its a base obtained a spectal exception from the 1961 Single Convertion with this governs international drugs, policy, allowing its a base obtained a spectal exception from the 1961 Single Convertion with this governs international drugs, policy, allowing its a base of the solid a spectal exception for ground and an exceptal and maintains that eggs in its provide native, and ancestral and maintains that eggs in its provide data with a spectral state syster's independent a spectral exception for global drugs-policy has isolar's independent a spectral exception from the 1962 Single Convention of that govern cline criterion from the 1962 Single Convention of that govern cline maintain from the 1962 Single Convention of that govern cline maintain from the 1962 Single Convention of that govern cline maintain from the 1962 Single Convention of that govern cline maintain from the 1962 Single Convention of the govern cline maintain from the 1962 Single Convention of the govern cline maintain from the 1962 Single Convention of the govern cline maintain from the 1962 Single Convention of the govern cline maintain from the 1962 Single Convention of the govern cline maintain from the single site of the convention was the opposition of the govern cline maintain for the convention was the opposition of the convention from the convention was the opposition	Information Seeking And a professional disagreegent with ride canging implications too global drops for try for emptied over the weak at ablivis's indigenous Indian trybes to they goal leaves, the principal ingred for in creating to be ablivis's indigenous Indian trybes to they goal leaves, the principal ingred for in creating to the for the training a special exemption from the 1961 Single Convention on Naroffic Drugs, the fragement that devent intercational drugs roblicy, allowing its indigenous people to themethe leaves. Bolivia has argued that the convention and sets to its and the formation and antifuture true there in the seturator take and an dangerbors marcotic." A what international drugs people convention of bolive danging implications for abbala drugs nolice has explored over the trypt of bolivia's indigenous Indian trybuses they cocal leaves, the principal indigenous people to change that documents interactional drugs policy. The original indigenous people to the fragement that operate indigenous Indian trybuses they cocal leaves, the principal indigenous people to change that documents interactional drugs policy. The trypt is indigenous people to change the togother indigenous Indian trybuses they cocal leaves, the principal indigenous people to change that documents interactional drugs policy. The indigenous indigenous indigenous people to change the fragements that governs interactional drugs policy. The indigenous indigenous indigenous indigenous indigenous indigenous indigenes policy allowing its indigenous people to change the fragements that and a section of the the convention and indigenous to the section the fragements that documents interactional drugs policy and and the indigenous		
950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966	CA major lintergational disage erubted over the Pipht of the ingredient of catane = Boff on Narcofic Drugs, the frame indigenous people to chew th to its new constitution, and coca as culturel-partmony dangegous markette. Agazige_international disagen erupted-over, the right-of-eer ingredient da cocaine - Infor- on Narcofic Drugs, other frame indigenous people to chew th to (<u>iss_newcongstitution</u> , adv coca as collevelal rectored	Urdinary Reading essent sitewide/enging implications for global drugs polify haf idvia's indigenous indian trabes to chearcers likes, the participal is has obtained a special ecception from the 1961 SingleGoweinBon work that governs international drugs,policy, allowing its a baser wolkida,bad argogi that the convegtion was in opposition pred fr 2009, which obligge it to "problet narive and ancestral and minimum to that eggs in its opposition for global drugs-policy has two is the deligned of the two problet narive and ancestral and minimum to that eggs in its opposition for global drugs-policy has two is the deligned of the two is for global drugs-policy has two is the deligned of the two is for global drugs-policy has two is the indigenous indian tribes to chip, cold toles indige tokention ork that governesinternational drugs policy, allowing its 	Information Sceking And a professional disagreegent with ride canging implications too classed drops bottly has empted over the weak an ablyin's indigenous Indian types to the gene leaves, the principal ingred det in coatter. Bottly that the international drops policy, allowing its indigenous people to cheestle leaves. Solid ablying that the gongention are in appendix coat agents at ablying in 2000, which gotteges in to "profest metric and and estimation on Narotic Drops, the frageness that governt international drops policy, allowing its indigenous people to cheestle leaves. Solid able angles in its duration and estimations to right an attraction of the profession of the solid and angles in the to "profest metric able and angle for international disgreement with other angles implications for abbott and estimation information to right of converte leaves. Solid and angles indigenous for the international drops and and angles and a effort for the right of converte indigenous Indian tradement for the top abbott action of the arige of information to right of converte indigenous Indian tradementer that the top and the solid and angles and the solid and angles and the formation of the fragments that governs international drops policy. The fragments that governs international drops policy allowing its indigenous people to cheer the there are indigenous Indian tradementer 1901 Shiple convention on Narctic beauge, the fragments that governs international drops policy. The fragments that core in the indigenous people to cheer the there are indicated a solid angle drops of the indigenous and angles in the indigenous and angles and angles and angle and the indigenous and the solid and angle indigenous indigenous and the indigenous and the indigenous and the indigenous and angles ang		
950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967	A major international disage erubted over the Pipht of Be ingreeten an construction, and indigenous people to chew th to its new constitution, and coca as cultural partingor dangenus markett. Apaige_international disager erupted-over the right of eo ingreeteness the right of eo ingreeteness the right of eo ingreeteness the right of eo indigenous page to chew th to its new constitution, who	Urdinary Reading essent sitewide/enging implications for global drugs polify has hista's indigenous indfan trobes to chewhood leaves, the principal is has obtained a special exception from the 1961 Single Convention with this governs international drugs_policy, allowing its e Teaves, eolida, had argued that the convention was in apposition pred fri 2000, which obliggs it to "probet native and ancestral and winntsche that eggs "in its optimal state	Differentiation of the register of the right		
950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968	A major lintergational disage erubted over the Pipht of de ingredient of colater - soft on Narcolic Drugs, the fram indigénous secolic to chew th to its new constitution, and cocca as coltural partingsp dangegous marcolic. Apalge_international disager erupted-over the right of eor ingredient in costs to boty on Narcogic Drugs, other fram indigénous pargie to chew th to its new constitution, sho to its new constitution, sho to its new constitution, sho to its new constitution, sho cocca as guitered to chew the to its new constitution, sho cocca as guitered to chew the dangergumangotic.	Urdinary Reading essent sidewide/enging implications for global drugs polify has http://www.science/org/lights/lig	Information Seeking An end interference of ingregation in the ingregation ingregation ingregation in the ingeneration is indigenous proble too when the ingeneration in the ingeneration in the ingeneration in the ingeneration in the ingeneration is indigenous ingeneration in the ingeneration is indigenous ingeneration ingeneration in the ingeneration in the ingeneration ingeneration in the inge		

С MODEL TRAINING AND HYPERPARAMETERS

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

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

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

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

ADDITIONAL RESULTS D

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



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

Table 3: Validation accuracy results aggregated across 10 cross-validation splits. 'Emb' stands for word embeddings, 'Ling. Feat.' for linguistic word properties. Model performance is compared to the Reading Time baseline using a linear mixed-effects model. In R notation: is_correct \sim $model + (model \mid participant) + (model \mid paragraph)$. Significant gains over this baseline are marked with '*' p < 0.05, '**' p < 0.01 and '***' p < 0.001 in superscript, and significant drops compared to the best model in each regime are marked in subscript with '+'.

Model	Gaze Representation	Text Representation	New Item	New Participant	New Item & Participant	All
Majority Class / Chance Reading Time	e None None	None None	$\begin{array}{c} 50.0 \pm 0.0 \\ 58.9 \pm 0.5_{+++} \end{array}$	50.0 ± 0.0 $58.9 \pm 1.0_{+++}$	50.0 ± 0.0 $60.4 \pm 1.3_{++}$	$\begin{array}{c} 50.0 \pm 0.0 \\ 58.9 \pm 0.5_{+++} \end{array}$
Logistic Regression BiEyeLSTM - No Text	Global Fixations	None None	$\begin{array}{c} 62.6 \pm 0.3^{**}_{+++} \\ 73.2 \pm 0.6^{***}_{+++} \end{array}$	$\begin{array}{c} 60.6 \pm 1.5_{\text{+++}} \\ 64.9 \pm 1.0 \end{array}$	$61.0 \pm 1.8_{++}$ 65.1 ± 1.4	$\begin{array}{c} 61.6 \pm 0.8^{*}_{+++} \\ 68.8 \pm 0.5^{***} \end{array}$
RoBERTa-Eye-W MAG-Eye	Words Words	Emb + Ling. Feat. Emb + Ling. Feat.	$\begin{array}{c} 66.3 \pm 0.4^{***}_{+++} \\ 53.7 \pm 0.2_{+++} \end{array}$	$65.4 \pm 1.1^{**}$ $53.5 \pm 0.5_{+++}$	64.1 ± 1.0 $52.6 \pm 0.6_{+++}$	$\begin{array}{c} 65.8 \pm 0.5^{***}_{+++} \\ 53.5 \pm 0.2_{+++} \end{array}$
PLM-AS Haller RNN	Fixations Order Fixations	Emb Emb	$59.1 \pm 0.5_{+++}$ $62.5 \pm 0.4_{+++}$	$61.1 \pm 0.7_{+++}$ $62.4 \pm 1.1_{++}$	$58.6 \pm 0.9_{+++}$ $61.4 \pm 1.4_{++}$	$\begin{array}{c} 60.1 \pm 0.4_{+++} \\ 62.3 \pm 0.5_{+++} \end{array}$
BEyeLSTM	Fixations	Ling. Feat.	$72.3 \pm 0.6^{***}_{+++}$	65.0 ± 1.3	66.1 ± 1.2	$68.5 \pm 0.6^{***}$
Eyettention RoBERTa-Eye-F PostFusion-Eye	Fixations Fixations Fixations	Emb + Ling. Feat. Emb + Ling. Feat. Emb + Ling. Feat.	$\begin{array}{c} 56.4 \pm 0.8_{+++} \\ \textbf{76.5} \pm \textbf{1.1}^{***} \\ 72.8 \pm 0.8_{+++}^{***} \end{array}$	$\begin{array}{c} 56.6 \pm 0.9_{+++} \\ 64.8 \pm 1.3^{*} \\ 64.6 \pm 1.4^{*} \end{array}$	$\begin{array}{c} 58.6 \pm 1.1_{+++} \\ 65.9 \pm 1.4 \\ 63.6 \pm 1.8 \end{array}$	$56.6 \pm 0.5_{+++} \\ 70.3 \pm 0.7^{***} \\ 68.4 \pm 0.9_{++}^{***}$



1080 E PAIRWISE AGREEMENT BETWEEN MODELS BY EVALUATION REGIME