Beyond the Average Reader: the Reader Embedding Approach

Calogero Jerik Scozzaro, Matteo Delsanto, Daniele P. Radicioni
Department of Computer Science, University of Turin
{calogerojerik.scozzaro, matteo.delsanto, daniele.radicioni}@unito.it

Abstract

Focus of this work is the prediction of reading times as the task is customarily dealt with in literature: that is, by collecting eye-tracking data that are averaged and employed to train learning models. We start by observing that systems trained on average values are ill-suited for the prediction of the reading times for specific subjects, as they fail to account for individual variability and accurately analyze the reading gestures of specific reader groups, or to target specific user needs. To overcome such limitation, that is to predict the reading times for a specific subject, we propose a novel approach based on creating an embedding to compactly describe her/his fixations. Embeddings are used to individuate readers that share same or similar reading behavior from a reference corpus. Models are then trained on values averaged over this subset of similar readers. Experimental results indicate that the proposed approach consistently outperforms its corresponding variants, in which predictions of reading times for specific readers are based on data from all subjects rather than from the most similar ones.[1]

1 Introduction

Eye-tracking reading data are acknowledged as an important resource for both natural language processing and psycholinguistics. The ability to model gaze features is crucial to advance our understanding of language processing (Hollenstein et al., 2021a). Eye-tracking data typically collect information on where readers look at in the form of timestamped fixations. These offer a glimpse on the path of attention deployed by readers providing valuable insights into what is the focus of the attention of the reader through time. Eye-tracking data mostly include measures to record the duration of the first fixation(s) on words, the number of fixations on a given word, and the overall time spent, going back and forth through the reading process, on a specific word. Such measures have been fruitfully employed to investigate linguistic processing mechanisms, such as, e.g., frequency and predictability effects on reading. Based on the analysis of reading times (RTs), we know that it typically takes longer to read infrequent words than frequent ones (Staub et al., 2010), and also context effects have been proven: in left-to-right languages such as English, words that can hardly be predicted based on their left context will also require longer RTs than those that are predictable (Staub, 2011).

Analyzing how text documents are read is relevant from a scientific viewpoint, and also has many applications. Precious insights can be gained about aspects as diverse as lexical access, semantic integration, individual differences including disorders and deficits (Rayner, 1998), and investigating the reading task may shed light on overall words predictability (Smith and Levy, 2013), on the incrementality hypothesis (Fossum and Levy, 2012), on the assessment of competing theories, such as dependency locality theory *vs.* surprisal (Demberg and Keller, 2008), on the interplay between reader's expertise and the features of the text documents (Ashby et al., 2005), and many other intriguing areas at the intersection of computational linguistics and cognitive science. Additionally, many sorts of applications may be drawn that benefit from this kind of research. Being able to predict RTs may impact on applications such as the assessment of linguistic complexity and text readability (Singh et al., 2016; Sarti et al., 2021; Scozzaro et al., 2024b), the design of text simplification strategies (De Martino, 2023) and readers profiling (Scozzaro et al., 2024a), the organization of personalized reading experiences, learning assessments, e.g., in L2 acquisition (Puimège

[1] The code is available at the URL https://github.com/calogero-jerik-scozzaro/beyond_the_average_reader.

Findings of the Association for Computational Linguistics: ACL 2025, pages 15231–15244
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2023), in problem solving skills acquisition (Tóthová et al., 2021).

Eye-tracking data can be complemented with the many analytical tools developed for language analysis to predict reading times and readers' behavior. Different approaches have been proposed, that are based on either boosting methods relying on tree-based algorithms, or on the use of neural approaches casting RTs prediction to a regression task, solved through the fine-tuning of transformer-based language models (Hollenstein et al., 2021a). Other approaches, mainly inspired by probabilistic accounts of language processing, provided evidence that reading times are correlated to the conditional word probabilities estimated by language models (Monsalve et al., 2012). In this view, unpredictability would act as as a major source of processing difficulty in language comprehension: increasing effort would produce a slower reader response (corresponding, e.g., to higher gaze duration) as a consequence of the increased distance intervening between the cognitive resource allocation and the actually encountered input (Levy, 2008; Monsalve et al., 2012).

However, readers may be differently affected by such sources of difficulty, and taking into account such factors is crucial to predict the RTs of specific readers, which would be beneficial to analyze the behavior of specific reader groups or to target specific user needs. One major limitation of the current approaches is that, to the best of our knowledge, all reported work relies on average measures, that can be extracted from reference *corpora* collecting data on readers fixations. For example, if the total reading time —that is, the overall duration of eye fixations received by each word— was considered, existing resources typically provide an average value across all readers. In point of fact, models that are trained on reading times obtained by averaging values over the subjects are not able to accurately account for individual differences in reading behavior. Provided that this approach is in general justified by the inter-subject consistency of the recorded data, it may overlook the fact that different readers adopt different reading strategies, and that the same text can be read in different ways by different readers. If such models are employed to predict RTs, while they will exhibit a good fit with average RTs, they will mostly obtain lower accuracy in predicting the RTs featuring specific readers.

The present study addresses this issue, and is concerned with improving RTs predictions for specific readers. The following novelties are introduced: *i*) a vectorial representation for describing different behavioral (reading) gestures and aptitudes; *ii*) the idea of predicting a specific subject reading times based on the RTs of similar readers; additionally, we explore the output of models based on widely varied assumptions and features, and prove that few readers' data are actually needed to build a vectorial user profile ensuring consistent improvements in the prediction of reading times. In all cases, we compare and contrast experimental results obtained by testing on averaged reading time values (as done in literature) *vs.* the actual values characterizing the reading performance of specific subjects.

We created an array of experiments to test different models and architectures: the proposed approach always outperforms average-based variants in predicting reading times for specific readers. Also remarkably, this approach only requires a fraction of data to overcome the traditional averaging-based variants, thereby allowing for an efficient, resource-saving, and effective way to predict reading times.

2 Related Work

Reading involves two main eye movements, *fixations* and *saccades*. Fixations are brief stops (ranging from 50 to 1500 ms) that typically occur at each word; sometimes even more stops are needed, also depending on words length and on lexical and syntactic complexity. Saccades are fast (from 10 to 100 ms) movements between each two fixations, and are used to reposition the point of focus. In general, individual words are fixated differently: e.g., a pioneering work by Carpenter and Just (1983) reports that 85% content words and 35% function words get fixations. Among the main variables that impact on eye movements, one must additionally consider *i*) words length: shorter (2-3 letter) words are skipped 75% of the time, while longer (8 letters or more) words are fixated almost always (Rayner, 1978); and *ii*) syntactic and conceptual difficulty of the considered text (Jacobson and Dodwell, 1979).

Several measures have been proposed to analyze text reading and processing times. While the total reading time (TRT) is supposed to grasp the time taken by the overall semantic integration (Radach and Kennedy, 2013), two partial and finer-grained measures have been also proposed: the duration of the first fixation (FFD) that allows estimating the cost underlying lexical access (Hofmann et al., 2022), and the number of fixations (NF), which is deemed to report about words integration in the context of what has been read so far (Frazier and Rayner, 1982).

Most work focused on the processes underlying lexical access and semantic integration falls into two broad approaches to model context. In the first case we have models concerned with the semantic relatedness between words and their context: in this setting, reading times are predicted based on the similarity between embeddings describing words and their context. Works adopting the second approach mostly rely on a probabilistic framework whereby words are predicted based on their (left) context. In this view, word predictability should be intended as a function of the probability of a word given the context, and the probability of that word may work, in turn, as a main predictor of reading times: in essence, the less likely the emission of a word, the higher the *surprisal* associated to that word, and the longer the time it requires for readers to process it. Such probabilistic device has also proven useful in distinguishing the linguistic productions of individuals with dementia from those produced by healthy controls (Cohen and Pakhomov, 2020; Colla et al., 2022; Sigona et al., 2025).

In the last few years neural language models gained a central role in analyzing reading as well, since they are able to acquire conditional probability distributions over the lexicon that are also predictive of human processing times. While word length and frequency are widely acknowledged as predictors for determining lexical access, different sorts of language models have been recently compared to analyze and explain syntactic and semantic factors (Hofmann et al., 2022): N-gram models have been found to succeed in capturing short-range lexical access, while models based on recurrent neural networks show better fit in predicting the next-word. Other studies found that surprisal scores are strong predictors of reading times and eye fixations obtained through eye-tracking (Smith and Levy, 2008; Goodkind and Bicknell, 2018), along with a substantial linear relationship between models' next-word prediction accuracy and their ability in predicting reading times (Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023).

A different line of research hypothesizes that the reading process may also be anticipatory (Pimentel et al., 2023): in this view, readers predict upcoming words and, based on their expectations, allocate time to process them. This anticipatory predictability effect is quantified as words' contextual entropy (Hale, 2006), and has been found to be predictive of reading times as well (Linzen and Jaeger, 2015; van Schijndel and Schuler, 2017).

The interaction between gaze and subjective hatefulness rating has been studied by Alacam et al. (2024), who observed that the annotator's gaze provides predictors of their subjective hatefulness rating. Specifically, the TRT and the NF correlate with annotators' subjective hate ratings and improve predictions of text-only hate speech models. The study by Haller et al. (2024) evaluates the predictive power of surprisal and entropy at the individual level by incorporating information on individuals' cognitive capacities and allowing them to modulate the magnitude of surprisal and entropy effects. The findings indicate that including cognitive capacities increases the predictive power of surprisal and entropy on reading times, and that high performance in the psychometric tests is associated with lower sensitivity to predictability effects.

3 Readers Profiling

Eye-tracking features capture different aspects of the reading process, such as fixation duration and saccade length (Schotter and Rayner, 2013). In the conventional approach, multiple participants read the same texts, and the values of each feature are averaged across readers (Hollenstein et al., 2021a, 2022) producing a representation that approximates the behavior of an *average reader*. For example, in the Provo Corpus (Luke and Christianson, 2018), the mean values (complemented by standard deviations) are: 198.14 (107.13) ms for TRT, 0.95 (0.47) for NF, and 139.80 (52.00) ms for FFD. Averaged features are assumed to provide a generalizable representation of reading behavior and are utilized for both training and evaluating models that predict reading times. However, no studies have assessed the performance of these models in predicting the behavior of specific readers. In Section 4.4 an evaluation on this task is presented, where performances of the common approach based on averaging RTs is compared to the results of three reader profiling methods: in this newly proposed approach, the values used to train the model are obtained using only readers who are *similar* to the target reader, rather than all

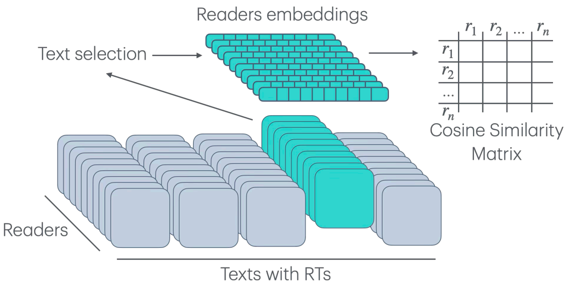

Figure 1: The process of text selection, creation of the readers embeddings, and creation of the cosine similarity matrix. The matrix is symmetric, with values in the range [0, 1], and elements on the main diagonal are by definition set to 1.

readers data. This procedure focuses on an individually carved representation, as the reading times predictions for each reader are based on a subset of readers exhibiting similar behavior. Moreover, this approach is applicable in real-world scenarios where the profiling (along with the selection of the most similar readers) of a specific reader is computed on the basis of a limited set of texts.

A reader embedding is created using the First Fixation Duration (FFD) and the Number of Fixations (NF) for all tokens in the selected texts.[2] The procedure employed to compute the new readers embeddings involves sampling their fixations while reading a text, which was also read by other subjects in a reference corpus. Given a text composed of n lexical items, for each item therein both FFD and NF data are collected, which results in a reader embedding with a size of $2n$ elements.

Cosine similarity is employed as the distance function to compare our embeddings: in particular, a cosine similarity matrix is constructed (Figure 1), where each cell (i,j) represents the cosine similarity score between the embeddings of reader i and reader j. Different selection strategies were considered to individuate the closest readers from the dataset: Top K Similar Readers, Above Threshold, and FFD-NF plot.

Top K Similar Readers The top k most similar readers of a reader i are determined from the cosine similarity matrix by selecting the k highest values in row i, excluding the cell (i,i), which corresponds to the reader itself.

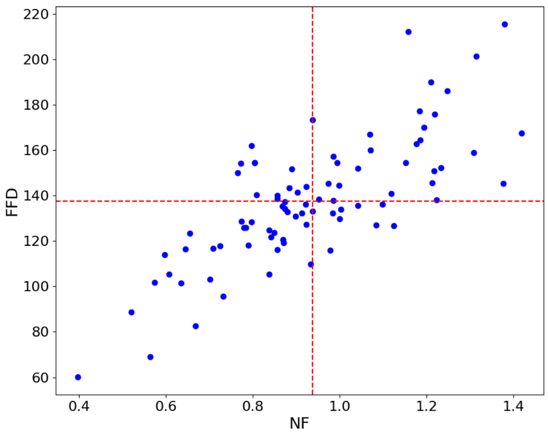

Figure 2: An example of the FFD-NF plot, where each point represents a reader, identified by its average FFD and NF. Red dashed lines report average values over the whole considered population.

Above Threshold A given reader i is associated only with the readers j whose cosine similarity value (cell (i,j)) in the cosine similarity matrix is above a predetermined threshold, thereby picking a variable number of readers ensuring at least some degree of similarity.

FFD – NF plot For each reader, the average FFD and NF on the selected texts are computed, and the results are plotted as points in a FFD-NF plot. Four classes are defined based on mean FFD and NF values, with readers within the same class considered as *similar readers*. In this setting, referring to Figure 2, we identify four classes: class 1 includes readers featured by FFD above average and NF below average; class 2 composed by readers that exhibit both FFD and NF above average; class 3 with subjects whose FFD and NF are both below average; and class 4, with subjects whose FFD is below average and NF above average.

4 Experiments

Our hypothesis is that the prediction of individual readers reading times may be improved by selecting few readers deemed as similar based on their FFDs and NFs. To test such hypothesis we devised an array of experiments in which we compared different approaches to select the most similar readers. Additionally, we tested different models relying on diverse learning and representational rationales. All results reported hereafter are referred to the prediction of the FFD measure.

[2] Eye-tracking data are actually collected based on geometric areas, called ‘Areas of Interest’ (AOIs) that are defined and individuated through an eye-tracking software. Each AOI is a polygon encompassing an attribute of interest within the image, a token in the case of text documents: in the present setting, we thus refer to ‘tokens’ with only marginal loss of descriptive precision.

4.1 Data

The Provo Corpus (Luke and Christianson, 2018) is a corpus of eye-tracking data accompanied by predictability norms. It consists of 55 English texts from a variety of sources, including online news articles, popular science magazines, and public-domain works of fiction, with an average length of 50 words. For this study, we focused exclusively on the eye-tracking data. The texts were read by 84 native English-speaking participants from the Brigham Young University (Provo, Utah, US), and eye-tracking features were recorded.

4.2 Experimental Design

Four experiments were devised.

Average RTs *vs.* user-specific RTs. This trial is aimed at assessing existing models and approaches when trying to predict the actual reading times of specific readers (SPECIFIC READER TIMES), which substantially differs from predicting average reading times (AVERAGE READING TIMES). In this trial we tested in two partially different conditions (always in a leave-one-out setting), whereby predictions were compared both with average values and with the specific user data. This test is of primary relevance for systems aimed at the prediction of real user reading times.

Selection strategy. This experiment is designed to compare different criteria to select the readers that are most useful to build a profile for a specific reader r_i: based on the uniform embedding representation, this step amounts to looking for those users whose reading behavior was most proximal to the reading times collected for r_i.

Profiling and models comparison. This experiment investigates the performance in the SPECIFIC READER TIMES setting of the best selection strategy *vs.* non-profiling strategy across various models.

Sizing of profiling data. In the last trial we tried to characterize the optimal amount of reading data to build effective profiles.

For an extensive evaluation of the entire dataset, we employed a leave-one-out strategy for readers (Figure 3). For each reader, we performed 5 runs with different random seeds, each time splitting the texts not previously involved in the reader embedding step (see Figure 1) into 90% training, 5% validation, and 5% test sets. We evaluated the predictive performance of several models and profiling configurations on the FFD measure. To generate the reader embeddings, we randomly selected 10 texts, corresponding to 481 tokens (17.91% of the total tokens). For the sake of creating reader profiles, the best parameter setting was experimentally determined as follows: $k = 20$ was used for the Top K selection strategy, and a similarity threshold of 0.60 was employed for the 'Above Threshold' method. All eye-tracking features were scaled to the $[0, 1]$ range to improve numerical stability and facilitate neural models training. Models performance was evaluated using three accuracy metrics recently introduced by Lento et al. (2024):

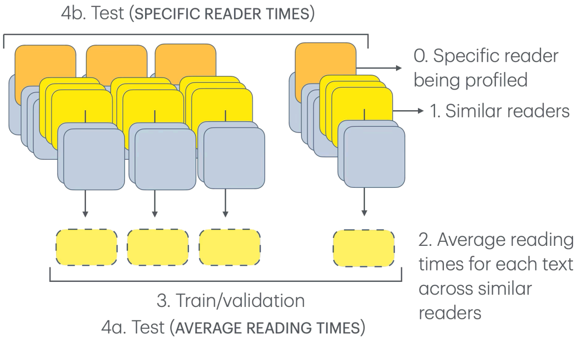
4b. Test (SPECIFIC READER TIMES)
0. Specific reader being profiled
1. Similar readers
2. Average reading times for each text across similar readers
3. Train/validation
4a. Test (AVERAGE READING TIMES)

Figure 3: Schema describing the implemented leave-one-out approach on the readers. After the reader profiling phase (0) and the selection of the most similar subjects (1), we compute values averaged on this group (2). We performed five runs with different seeds, each using a 90/5/5 split on texts: models were trained on the average values (3), and tested on both AVERAGE READING TIMES (4a) and SPECIFIC READER TIMES (4b).

Loss Accuracy (*accL*) reflects the overall similarity between predicted and target values, computed as $1-$MAE (Mean Absolute Error).

Threshold Accuracy (*accT*) evaluates the frequency with which the predicted value falls within a fixed neighborhood threshold of the target value, set to 50 ms, following Lento et al. (2024).

Sensitivity Accuracy (*accS*) measures how often the predicted value is within a dynamically determined threshold, computed as 10% variation of the target value. To handle zero-valued targets, a fixed offset of 25 ms was applied in place of the 10% threshold.

All these metrics are expressed as percentages, with higher values indicating better performance.

Model	AVERAGE READING TIMES accL (σ)	accT (σ)	accS (σ)	SPECIFIC READER TIMES accL (σ)	accT (σ)	accS (σ)
LGBM	93.05 (0.06)	95.29 (0.27)	89.04 (0.44)	67.93 (4.89)	31.25 (6.70)	26.41 (5.73)
MLP	92.46 (0.07)	94.29 (0.30)	87.22 (0.34)	67.76 (4.95)	30.55 (6.78)	26.53 (5.77)
LSTM	84.54 (0.15)	62.89 (0.62)	51.51 (0.84)	64.44 (5.16)	30.38 (8.32)	25.98 (7.35)
LSTM-MLP	92.73 (0.14)	95.42 (0.39)	87.22 (1.07)	67.76 (4.93)	30.97 (6.71)	26.62 (5.74)
BERT	79.18 (0.12)	51.41 (0.28)	41.40 (0.35)	61.79 (4.60)	26.27 (5.07)	20.19 (4.39)
BERT-FT	90.45 (0.07)	87.22 (0.33)	77.38 (0.31)	67.18 (4.82)	31.03 (6.33)	25.74 (5.60)

Table 1: Accuracy scores obtained by employing models acquired through average reading times (all readers data) and tested on AVERAGE READING TIMES and on SPECIFIC READER TIMES. The reported values represent percentage accuracy scores for the three metrics, with standard deviations in parentheses.

4.3 Models

We developed various models for predicting the First Fixation Duration (FFD), including a LightGBM (LGBM) regressor, a basic Multi-Layer Perceptron (MLP), a sequential Long Short-Term Memory (LSTM) network, and BERT models, from a simple architecture with a linear layer stacked on top of a fully fine-tuned version. For all neural models, we used the AdamW optimizer (Loshchilov and Hutter, 2017) with a linear learning rate scheduler based on the number of epochs and Mean Squared Error (MSE) loss. Additionally, we applied gradient clipping with a threshold of 1 and implemented early stopping after 10 epochs without improvement in validation accuracy.

LGBM The LightGBM regressor[3] is based on the gradient boosting framework, and was proven successful in the CMCL 2021 Shared Task on Eye-Tracking Prediction (Hollenstein et al., 2021a; Bestgen, 2021). The employed features include word length, previous word length, word position in the sentence, word frequency, previous word frequency, and word surprisal (Hale, 2016). Surprisal associated to a word w_n is defined as the negative logarithm of the probability of emitting w_n given its history $h = \{w_0, w_1, \ldots, w_{n-1}\}$: $\text{SUR}(w_n) = -\log P(w_n|w_0, w_1, \ldots, w_{n-1})$.

MLP A Multi-Layer Perceptron (MLP), consisting of a single hidden layer was implemented. The input features for each word involve word length, word position, word frequency, and word surprisal, as well as these statistics for the two preceding and two following words. The architectural details are similar to those presented in (Lento et al., 2024), including a single hidden layer with 10 units, sigmoid activation functions, an initial learning rate set to $5 \cdot 10^{-3}$, a batch size of 8, and 1000 epochs.

LSTM Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) represent a more cognitively plausible architecture for this task, as reading is inherently a sequential process (Aurnhammer and Frank, 2019). The input features used are the same as those in the MLP. The LSTM consists of two layers with a hidden dimension of 64, a batch size of 4, and a total of 3000 training epochs.

LSTM-MLP An LSTM model with an MLP on top was also trained using the same input features. The LSTM has a single layer with 64 hidden units, followed by a feed-forward network with tanh activation functions. All other hyperparameters remain consistent with the previously described LSTM model.

BERT Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a Transformer-based encoder architecture. We used a pretrained checkpoint[4] and added a final dense layer on top. This layer, shared across all tokens, projects the model's hidden representation (768-dimensional) to a single output. We evaluated two training configurations: *i*) training only the final layer while keeping the pretrained parameters frozen (referred to as 'BERT' in the following); *ii*) fine-tuning the entire model ('BERT-FT'). For both configurations, we used a learning rate set to $5 \cdot 10^{-5}$, trained for 100 epochs, and set the batch size to 16, as described in (Hollenstein et al., 2021b).

4.4 Results

In the following we report the results obtained through four experiments: Average RTs *vs.* user-specific RTs; Selection strategy; Profiling and mod-

[3] https://lightgbm.readthedocs.io

[4] https://huggingface.co/google-bert/bert-base-cased

SPECIFIC READER TIMES

Sel. Strat.	HM($accL, accT, accS$) (σ)
All	$34.39(5.35)$
Top K	$\mathbf{37.07(5.09)}$
FFD-NF	$34.80(6.38)$
Above T.	$35.50(5.53)$

Table 2: Comparison of the accuracy scores obtained by applying the selection strategies (Sel. Strat.) at stake: harmonic mean of accuracy scores (HM($accL, accT, accS$)) across all models listed in Table 1; more specifically, each number herein represents the harmonic mean of the 3 values listed in SPECIFIC READER TIMES sections of Tables 4–9 averaged over all such tables. Percentage score are complemented by standard deviation values.

els comparison; and Sizing of profiling data.

4.4.1 AVERAGE READING TIMES *vs.* SPECIFIC READER TIMES

We first evaluated the models using the standard pipeline, in which the average reader was computed from all readers data, and training and testing were conducted by employing such averaged representation (in the following, such datum is referred to as AVERAGE READING TIMES). We then assessed their performance in the new RTs prediction task, in which models were trained on the average reader and evaluated on specific readers, not included in the training set (referred to as SPECIFIC READER TIMES). The results of these evaluations are presented in Table 1: while the performance in the AVERAGE READING TIMES is in line with the literature,[5] it significantly drops when trying to predict SPECIFIC READER TIMES. Additionally, the increase in standard deviation values indicates variability in individual reading patterns, suggesting the need for more precise models.

4.4.2 Selection strategy

We then evaluated the three methods for selecting the readers most similar to the specific reader r_i. We randomly picked 10 texts from the entire dataset to generate the embeddings (for r_i and all other readers in the corpus). Table 2 presents the harmonic mean of the values obtained for each profiling configuration (including the non-profiling selection strategy, labeled as 'All'). For each considered strategy, the harmonic mean was calculated across all models and all evaluation metrics, and should be thus intended as a synthetic value for all figures provided in Tables 4-9 (Appendix A). These results reveal that the Top K strategy performs best in the SPECIFIC READER TIMES prediction, and also exhibits the lowest standard deviations, which seems to confirm that it is the most suited to capture the diverse aspects of individual reading behaviors.

SPECIFIC READER TIMES

Model	Sel. Strat.	HM($accL, accT, accS$) (σ)	HM(Δ)
LGBM	All	$36.76(5.77)$	-
(Table 4)	Top K	$39.58(5.06)$	$+7.78$
MLP	All	$37.07(5.86)$	-
(Table 5)	Top K	$39.71(5.39)$	$+7.27$
LSTM	All	$36.01(6.72)$	-
(Table 6)	Top K	$\mathbf{40.08(6.12)}$	$\mathbf{+10.95}$
LSTM-MLP	All	$36.81(5.76)$	-
(Table 7)	Top K	$39.52(5.08)$	$+7.44$
BERT	All	$26.93(4.57)$	-
(Table 8)	Top K	$28.89(4.15)$	$+7.45$
BERT-FT	All	$35.46(5.73)$	-
(Table 9)	Top K	$37.72(5.13)$	$+6.56$

Table 3: Comparison between the non-profiling ('All': by taking into account data from all readers, no profiling is *de facto* adopted) and 'Top K' strategies. The values represent the harmonic mean of the three evaluation metrics (*accL*, *accT*, and *accS*) for the test on SPECIFIC READER TIMES, with the harmonic mean of the standard deviation values shown in parentheses. Delta values indicate the harmonic mean of the performance difference. Complete results for each model are provided in the Appendix (Tables 4–9).

4.4.3 Profiling and models comparison

Given these results, we performed an extensive comparison between the non-profiling and Top K selection strategy, using the harmonic mean of the three evaluation metrics for each model. As shown in Table 3, the Top K strategy consistently outperforms the non-profiling one across all tested models. The improvement in performance is noticeable, with Δ values ranging from $+6.56\%$ in the BERT-FT model to $+10.95\%$ in the LSTM model, with the last model also achieving the highest absolute accuracy rates. To further assess the reliability of these differences, we computed 95% confidence intervals and conducted significance tests. These analyses confirmed that the improvements obtained with the Top K strategy over the non-profiling strategy are statistically significant, with p-values less than 0.05.

[5] The only comparison can be made for the BERT and BERT-FT conditions with the work by Hollenstein et al. (2021b) that report slightly lower figures by experimenting with different datasets.

4.4.4 Sizing of profiling data

Given the significant improvement obtained through the LSTM model, we further investigated the optimal number of texts required for generating reader embeddings. To explore this point, we analyzed the effect of varying the number of texts used in the embedding process. As shown in Figure 4, we found that a set of 10 texts, representing approximately 18% of the total tokens in the dataset, provides an optimal balance between the need of selecting data to find out the closest (and thus more homogeneous) subjects, and ensuring enough data for training purposes. All metrics assessing the output of the Top K strategy confirm that for the present setting adopting the 10 text documents provides the best accuracy; conversely, in the 'All' setting we observe a trend that is independent on the size of the samples for the readers profiling step, and on the balance between such split and training data. More specifically, the curves portrayed in the Figure may be explained with the fact that by resorting to more than 10 texts we end up with a reduced number of training examples and, as a consequence, with decreasing accuracy. Interestingly enough, even when the Top K strategy is employed with a reader embedding created from a single text (comprising just 42 tokens), there is still an improvement in the model's ability to predict the behavior of specific subjects. This suggests that even a small amount of text can provide valuable information for profiling, enhancing the model's ability to cope with previously unseen readers.

5 Conclusions

We started by identifying a key issue: the RTs predictions delivered through systems trained on average values (as is customarily done in literature) are not suited for the prediction of specific subjects. Models trained on average values will mostly learn to predict average values, disregarding (since they were not exposed to them) elements related to individual variability. Being able to predict how a specific user behaves while reading a text excerpt may be beneficial for many different purposes, such as to improve text accessibility (e.g., for specific groups, either based on age or suffering from the same disorder or disease), to stimulate learning and engagement with text, and more in general to improve the overall reading experience in a personalized way.

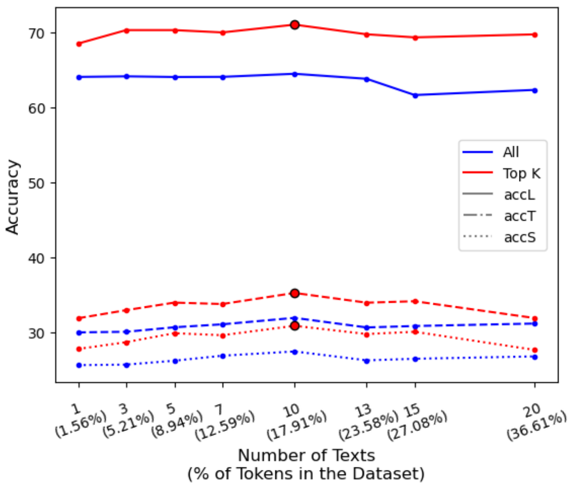

Figure 4: Analysis of the effect of varying the number of texts used to generate embeddings in the LSTM model for the SPECIFIC READER TIMES prediction. Using 10 texts (17.91% of the total tokens in the dataset) yields the optimal balance.

We introduced a novel approach to improve RT predictions on specific readers. First, it builds on a vectorial representation whereby subjects are represented through a collection of fixation data (FFD and NF information is embedded to build such vectors) reporting about how they read texts: this allows creating a shared representational space where readers can be placed, directly compared, and selected based on similarity accounts.[6] In the same spirit as for word embeddings, such vectors can be thought of as points over a multidimensional Euclidean space, where distance acts like a proxy for similarity. Secondly, few similar readers are identified, and for each such subject a small fraction of all available data (10 texts, along with their RTs) are then employed to select the most similar readers to a specific reader. Once these profiles have been created, and the reader under consideration has been equipped with a set of readers with analogous reading behavior, various models have been trained that ensure up to 10% improvement (averaged over three measures, *accL*, *accT*, *accS*) in the accuracy of RTs predictions.

As acknowledged in the first paragraph of Section 4, all our experiments addressed first fixation durations (FFDs); however, we set up further ex-

[6] As regards as the contribution of the FFD and NF metrics, we ran a preliminary ablation experiment, involving the base LSTM architecture, whereby NF information was dropped: in this setting we observed slightly lower figures (overall −0.10 over the three considered evaluation metrics). While we defer to future work a conclusive and extensive experimentation on this point, we presently employ NF information as this allows for better predictions on the number of fixations.

periments, not reported here for the sake of brevity, that show encouraging results also when testing on complementary measures, such as number of fixations (NFs) and total reading times (TRTs). In both cases the improvement in the accuracy in the prediction of the SPECIFIC READER TIMES is in the order of $+4\%$ *accL* (that is, $1-\text{MAE}$), as reported in Table 10 in Appendix. The proposed approach is thus general enough to deal with such additional measures. We also explored the possibility of training models directly on the target reader's own data. Although intuitively appealing, this approach consistently underperformed. Training the LSTM model solely on the reader's data resulted in a harmonic mean of 32.52% ($\pm 4.39\%$) across the three evaluation metrics, which is 9.71% lower than the 'All' selection strategy, and 18.71% lower than the best-performing model. We also tested a variant in which the target reader's data were used for fine-tuning the best model, yielding a harmonic mean of 33.16% ($\pm 4.34\%$), that is 17.27% lower than the same model without fine-tuning. These results suggest that, based on current evidence, using the target reader's data is less effective than leveraging patterns learned from similar readers.

We reported the results obtained in four experiments testing different models and architectures: the proposed approach always outperforms average-based variants in predicting reading times for specific readers, providing evidence that the approach itself is beneficial, regardless of the specific learning algorithm used. Also remarkably, this approach only requires a limited quantity of data to overcome the traditional averaging-based variants, thereby allowing for an efficient, resource-saving, and effective way to predict reading times.

6 Limitations

This study relies on first fixation duration (FFD) as the primary measure: provided that the embeddings employed to pick similar readers are composed of both FFDs and NFs, the models training and testing is conducted on FFDs. While this measure is acknowledged to be associated with lexical access in early cognitive processing, it does not fully capture attentional mechanisms, that would require integrating additional eye-tracking measures, such as NF, TRT, skipping rate, and regression rate. Such limitation will be addressed in future work. Expanding the scope to predict a wider range of eye movement measures would be helpful to better assess the accuracy of profiling models with respect to non-profiling ones, as well as to improve the understanding of the underlying cognitive processes.

Another limitation of this study is the use of different input features for the various models employed in Section 4. This choice was made to align with configurations previously tested in the literature (Bestgen, 2021; Scozzaro et al., 2024a; Lento et al., 2024). While employing different feature sets limits direct comparability between models, our primary objective was to demonstrate the benefits of the reader embedding approach across different architectures, rather than to compare the architectures directly. Future work could explore a more controlled feature selection to enable fairer model comparisons, and to better identify the best-performing architecture.

Finally, the embedding process itself should be considered as a first attempt at building a vectorial representation to describe readers; it might be extended by incorporating information on saccadic features, regressions, and skipping behavior to provide a richer representation of reading patterns and therefore better quality for the resulting embeddings. Despite the fact that the proposed approach relies on a simple (and to some extent limited) representation— which is, strictly speaking, a limitation of this work— its consistent improvement over the traditional counterpart highlights its robustness.

References

Özge Alacam, Sanne Hoeken, and Sina Zarrieß. 2024. Eyes Don't Lie: Subjective Hate Annotation and Detection with Gaze. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 187–205, Miami, Florida, USA. Association for Computational Linguistics.

Jane Ashby, Keith Rayner, and Charles Clifton Jr. 2005. Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A*, 58(6):1065–1086.

Christoph Aurnhammer and Stefan L. Frank. 2019. Comparing Gated and Simple Recurrent Neural Network Architectures as Modelsof Human Sentence Processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 41(0).

Yves Bestgen. 2021. LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 90–96, Online. Association for Computational Linguistics.

Patricia A Carpenter and Marcel Adam Just. 1983. What your eyes do while your mind is reading11this research was partially supported by grant g-79-0119 from the national institute of education and grant mh-29617 from the national institute of mental health. *Eye movements in reading*, pages 275–307.

Trevor Cohen and Serguei Pakhomov. 2020. A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the alzheimer's type. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1946–1957.

Davide Colla, Matteo Delsanto, Marco Agosto, Benedetto Vitiello, and Daniele P Radicioni. 2022. Semantic coherence markers: The contribution of perplexity metrics. *Artificial Intelligence in Medicine*, 134:102393.

Maria De Martino. 2023. Processing effort during reading texts in young adults: Text simplification, readability assessment and preliminary eye-tracking data. In *Proceedings of the 9th Italian Conference on Computational Linguistics, CLiC-it*.

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

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

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 61–69, Montréal, Canada. Association for Computational Linguistics.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2):178–210.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Patrick Haller, Lena Bolliger, and Lena Jäger. 2024. Language models emulate certain cognitive profiles: An investigation of how predictability measures interact with individual differences. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7878–7892, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:730570.

Nora Hollenstein, Emmanuele Chersoni, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2022. CMCL 2022 Shared Task on Multilingual and Crosslingual Prediction of Human Reading Behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 121–129, Dublin, Ireland. Association for Computational Linguistics.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021a. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78, Online. Association for Computational Linguistics.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021b. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 106–123, Online. Association for Computational Linguistics.

J Zachary Jacobson and Peter C Dodwell. 1979. Saccadic eye movements during reading. *Brain and Language*, 8(3):303–314.

Alessandro Lento, Andrea Nadalini, Nadia Khlif, Vito Pirrelli, Claudia Marzi, and Marcello Ferro. 2024. Comparative evaluation of computational models predicting eye fixation patterns during reading: Insights from transformers and simpler architectures.

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

Tal Linzen and T. Florian Jaeger. 2015. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 40(6):1382–1411.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Irene Fernandez Monsalve, Stefan L Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the Effect of Anticipation on Reading Times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

Eva Puimège, Maribel Montero Perez, and Elke Peters. 2023. Promoting l2 acquisition of multiword units through textually enhanced audiovisual input: An eye-tracking study. *Second Language Research*, 39(2):471–492.

Ralph Radach and Alan Kennedy. 2013. Eye movements in reading: Some theoretical context. *The Quarterly journal of experimental psychology*, 66(3):429–452.

Keith Rayner. 1978. Eye movements in reading and information processing. *Psychological bulletin*, 85(3):618.

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

Gabriele Sarti, Dominique Brunato, and Felice Dell'Orletta. 2021. That looks hard: Characterizing linguistic complexity in humans and language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–60.

Elizabeth Schotter and Keith Rayner. 2013. Eye movements in reading. In *Eye tracking in audiovisual translation*.

Calogero J Scozzaro, Davide Colla, Matteo Delsanto, Antonio Mastropaolo, Enrico Mensa, Luisa Revelli, Daniele P Radicioni, et al. 2024a. Legal text reader profiling: Evidences from eye tracking and surprisal based analysis. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024*, pages 114–124. ELRA and ICCL.

Calogero J Scozzaro, Matteo Delsanto, Antonio Mastropaolo, Enrico Mensa, Luisa Revelli, Daniele P Radicioni, et al. 2024b. On the reform of the italian constitution: an interdisciplinary text readability analysis. In *CEUR WORKSHOP PROCEEDINGS*, 3877. CEUR-WS.

Francesco Sigona, Daniele P Radicioni, Barbara Gili Fivela, Davide Colla, Matteo Delsanto, Enrico Mensa, Andrea Bolioli, and Pietro Vigorelli. 2025. A computational analysis of transcribed speech of people living with dementia: The anchise 2022 corpus. *Computer Speech & Language*, 89:101691.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajakrishnan Rajkumar. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*, pages 202–212.

Nathaniel J Smith and Roger Levy. 2008. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

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

Adrian Staub. 2011. The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic bulletin & review*, 18:371–376.

Adrian Staub, Sarah J White, Denis Drieghe, Elizabeth C Hollway, and Keith Rayner. 2010. Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5):1280.

Martina Tóthová, Martin Rusek, and Vlastimil Chytry. 2021. Students' procedure when solving problem tasks based on the periodic table: An eye-tracking study. *Journal of chemical education*, 98(6):1831–1840.

Marten van Schijndel and William Schuler. 2017. Approximations of predictive entropy correlate with reading times. In *Proceedings of the 39th Annual*

Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017. cognitivesciencesociety.org.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *CoRR*, abs/2006.01912.

A Appendix

This appendix provides additional performance details for all the models evaluated in this study. Tables 4, 5, 6, 7, 8, and 9 display the performance of the LGBM, MLP, LSTM, LSTM-MLP, BERT, and BERT-FT models, respectively.

Table 10 shows a comparison between the 'Top K' and 'All' selection strategies using the best-performing model (LSTM) for predicting total reading time (TRT) and number of fixations (NF) in the SPECIFIC READER TIMES setting.

Sel. Strat.	AVERAGE READING TIMES			SPECIFIC READER TIMES			Δ SPECIFIC READER TIMES		
	accL (σ)	accT (σ)	accS (σ)	accL (σ)	accT (σ)	accS (σ)	accL	accT	accS
All	**92.39 (0.07)**	**93.23 (0.31)**	**87.23 (0.47)**	67.88 (4.90)	32.31 (6.82)	27.83 (5.90)	-	-	-
Top K	91.71 (0.46)	85.94 (1.33)	76.87 (1.79)	**73.52 (3.97)**	**34.96 (6.19)**	**29.77 (5.57)**	**+8.31**	**+8.20**	**+6.97**
FFD-NF	90.99 (0.96)	86.24 (6.79)	76.64 (8.38)	70.78 (5.17)	34.00 (6.68)	27.78 (6.84)	+4.27	+5.23	−0.18
Above T.	92.26 (0.89)	90.73 (8.90)	84.19 (9.57)	69.89 (4.79)	33.03 (6.28)	28.50 (5.69)	+2.96	+2.23	+2.41

Table 4: Performance of the LGBM model in predicting both the AVERAGE READING TIMES and SPECIFIC READER TIMES. The last three columns report the difference (Δ) in performance for each selection strategy compared to the 'All' strategy in the SPECIFIC READER TIMES setting. Percentage score are complemented by standard deviation values. The best value in each column is highlighted in bold.

Sel. Strat.	AVERAGE READING TIMES			SPECIFIC READER TIMES			Δ SPECIFIC READER TIMES		
	accL (σ)	accT (σ)	accS (σ)	accL (σ)	accT (σ)	accS (σ)	accL	accT	accS
All	**92.28 (0.07)**	**92.72 (0.26)**	**86.60 (0.35)**	67.86 (4.94)	32.35 (6.98)	28.35 (6.01)	-	-	-
Top K	91.13 (0.79)	82.98 (3.17)	74.17 (2.93)	**73.45 (3.99)**	**34.67 (6.88)**	**30.22 (6.22)**	**+8.24**	**+7.17**	**+6.60**
FFD-NF	90.62 (0.45)	85.03 (6.25)	76.07 (6.87)	70.72 (5.14)	34.62 (7.03)	28.26 (7.78)	+4.21	+7.02	−0.32
Above T.	91.91 (0.91)	89.44 (8.68)	82.76 (9.44)	69.83 (4.83)	33.19 (6.83)	29.02 (5.88)	+2.90	+2.60	+2.36

Table 5: Performance of the MLP model in predicting both the AVERAGE READING TIMES and SPECIFIC READER TIMES. The last three columns report the difference (Δ) in performance for each selection strategy compared to the 'All' strategy in the SPECIFIC READER TIMES setting. Percentage score are complemented by standard deviation values. The best value in each column is highlighted in bold.

Sel. Strat.	AVERAGE READING TIMES			SPECIFIC READER TIMES			Δ SPECIFIC READER TIMES		
	accL (σ)	accT (σ)	accS (σ)	accL (σ)	accT (σ)	accS (σ)	accL	accT	accS
All	83.91 (0.11)	**60.03 (0.55)**	**49.18 (0.45)**	64.54 (5.17)	31.91 (8.36)	27.42 (7.48)	-	-	-
Top K	**84.70 (1.11)**	54.27 (3.67)	45.73 (4.11)	**71.11 (4.12)**	**35.24 (8.51)**	**30.85 (7.69)**	**+10.18**	**+10.44**	**+12.51**
FFD-NF	83.83 (1.85)	56.58 (3.66)	45.52 (6.18)	67.37 (6.20)	30.86 (11.20)	26.30 (10.35)	+4.38	−3.29	−4.08
Above T.	84.21 (0.75)	58.36 (4.74)	47.69 (4.01)	66.93 (5.28)	33.36 (8.08)	28.93 (7.32)	+3.70	+4.54	+5.51

Table 6: Performance of the LSTM model in predicting both the AVERAGE READING TIMES and SPECIFIC READER TIMES. The last three columns report the difference (Δ) in performance for each selection strategy compared to the 'All' strategy in the SPECIFIC READER TIMES setting. Percentage score are complemented by standard deviation values. The best value in each column is highlighted in bold.

Sel. Strat.	AVERAGE READING TIMES			SPECIFIC READER TIMES			Δ SPECIFIC READER TIMES		
	accL (σ)	accT (σ)	accS (σ)	accL (σ)	accT (σ)	accS (σ)	accL	accT	accS
All	**92.01 (0.16)**	**91.66 (0.64)**	**84.75 (0.72)**	67.76 (4.90)	32.74 (6.80)	27.63 (5.90)	-	-	-
Top K	91.47 (0.50)	84.81 (1.57)	75.80 (2.10)	**73.44 (3.96)**	**34.91 (6.26)**	**29.71 (5.60)**	**+8.38**	**+6.63**	**+7.53**
FFD-NF	90.82 (0.96)	85.25 (6.68)	75.96 (8.04)	70.72 (5.15)	33.97 (6.53)	27.73 (6.91)	+4.37	+3.76	+0.36
Above T.	91.94 (0.84)	89.21 (8.42)	82.33 (8.90)	69.83 (4.78)	33.36 (6.54)	28.68 (5.72)	+3.05	+1.89	+3.80

Table 7: Performance of the LSTM-MLP model in predicting both the AVERAGE READING TIMES and SPECIFIC READER TIMES. The last three columns report the difference (Δ) in performance for each selection strategy compared to the 'All' strategy in the SPECIFIC READER TIMES setting. Percentage score are complemented by standard deviation values. The best value in each column is highlighted in bold.

Sel. Strat.	AVERAGE READING TIMES			SPECIFIC READER TIMES			Δ SPECIFIC READER TIMES		
	accL (σ)	accT (σ)	accS (σ)	accL (σ)	accT (σ)	accS (σ)	accL	accT	accS
All	77.13 (0.11)	**45.70 (0.21)**	**38.12 (0.31)**	60.57 (4.77)	23.91 (4.95)	18.84 (4.08)	-	-	-
Top K	**78.06 (0.81)**	39.56 (1.20)	31.95 (1.26)	**67.17 (3.57)**	**25.25 (4.71)**	**20.26 (4.33)**	**+10.90**	**+5.60**	**+7.54**
FFD-NF	77.58 (0.84)	44.28 (4.19)	33.99 (2.13)	63.69 (5.08)	24.17 (5.43)	18.61 (5.64)	+5.15	+1.09	−1.22
Above T.	77.43 (1.06)	43.75 (3.06)	35.76 (2.95)	62.83 (5.34)	24.75 (4.71)	19.57 (3.84)	+3.73	+3.51	+3.87

Table 8: Performance of the BERT model in predicting both the AVERAGE READING TIMES and SPECIFIC READER TIMES. The last three columns report the difference (Δ) in performance for each selection strategy compared to the 'All' strategy in the SPECIFIC READER TIMES setting. Percentage score are complemented by standard deviation values. The best value in each column is highlighted in bold.

Sel. Strat.	AVERAGE READING TIMES			SPECIFIC READER TIMES			Δ SPECIFIC READER TIMES		
	accL (σ)	accT (σ)	accS (σ)	accL (σ)	accT (σ)	accS (σ)	accL	accT	accS
All	**89.62 (0.08)**	**83.65 (0.48)**	**73.45 (0.37)**	67.02 (4.94)	31.29 (6.82)	26.51 (5.73)	-	-	-
Top K	89.28 (0.60)	74.74 (2.16)	64.79 (2.71)	**72.72 (3.92)**	**33.08 (6.38)**	**28.12 (5.80)**	**+8.50**	**+5.72**	**+6.07**
FFD-NF	88.34 (1.01)	75.75 (6.08)	64.18 (7.02)	69.75 (5.23)	31.97 (7.16)	26.34 (7.27)	+4.07	+2.17	−0.64
Above T.	89.59 (0.72)	80.80 (7.87)	70.69 (7.31)	69.06 (4.81)	31.99 (6.52)	27.31 (5.55)	+3.04	+2.24	+3.02

Table 9: Performance of the BERT-FT model in predicting both the AVERAGE READING TIMES and SPECIFIC READER TIMES. The last three columns report the difference (Δ) in performance for each selection strategy compared to the 'All' strategy in the SPECIFIC READER TIMES setting. Percentage score are complemented by standard deviation values. The best value in each column is highlighted in bold.

Sel. Strat.	TRT	NF
	accL (σ)	accL (σ)
All	82.53 (3.13)	81.47 (2.71)
Top K	85.80 (2.10)	84.54 (2.06)

Table 10: Performance and standard deviation of the LSTM model in predicting total reading time (TRT) and number of fixations (NF) using the 'All' and 'Top K' selection strategies in the SPECIFIC READER TIMES setting. Results show a 3.96% improvement in TRT and a 3.77% improvement in NF when using the 'Top K' strategy over 'All'; both differences are statistically significant for $p < 0.05$.