# How Well Can AI Models Generate Human Eye Movements During Reading?

**Ivan Stebakov and Ilya Pershin**
Research Center of the Artificial Intelligence Institute
Innopolis University
Innopolis, Russia
i.stebakov@innopolis.ru, i.pershin@innopolis.ru

## Abstract

Eye movement analysis has become an essential tool for studying cognitive processes in reading, serving both psycholinguistic research and natural language processing applications aimed at enhancing language model performance. However, the scarcity of eye-tracking data and its limited generalizability constrain data-driven approaches. Synthetic scanpath generation offers a potential solution to these limitations. While recent advances in scanpath generation show promise, current literature lacks systematic evaluation frameworks that comprehensively assess models' ability to reproduce natural reading gaze patterns. Existing studies often focus on isolated metrics rather than holistic evaluation of cognitive plausibility. This study presents a systematic evaluation of contemporary scanpath generation models, assessing their capacity to replicate natural reading behavior through comprehensive scanpath analysis. We demonstrate that while synthetic scanpath models successfully reproduce basic gaze patterns, significant limitations persist in capturing part-of-speech dependent gaze features and reading behaviors. Our cross-dataset comparison reveals performance degradation in three key areas: generalization across text domains, processing of long sentences, and reproduction of psycholinguistic effects. These findings underscore the need for more robust evaluation protocols and model architectures that better account for psycholinguistic complexity. Through detailed analysis of fixation sequences, durations, and reading patterns, we identify concrete pathways for developing more cognitively plausible scanpath generation models.

## 1 Introduction

Eye movements during reading reflect readers' attention (Rayner, 1998), processing difficulty, and information integration (Rayner, 2009; Clifton et al., 2016). Thus, eye-tracking data provides a rich source of insights into human language processing. Models derived from gaze data not only shed light on attention and comprehension but also have practical applications in readability estimation (Klein et al., 2025), educational technology (da Silva Soares Jr et al., 2023), and cognitively plausible NLP (Barrett et al., 2018). However, the utility of such data is constrained by its limited availability. Synthetic data generation has emerged as a critical solution across domains, particularly for enhancing deep learning models in data-scarce scenarios. Recently, eye-tracking models for reading have gained traction in machine learning research.

Studies suggest that cognitive models like E-Z Reader (Reichle et al., 2003), which simulate gaze patterns during reading, can improve language models in standard NLP tasks (Sood et al., 2020). Modern approaches follow two key paradigms: Predicting aggregated eye-tracking features (e.g., fixation durations) (Li and Rudzicz, 2021; Hollenstein et al., 2021; Srivastava, 2022); Generating scanpaths—temporal sequences of word fixations with durations (Deng et al., 2023b; Khurana et al., 2023; Bolliger et al., 2025). For instance, Lopez-Cardona et al. (2024) used a gaze feature prediction model (Li and Rudzicz, 2021) to train a reward model by concatenating predicted eye-tracking features with contextual embeddings. Evaluations on the OASST1 and Helpsteer2 datasets showed significant accuracy improvements over baselines. By generating scanpaths, these models can additionally compute reading-related gaze features, thereby increasing their utility. Scanpaths enable modeling of gaze phenomena such as refixations (repeated word fixations) and regressive saccades (backward eye movements). The latter has drawn increasing attention, as it not only enhances the performance of established models like E-Z Reader (Reichle et al., 2003) and SWIFT (Engbert et al., 2002) but also shows promise for downstream NLP applica-

*Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 148–162
November 9, 2025 ©2025 Association for Computational Linguistics

tions.

Despite progress, existing studies lack a comprehensive analysis of generated scanpaths and standardized evaluation metrics. For example: Deng et al. (2023b) proposed Eyettention, evaluated using Normalized Levenshtein Distance (NLD) (Levenshtein, 1966). However, NLD ignores fixation durations, lacks spatial sensitivity, and has limited interpretability. Eyettention has been applied to improve NLP task performance on the GLUE benchmark (Wang et al., 2019) by reordering text to mimic natural reading patterns (Deng et al., 2023a, 2024; Kiegeland et al., 2024). Khurana et al. (2023) introduced ScanTextGAN, employing both NLD and MultiMatch (Jarodzka et al., 2010). Yet, Kümmerer and Bethge (2021) demonstrated that MultiMatch can favor incorrect models over ground truth. ScanTextGAN's integration of predicted scanpaths (via LSTM and multi-head attention) improved performance on GLUE, sentiment analysis, and sarcasm detection (Mishra et al., 2016). Bolliger et al. (2023) developed ScanDL later extended to ScanDL 2.0 (Bolliger et al., 2025), using two separate models for fixation sequences and durations. They use ScaSim (von der Malsburg and Vasishth, 2011), a metric penalizing spatial/temporal deviations between fixations. While ScaSim addresses NLD's limitations, their reproducibility analysis excluded fixation durations, and no comparison was made against randomly generated scanpaths for ScaSim or gaze features.

This work synthesizes prior research on scanpath generation models and addresses their limitations. Our contributions are: 1) A unified evaluation framework for scanpath generation models, covering critical gaze properties. 2) Quantitative benchmarking of publicly available models using this framework. 3) Analysis of scanpath generation models weaknesses to guide future improvements.

## 2 Methodology

The core task involves predicting a complete scanpath representation $\mathbf{S} = \langle s_1, ..., s_n \rangle$, where each point $s_i$ consists of both fixation positions $\mathbf{F} = \langle f_1, ..., f_n \rangle$ and corresponding durations $\mathbf{D} = \langle d_1, ..., d_n \rangle$, given an input sentence $\mathbf{W} = \langle w_1, ..., w_m \rangle$. Here, each fixation position $f_i$ corresponds to the index $j$ (where $1 \leq j \leq m$) of the fixated word $w_j$ in the sentence. Contemporary models demonstrate the capability to generate diverse scanpaths for identical text inputs, effectively simulating individual differences in reading patterns among human subjects. Our analysis focuses on publicly available implementations of three existing approaches.

The E-Z Reader model[1] implements a cognitive architecture that incorporates multiple psycholinguistic variables including lexical frequency, word predictability, and integration time parameters. This framework provides a comprehensive computational account of the interaction between perceptual, cognitive, and oculomotor processes during reading, explicitly modeling the mechanisms underlying saccade programming and execution that produce characteristic eye movement patterns.

Eyettention[2] adopts a probabilistic approach to predict subsequent fixation locations through the conditional distribution $P(f_i|\mathbf{W}, s_1, ..., s_{i-1})$, where the model considers both the textual input $\mathbf{W}$ and the preceding scanpath segment $\langle s_1, ..., s_{i-1} \rangle$ that includes landing position information. During inference, the model utilizes only the fixation position component of this history. The model architecture employs parallel processing streams: A Word-Sequence Encoder leveraging BERT embeddings (Devlin et al., 2019) with word-level aggregation, enhanced through bidirectional LSTM processing and supplemented with explicit word length features; A Fixation-Sequence Encoder implemented as a unidirectional LSTM that processes concatenated representations of fixation word embeddings, normalized duration values, and within-word landing positions. These parallel representations are integrated through a cross-attention mechanism, with final predictions generated by a ReLU-activated fully-connected decoder network. The model produces scanpaths through iterative sampling from a probability distribution over possible saccade targets, including both progressive (forward) and regressive (backward) movements within the range $-M + 1, ..., M$ (where $M$ denotes maximum sentence length), plus an additional end-of-scanpath marker class, resulting in a $2M + 1$-dimensional output space. Training optimizes the mean negative log-likelihood objective.

ScanDL 2.0[3] introduces a modular architecture comprising two specialized components: The ScanDL Module implements a discrete diffusion sequence-to-sequence model for sequence genera-

[1] https://github.com/jakdot/ezreader-python

[2] https://github.com/aeye-lab/Eyettention

[3] https://github.com/DiLi-Lab/ScanDL-2.0

| Dataset | # Uniuqe sentence | # Readers | Sentence length | # Samples |
|---|---|---|---|---|
| CELER | 5486 | 69 | up to 22 | ∼10.7k |
| ZuCO | 700 | 12 | 3-62 | ∼8.4k |

Table 1: Summary of the eye-tracking while reading datasets.

tion, transforming input text (represented through word indices, BERT embeddings, and positional encodings) into realistic fixation sequences through iterative noise addition and denoising via transformer encoder; The Fixation Duration Module employs a transformer-based sequence-to-sequence architecture to predict temporal durations for fixations, using GPT-2-derived contextual embeddings that are dynamically reordered according to the scanpath. The ScanDL Module's training incorporates both variational lower bound (VLB) optimization and mean squared error minimization between predicted and ground truth embeddings. The Fixation Duration Module utilizes a 12-layer transformer encoder with self-attention mechanisms, followed by ReLU-activated fully-connected layers, trained via mean squared error minimization on duration predictions. This decoupled architecture permits independent training and deployment of each module, offering significant flexibility in practical applications.

# 3 Experiments

## 3.1 Datasets

The models were trained using the CELER dataset. The CELER dataset includes eye-tracking while reading data from 69 readers for 5,486 sentences. Each participant in CELER read 156 newswire sentences from the Wall Street Journal. Of these, 78 sentences were common to all readers, while the remaining 78 were unique to each individual reader. The maximum sentence length is 22 words. The CELER dataset contains approximately 10,700 samples.

For additional verification, the ZuCO dataset (Laurinavichyute et al., 2019) was used. The ZuCO dataset includes eye-tracking while reading data from 12 readers for 400 sentences from movie reviews (positive, negative or neutral) and 300 Wikipedia sentences with specific relations. The sentence length ranges from 3 to 62 words. The ZuCO dataset contains approximately 8,400 samples. Table 1 presents a summary of the eye-tracking datasets used in this study.

The CELER dataset was divided into 5 folds and a test set, following a new reader/new sentence split. Each fold and the test set included approximately 11-12 readers and 13 sentences. Unique sentences were used only in the training set. The same split was used for all models. Metrics for Within-Dataset Evaluation (Section 3.4) were calculated on the test set. Metrics for Cross-Dataset Evaluation (Section 3.5) were calculated on the entire ZuCO dataset.

## 3.2 Metrics

As mentioned earlier, the ScaSim metric (von der Malsburg and Vasishth, 2011), specifically designed for quantitative assessment of differences between scanpaths, represents the preferred choice. Following (Bolliger et al., 2025), we configured ScaSim Base for a constant y-coordinate and computed two normalized versions: ScaSim Fix (normalized by the number of fixations in both scanpaths) and ScaSim Dur (normalized by the total duration of all fixations). To evaluate the reproducibility of gaze features based on predictions, we calculated the mean absolute error (MAE) and Pearson correlation coefficient (PCC). We examined 23 distinct gaze features capturing various eye movement characteristics: fixation duration, reading time, saccade amplitude, fixation count, regressions, and word skipping. The complete list and description of features appears in Appendix A. The MAE and PCC metrics were applied to features computed in three processing modes: without aggregation (Base), word aggregation across readers (Word), and sentence aggregation across readers and sentences (Sent). All feature values were normalized to a 0-100 scale. For improved readability, we report prediction accuracy as $100 - MAE$ in all experimental results. We additionally employed Normalized Levenshtein Distance (NLD) to assess fixation sequence similarity. The Levenshtein distance was normalized by the maximum sequence length: $NLD = LD(S_1, S_2)/\max(|S_1|, |S_2|)$. All reported metrics represent averages across models trained on the 5 folds.

To assess the models' ability to replicate human-like gaze behavior, we analyzed their capacity to reproduce established psycholinguistic phenomena. We evaluated correlations between gaze features and three key predictors: word length, surprisal (computed using GPT-2 base (Radford et al., 2019)), and lexical frequency (obtained via the wordfreq library[4]). Furthermore, we investigated part-of-speech effects on gaze distribution using the NLTK library[5], calculating average gaze features per grammatical category. The analysis focused on six core measures: first-pass reading time (FPRT), re-reading time (RRT), total fixation time (TFT), first-pass fixation count (FPFC), first-pass regression (FPReg), and skipping rate (SR). These word-aggregated features capture fundamental reading patterns: word processing time, fixation frequency, word skipping probability, and regression likelihood.

Model comparisons employed two human baselines: Human Shuffled (shuffled test set scanpaths) and Human Train-Val (random scanpaths from 5-fold readers). The Human Shuffled baseline reveals differences in gaze patterns among random readers within the test sample. However, word- and sentence-level aggregated metrics become unavailable for this mode, as gaze features are calculated across all readers from the test set. To address this gap in evaluation, the Human Train-Val baseline is employed. In this case, for each fold, a random set of readers is selected, matching the number of readers in the test set. Regarding the remaining metrics, both Human Shuffled and Human Train-Val demonstrate variations in metrics depending on the reader set. The Human Train-Val baseline enables MAE/PCC comparison for reader-averaged gaze features. We also included two random baselines: Uniform Fixations - random uniform fixation positions with dataset-derived scanpath lengths; Random Saccades - random saccades ranging from -1 to +2 words, terminating at sentence end. The probability of saccades of length -1 and 0 is 13%, and the probability of direct saccades of length 1 and 2 is 37%. Both random baselines generated fixation durations from normal distributions parameterized by training data statistics. The objective of evaluating random predictions is to demonstrate that the generated gaze sequences from models are not random and differ significantly from random predictions. Furthermore, such evaluation can establish a baseline of adequacy for generative models. For metrics that provide an indirect assessment of quality, evaluation on random predictions can shed light on the utility of the metric itself.

### 3.3 Gaze model

The E-Z Reader model requires three key word parameters to be specified: frequency, predictability, and integration time. The lexical frequency values were obtained using the wordfreq library[6]. Predictability values were derived using GPT-2 base (Radford et al., 2019). The integration time parameter was set to the average value of 25 ms as reported in (Reichle and Sheridan, 2015).

Since ScanDL 2.0 comprises two independent models - the ScanDL Module and Fixation Duration Module - we analyze them separately in this study. For clarity, we refer to the ScanDL Module simply as ScanDL, and the Fixation Duration Module as Scan2Dur. Notably, Scan2Dur is also applied to enhance the predictions of the Eyettention model. This approach combines the fixation position predictions from both Eyettention and ScanDL with duration predictions from Scan2Dur. For model implementation, we used the original code published in the respective papers for Eyettention, ScanDL and Scan2Dur. The only modifications made involved adapting the training and testing samples to our experimental setup while maintaining all other parameters and architectural choices as specified in the original implementations. ScanDL also was chosen as the reference model since it achieves the strongest overall performance in the available studies.

### 3.4 Within-Dataset Evaluation

The results are presented in Table 2. It should be noted that significant improvements in metrics compared to Human Baselines may indicate insufficient diversity in generated scanpaths rather than superior performance. However, in this case, the differences are not substantial. Moreover, it would be incorrect to claim that generation models surpass human performance, as eye movements represent a natural cognitive process.

The metrics show that Human Train-Val and Human Shuffled demonstrate minor differences, suggesting that even small samples of readers can exhibit noticeable variations in gaze patterns. For the NLD metric, both E-Z Reader and ScanDL outperform Human Train-Val and show comparable results, though further analysis reveals signifi-

[4] https://github.com/rspeer/wordfreq

[5] https://github.com/nltk/nltk

[6] https://github.com/rspeer/wordfreq

|  | NLD | ScaSim |  |  | MAE |  |  | PCC |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  |  | Base | Dur | Fix | Base | Word | Sent | Base | Word | Sent |
| Random Uniform | 0.86↑ | 3615↑ | 0.52↑ | 111.80↑ | 79.30↓ | 72.24↓ | 24.76↓ | 0.13↓ | 0.25↓ | 0.68↑ |
| Random Saccades | 0.66↑ | 2872↑ | 0.45↑ | 96.77 | 83.20↓ | 82.85↓ | 72.45↑ | 0.00↓ | 0.00↓ | -0.06↓ |
| E-Z reader | 0.58 | 3705↑ | 0.47↑ | 146.77↑ | 84.76↓ | 79.80↓ | 45.37↓ | 0.10↓ | 0.33↓ | 0.20 |
| Eyettention | 0.65↑ | 2544↑ | 0.45 | 84.27 | 84.55↓ | 84.83 | 61.36↑ | 0.11↓ | 0.44↓ | 0.55 |
| ScanDL | 0.58* | 2395* | 0.44* | 85.45* | 86.43* | 84.94* | 54.45* | 0.16* | 0.50* | 0.44* |
| Human Train-Val | 0.60 | 2689↑ | 0.42↓ | 92.20 | 85.95 | 88.84↑ | 73.35↑ | 0.20↑ | 0.71↑ | 0.80↑ |
| Human Shuffled | 0.56↓ | 2814↑ | 0.39 | 86.76↑ | 86.47 | - | - | 0.23↑ | - | - |

Table 2: Metrics for the predicted scanpaths on the CELER dataset. To assess statistical reliability, we conducted paired t-tests (p<0.05) on metric values across folds, using ScanDL as the reference model. Significant differences are indicated with ↑/↓, where ↑ denotes an increase and ↓ a decrease relative to ScanDL (marked with *)

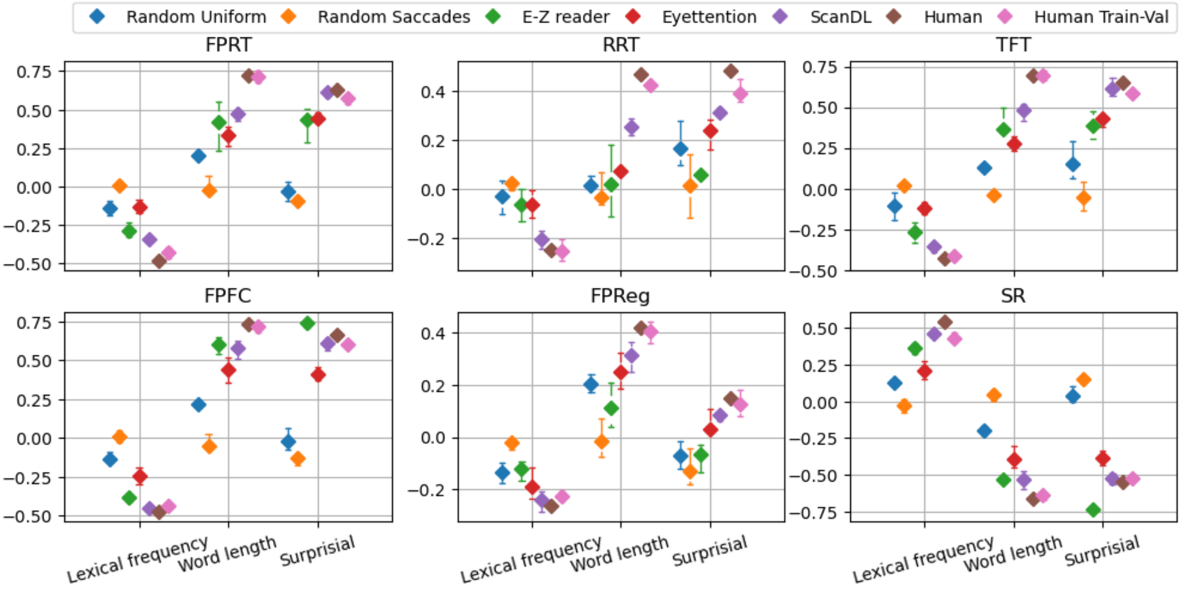

Figure 1: Pearson correlation coefficient between word features and gaze features on CELER dataset.

cant differences in their performance. The ScanDL model achieves the best results for ScaSim, ScaSim Fix, and MAE metrics, while Human Baselines remain superior for other metrics. The Eyettention model shows performance similar to ScanDL with minor variations: ScanDL leads in MAE Base, both models are comparable in MAE word, while Eyettention leads in MAE sent. However, Eyettention underperforms in NLD. Compared to Human Baselines, both Eyettention and ScanDL show noticeable gaps in PCC and MAE Sent metrics, with smaller differences in MAE Word, and only Eyettention trailing in MAE Base. The E-Z Reader model underperforms in all metrics except NLD and MAE Base.

The Random Saccades baseline performs worse than ScanDL and Eyettention across most metrics, with PCC approaching zero, yet shows comparable results for ScaSim Dur and MAE. While Random Fixations generally underperforms, it achieves results similar to the main models in PCC Base and PCC Sent. These observations demonstrate that relying on individual metrics may lead to incorrect model evaluations. Considering all metrics collectively, both ScanDL and Eyettention show the closest alignment with Human Baselines, with ScanDL performing slightly better. However, all models demonstrate challenges in accurately reproducing gaze features, highlighting the importance of considering multiple gaze feature metrics. Detailed metrics for individual features are provided in Appendix B.

Figure 1 presents PCC values between word features and gaze characteristics. The plot shows that Random Models demonstrate near-zero correlations. Among the evaluated models, ScanDL shows the closest alignment with Human Baselines, while E-Z Reader and Eyettention show varying

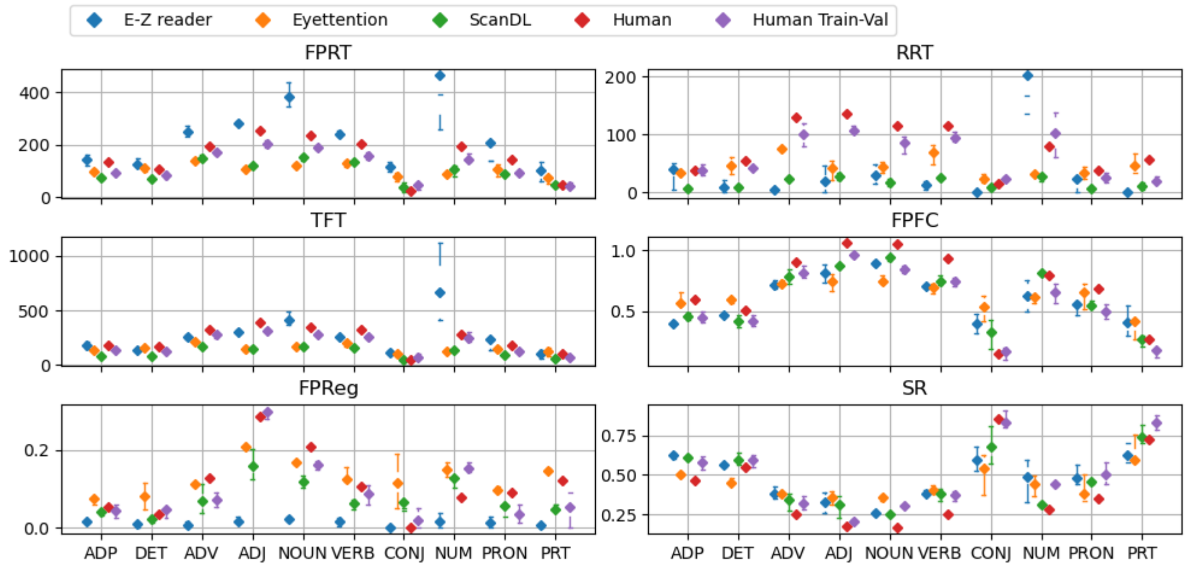

Figure 2: Mean gaze features with respect to POS tagging for CELER dataset.

| | NLD | ScaSim | | | MAE | | | PCC | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Base | Dur | Fix | Base | Word | Sent | Base | Word | Sent |
| Random Uniform | 0.90↑ | 4479.38↑ | 0.66↑ | 106.93↑ | 84.18↓ | 82.11↓ | 66.51↓ | 0.10↑ | 0.20↑ | 0.52↑ |
| Random Saccades | 0.70 | 4052.95↑ | 0.57↑ | 94.81↑ | 84.42↓ | 83.00↓ | 67.52↓ | 0.01↓ | 0.04↓ | -0.05↓ |
| E-Z reader | 0.64↓ | 10061.06↑ | 0.66↑ | 266.74↑ | 80.16↓ | 65.36↓ | 16.91↓ | 0.06↓ | 0.15↓ | 0.08↓ |
| Eyettention | 0.74↑ | 2609.36↑ | 0.54 | 66.68 | 85.93↓ | 85.98 | 82.76↑ | 0.04↓ | 0.13↓ | 0.32 |
| ScanDL | 0.70* | 2285.85* | 0.52* | 66.24* | 87.20* | 85.88* | 80.78* | 0.07* | 0.18* | 0.33* |
| Human Train-Val | 0.66↓ | 2515.39↑ | 0.46↓ | 53.04↓ | 88.02↑ | 90.48↑ | 85.95↑ | 0.22↑ | 0.60↑ | 0.58↑ |
| Human Shuffled | 0.52↓ | 1674.67↓ | 0.37↓ | 41.15↓ | 90.73↑ | - | - | 0.34↑ | - | - |

Table 3: Metrics for the predicted scanpaths on the ZuCO dataset. To assess statistical reliability, we conducted paired t-tests ($p<0.05$) on metric values across folds, using ScanDL as the reference model. Significant differences are indicated with ↑/↓, where ↑ denotes an increase and ↓ a decrease relative to ScanDL (marked with *)

degrees of approximation to human performance.

Figure 2 displays average gaze features by part of speech. The results indicate that E-Z Reader shows the largest deviations from Human Baselines. While ScanDL and Eyettention often produce results closer to human baselines, they still fail to fully reproduce the characteristic differences in gaze patterns across grammatical categories.

Despite its shortcomings, E-Z reader shows good results for the NLD metric and the analysis of psycholinguistic predictors and parts of speech based on FPFC and SR gaze features.

### 3.5 Cross-Dataset Evaluation

The results are presented in Table 3. The metrics for Human Train-Val and Human Shuffled show greater differences compared to the CELER dataset, confirming our earlier observations. While the E-Z reader model outperforms Human Train-Val on the NLD metric, it demonstrates inferior performance on most other metrics. Random Saccades achieves better NLD scores than ScanDL and Eyettention, but underperforms on all other metrics. ScanDL and Eyettention show performance relative to Human Baselines similar to their results on the CELER dataset, but exhibit more noticeable shortcomings in NLD and PCC metrics. Random Fixations underperforms compared to ScanDL and Eyettention on most metrics but achieves better PCC scores. For PCC Base and PCC word, this results from limitations in ScanDL and Eyettention, while for PCC Sent it stems from using averaged human data for scanpath generation.

Figure 3 displays the PCC between word features and gaze characteristics. Among the evaluated models, ScanDL again shows the closest alignment with Human Baselines, though with more pronounced differences in some cases. The E-Z reader and Eyettention models demonstrate weaker performance in this analysis.

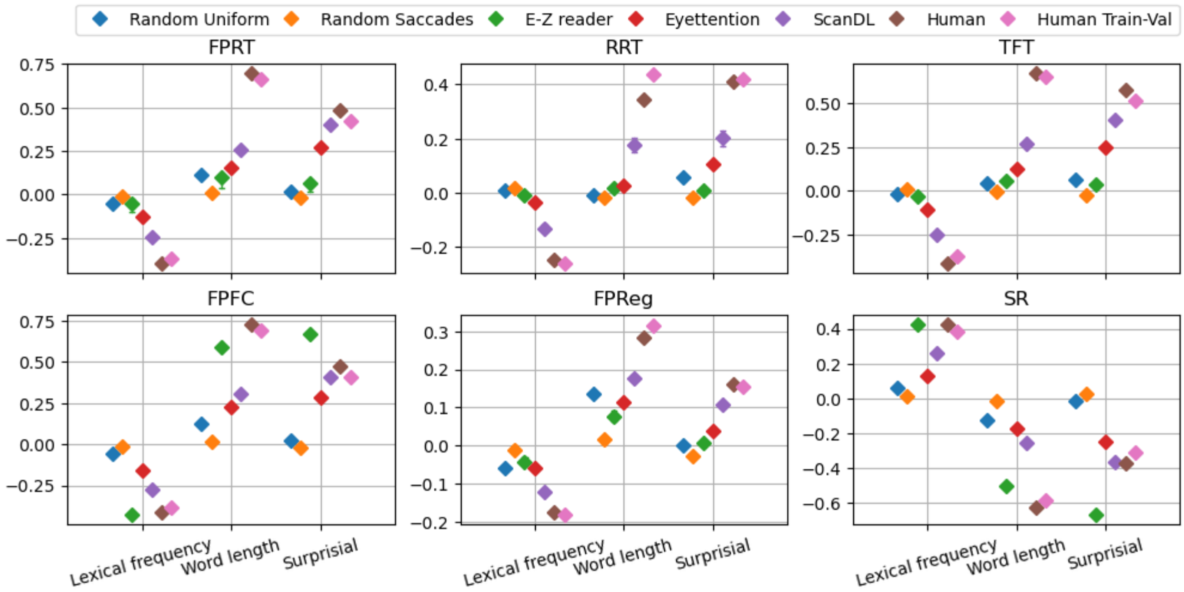

Figure 3: Pearson correlation coefficient between word features and gaze features on ZuCO dataset.

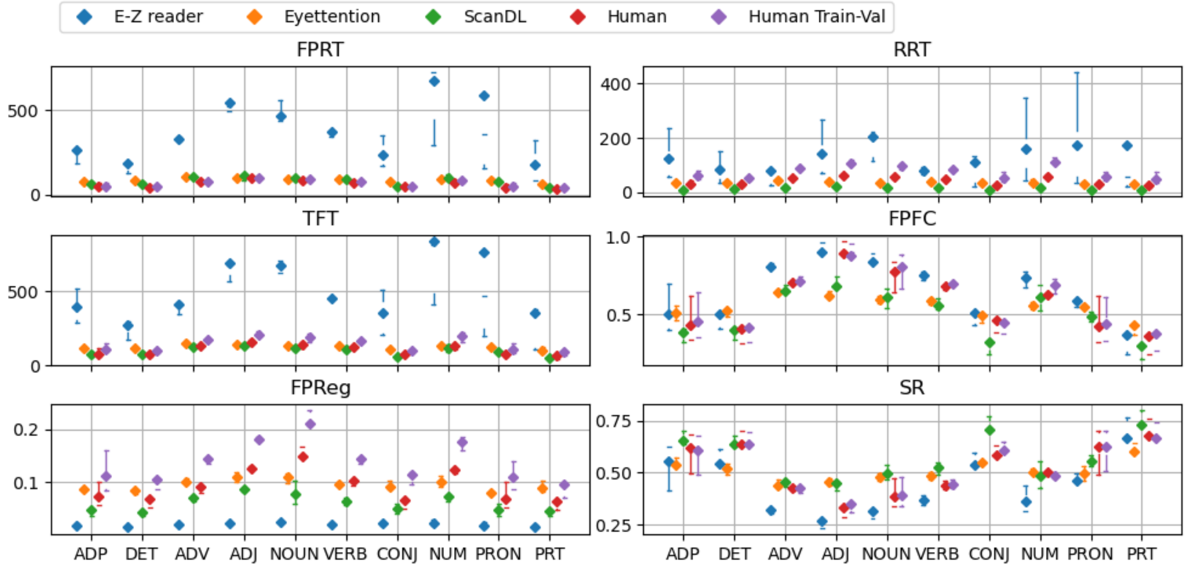

Figure 4: Mean gaze features with respect to POS tagging for ZuCO dataset.

Figure 4 presents average gaze features by part of speech. The deviations of E-Z reader have become much more substantial compared to the CELER dataset. Otherwise, the results remain comparable to those obtained for CELER.

As with the CELER dataset, E-Z reader shows good performance for the NLD metric and in analyzing psycholinguistic predictors and part-of-speech effects for the FPFC and SR gaze features. The model's primary limitation remains its inability to accurately reproduce regressions and fixation durations.

## 4 Conclusions

This study systematically evaluates contemporary approaches to scanpath generation and comprehensively compares their capabilities and limitations against authentic human gaze patterns. Our analysis of two distinct eye-tracking datasets reveals several important patterns that advance our understanding of current modeling paradigms. The ScanDL model for fixation sequence generation combined with the Fixation Duration Module proves to be the most robust among evaluated models, demonstrating consistent performance across multiple evaluation metrics while maintaining reasonable proximity to the Human Baseline. However, even this model shows notable deficiencies in reproducing certain aspects of natural gaze behavior, particularly when evaluated on a new dataset containing longer sentences of different domains. The primary limitation is insufficiently accurate reproduction of gaze features, especially in correlation metrics. The model also fails to fully capture part-of-speech-dependent variations in gaze patterns, particularly for re-reading time. While it performs well in assessing psycholinguistic predictors for Within-Dataset Evaluation, its performance degrades in Cross-Dataset Evaluation.

Eyettention represents an alternative approach that achieves competitive results. Although it matches ScanDL on several key metrics, it underperforms in overall evaluation. When evaluated with the Fixation Duration Module, Eyettention shows deterioration in gaze latency-based features compared to ScanDL. This outcome highlights the importance of fixation sequence quality for the Fixation Duration Module's performance. The E-Z Reader model, representing more traditional cognitive modeling approaches, demonstrates an interesting performance dichotomy. It performs similarly to ScanDL in Within-Dataset Evaluation of fixation sequences regarding similarity, word skipping, and fixation counts, and outperforms ScanDL in Cross-Dataset Evaluation. However, E-Z Reader shows significant difficulties with more complex gaze phenomena like regressions and fixation duration modeling. Initially, the E-Z Reader model accepts parameters derived empirically, which complicates its application for generating synthetic data. Consequently, the use of averaged and simulated parameters inevitably leads to a degradation in the quality of the generated gaze sequences. This pattern suggests that while symbolic cognitive models retain value for certain theoretical applications, they may require substantial enhancement to compete with data-driven approaches in practical implementations.

Comparative dataset analysis yields particularly valuable insights. The increased performance variability observed in the ZuCO dataset, with its more diverse text domains and longer sentences, underscores a critical challenge in gaze modeling - the need for systems capable of generalizing across different text types. This finding has important implications for practical applications, suggesting that future models will need to incorporate more diverse text domains. The persistent gap between model performance and human baselines across both datasets, particularly in correlation metrics, points to fundamental limitations in how current architectures represent the cognitive processes underlying reading.

ScanDL was chosen as the reference model since it achieves the strongest overall performance. The results show that ScanDL significantly outperforms other models and random baselines on most metrics. However, some metrics highlight weaknesses of the model: for example, gaze feature metrics aggregated at the sentence level are significantly worse than those of other models. Compared to the Human baseline, ScanDL generally performs significantly worse, indicating the need for further modifications of scanpath generation models.

Several promising directions for improving scanpath generation systems emerge from these results. Integrating multi-task learning objectives could help bridge the gap between gaze prediction and higher-level language understanding. Incorporating psycholinguistic and other features may enhance models' ability to capture nuances of reading behavior. Developing more comprehensive evaluation protocols, particularly those assessing models' capacity to reproduce known psycholinguistic phenomena across text domains, could drive significant improvements in model architectures and training approaches.

## Limitations

While this study provides a thorough examination of contemporary approaches to scanpath generation, several limitations must be acknowledged that both contextualize our findings and indicate important directions for future research. The exclusive focus on English-language datasets, while providing controlled comparison points, inevitably limits the generalizability of our conclusions. It is well-established that reading behaviors and eye movement patterns vary significantly across writing systems and linguistic structures: from alphabetic systems like English to logographic systems like Chinese or right-to-left scripts like Arabic. Future work should prioritize multilingual evaluation to determine whether the observed patterns hold across different languages and whether certain architectural approaches demonstrate particular advantages for specific writing systems.

The nature of our evaluation datasets, despite their careful construction, imposes certain limitations. Both CELER and ZuCO, despite their differences, consist predominantly of formal written language samples. This leaves open questions about how current models would perform with more informal or interactive text types, such as social media content or real-world reading scenarios where visual layout and task demands play important roles. The controlled laboratory conditions in which the eye-tracking data were collected may also limit applicability to more natural reading environments.

Our evaluation does not account for potential scaling effects, as we maintained fixed dataset sizes across experiments. Future work should examine how increasing training data volume impacts the reproduction of psycholinguistic gaze patterns. The question of which model characteristics influence the cognitively plausible reproduction of specific gaze properties remains open. A detailed analysis of this issue will facilitate a deeper understanding of gaze generation models and lay the theoretical groundwork for future models.

Our evaluation framework, while comprehensive, inevitably emphasizes certain aspects of gaze behavior over others. Current metrics focus primarily on low-level temporal and spatial patterns of eye movements. While this provides important quantitative benchmarks, they may not fully capture higher-level cognitive aspects of reading, such as comprehension monitoring or cross-sentence information integration. The development of more sophisticated evaluation protocols that account for these parameters remains an important challenge for the field.

## Acknowledgments

The study was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dd. 19.06.2025, IGK 000000C313925P4D0002).

## References

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd conference on computational natural language learning*, pages 302–312.

Lena Bolliger, David Reich, Patrick Haller, Deborah Jakobi, Paul Prasse, and Lena Jäger. 2023. ScanDL: A diffusion model for generating synthetic scanpaths on texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15513–15538, Singapore. Association for Computational Linguistics.

Lena S. Bolliger, David R. Reich, and Lena A. Jäger. 2025. Scandl 2.0: A generative model of eye movements in reading synthesizing scanpaths and fixation durations. *Proc. ACM Hum.-Comput. Interact.*, 9(3).

Charles Clifton, Fernanda Ferreira, John M. Henderson, Albrecht W. Inhoff, Simon P. Liversedge, Erik D. Reichle, and Elizabeth R. Schotter. 2016. Eye movements in reading and information processing: Keith rayner's 40year legacy. *Journal of Memory and Language*, 86:1–19.

da Silva Soares Jr, Raimundo, Oku, Amanda Yumi Ambriola, Barreto Cândida da Silva Ferreira, and Sato João Ricardo. 2023. Exploring the potential of eye tracking on personalized learning and real-time feedback in modern education. *Progress in Brain Research*, 282:49–70.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2023a. Pre-trained language models augmented with synthetic scanpaths for natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6500–6507, Singapore. Association for Computational Linguistics.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2024. Fine-tuning pre-trained language models with gaze supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–224, Bangkok, Thailand. Association for Computational Linguistics.

Shuwen Deng, David R. Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A. Jäger. 2023b. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. 7(ETRA).

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.

Ralf Engbert, André Longtin, and Reinhold Kliegl. 2002. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42(5):621–636.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. 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.

Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. 2010. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10, page 211–218, New York, NY, USA. Association for Computing Machinery.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia. Association for Computational Linguistics.

Samuel Kiegeland, David Robert Reich, Ryan Cotterell, Lena Ann Jäger, and Ethan Wilcox. 2024. The pupil becomes the master: Eye-tracking feedback for tuning LLMs. In *ICML 2024 Workshop on LLMs and Cognition*.

Keren Gruteke Klein, Shachar Frenkel, Omer Shubi, and Yevgeni Berzak. 2025. Eye tracking based cognitive evaluation of automatic readability assessment measures. *arXiv preprint arXiv:2502.11150*.

Matthias Kümmerer and Matthias Bethge. 2021. State-of-the-art in human scanpath prediction. *Preprint*, arXiv:2102.12239.

Anna K. Laurinavichyute, Irina A. Sekerina, Svetlana Alexeeva, Kristine Bagdasaryan, and Reinhold Kliegl. 2019. Russian sentence corpus: Benchmark measures of eye movements in reading in russian. *Behavior Research Methods*, 51(3):1161–1178.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Bai Li and Frank Rudzicz. 2021. TorontoCL at CMCL 2021 shared task: RoBERTa with multi-stage fine-tuning for eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 85–89, Online. Association for Computational Linguistics.

Angela Lopez-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. 2024. Seeing eye to ai: Human alignment via gaze-based response rewards for large language models. *Preprint*, arXiv:2410.01532.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Keith Rayner. 2009. The thirty fifth sir frederick bartlett lecture: Eye movements and attention during reading, scene perception, and visual search. quarterly journal of experimental psychology, 62, 1457-1506. *Quarterly journal of experimental psychology (2006)*, 62:1457–506.

Erik Reichle and Heather Sheridan. 2015. E-z reader: An overview of the model and two recent applications. *Oxford handbook of reading*, pages 277–292.

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

Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 6327–6341. Curran Associates, Inc.

Harshvardhan Srivastava. 2022. Poirot at CMCL 2022 shared task: Zero shot crosslingual eye-tracking data prediction using multilingual transformer models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 102–107, Dublin, Ireland. Association for Computational Linguistics.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

## A Gaze features nomenclature

Below is a list of gaze features that were used for the calculation:

FFD - first-fixation duration
SFD - single-fixation duration
FD - first duration
FPRT - first-pass reading time
FRT - first-reading time
TFT - total-fixation time
RRT - re-reading time
$\text{RPD}_{\text{inc}}$ - inclusive regression-path duration
$\text{RPD}_{\text{exc}}$ - exclusive regression-path duration
RBRT - right-bounded reading time
Fix - fixation (binary)
SR - skipping rate (binary)

FPF - first-pass fixation (binary)
RR - re-reading (binary)
FReg - first regression (binary)
FPReg - first-pass regression (binary)
$\text{TRC}_{\text{out}}$ - total count of outgoing regressions
$\text{TRC}_{\text{in}}$ - total count of incoming regressions
$\text{SL}_{\text{in}}$ - incoming saccade length
$\text{SL}_{\text{out}}$ - outgoing saccade length
FFC - first fixation count
FPFC - first-pass fixation count
TFC - total fixation count

## B Gaze features metrics

Table 4 presents MAE Word metrics for the CELER dataset for all gaze features.

Table 5 presents PCC Word metrics for the CELER dataset for all gaze features.

Table 6 presents MAE Word metrics for the ZUCO dataset for all gaze features.

Table 7 presents PCC Word metrics for the ZUCO dataset for all gaze features.

In the tables presented below, the Human column corresponds to the Human Train-Val baseline.

|  | Random<br>Uniform | Saccades | E-Z reader | Eyettention | ScanDL | Human |
|---|---|---|---|---|---|---|
| FD | 80.62 | 80.31 | 61.18 | 78.59 | 77.35 | 86.10 |
| FFC | 83.90 | 83.46 | 85.98 | 87.58 | 87.67 | 90.15 |
| FFD | 62.79 | 78.63 | 66.45 | 78.11 | 79.02 | 85.28 |
| FPF | 47.80 | 73.10 | 81.28 | 79.77 | 82.80 | 84.69 |
| FPFC | 70.14 | 82.23 | 86.97 | 86.36 | 87.87 | 89.50 |
| FPRT | 70.30 | 81.24 | 73.24 | 81.96 | 82.97 | 87.59 |
| FPReg | 87.21 | 85.77 | 86.87 | 88.89 | 88.94 | 91.71 |
| FRT | 82.57 | 82.22 | 68.60 | 82.16 | 81.79 | 88.11 |
| FReg | 73.84 | 84.38 | 83.65 | 85.52 | 87.18 | 89.52 |
| Fix | 77.27 | 76.38 | 79.88 | 82.68 | 82.76 | 86.85 |
| RBRT | 76.22 | 84.02 | 80.46 | 85.21 | 85.64 | 89.97 |
| RPD$_{exc}$ | 88.96 | 93.22 | 93.71 | 94.59 | 94.73 | 95.69 |
| RPD$_{inc}$ | 84.82 | 89.80 | 90.16 | 91.23 | 90.81 | 93.68 |
| RR | 60.48 | 76.55 | 65.43 | 79.30 | 73.94 | 82.87 |
| RRT | 69.87 | 83.57 | 76.75 | 85.86 | 82.71 | 88.82 |
| SFD | 63.81 | 78.82 | 59.87 | 77.01 | 77.63 | 82.45 |
| SL$_{in}$ | 42.47 | 88.14 | 90.94 | 88.89 | 90.79 | 90.79 |
| SL$_{out}$ | 78.15 | 92.29 | 92.05 | 93.47 | 92.86 | 93.84 |
| SR | 48.30 | 73.37 | 81.28 | 80.04 | 82.87 | 84.69 |
| TFC | 83.97 | 82.95 | 82.84 | 87.12 | 85.84 | 90.53 |
| TFT | 83.38 | 82.15 | 78.14 | 82.80 | 80.70 | 89.23 |
| TRC$_{in}$ | 75.85 | 88.54 | 87.37 | 88.49 | 89.10 | 91.31 |
| TRC$_{out}$ | 68.77 | 84.34 | 84.36 | 85.45 | 87.62 | 90.04 |

Table 4: MAE for the predicted gaze features on the CELER dataset.

|  | Random |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  | Uniform | Saccades | E-Z reader | Eyettention | ScanDL | Human |
| FD | 0.09 | -0.02 | 0.35 | 0.36 | 0.51 | 0.67 |
| FFC | 0.12 | -0.08 | 0.66 | 0.55 | 0.70 | 0.81 |
| FFD | 0.06 | -0.04 | 0.43 | 0.41 | 0.53 | 0.69 |
| FPF | -0.04 | -0.11 | 0.69 | 0.57 | 0.72 | 0.80 |
| FPFC | -0.00 | -0.13 | 0.68 | 0.58 | 0.71 | 0.82 |
| FPRT | 0.07 | -0.07 | 0.46 | 0.46 | 0.60 | 0.74 |
| FPReg | 0.57 | 0.05 | 0.13 | 0.52 | 0.41 | 0.71 |
| FRT | 0.11 | -0.04 | 0.38 | 0.43 | 0.59 | 0.73 |
| FReg | 0.36 | 0.04 | 0.13 | 0.43 | 0.41 | 0.65 |
| Fix | 0.11 | -0.09 | 0.68 | 0.50 | 0.69 | 0.80 |
| RBRT | 0.13 | -0.06 | 0.45 | 0.44 | 0.57 | 0.77 |
| RPD$_{exc}$ | 0.86 | 0.01 | 0.20 | 0.33 | 0.18 | 0.71 |
| RPD$_{inc}$ | 0.69 | -0.04 | 0.41 | 0.26 | 0.18 | 0.73 |
| RR | 0.30 | 0.05 | -0.09 | 0.30 | 0.23 | 0.60 |
| RRT | 0.34 | 0.03 | -0.03 | 0.31 | 0.24 | 0.65 |
| SFD | 0.04 | 0.02 | 0.23 | 0.22 | 0.43 | 0.51 |
| SL$_{in}$ | 0.16 | -0.10 | 0.48 | 0.35 | 0.54 | 0.61 |
| SL$_{out}$ | 0.51 | 0.57 | 0.52 | 0.75 | 0.63 | 0.73 |
| SR | -0.04 | -0.14 | 0.69 | 0.57 | 0.72 | 0.80 |
| TFC | 0.25 | -0.05 | 0.60 | 0.55 | 0.64 | 0.84 |
| TFT | 0.23 | -0.02 | 0.30 | 0.50 | 0.60 | 0.81 |
| TRC$_{in}$ | 0.44 | 0.11 | -0.12 | 0.32 | 0.21 | 0.57 |
| TRC$_{out}$ | 0.41 | 0.04 | 0.10 | 0.44 | 0.40 | 0.66 |

Table 5: PCC for the predicted gaze features on the CELER dataset.

| | Random | | | | | |
|---|---|---|---|---|---|---|
| | Uniform | Saccades | E-Z reader | Eyettention | ScanDL | Human |
| FD | 71.70 | 64.49 | -39.97 | 76.87 | 77.01 | 89.33 |
| FFC | 94.65 | 94.42 | 95.32 | 94.60 | 94.03 | 96.29 |
| FFD | 81.15 | 64.01 | -0.36 | 76.39 | 75.74 | 89.89 |
| FPF | 58.42 | 69.86 | 72.21 | 68.07 | 65.52 | 80.75 |
| FPFC | 86.90 | 89.29 | 90.99 | 89.83 | 88.96 | 93.11 |
| FPRT | 86.57 | 74.85 | 23.88 | 83.74 | 82.97 | 92.25 |
| FPReg | 87.99 | 84.79 | 88.71 | 87.14 | 87.92 | 87.78 |
| FRT | 89.44 | 85.04 | 37.55 | 90.45 | 90.31 | 95.13 |
| FReg | 72.68 | 82.25 | 84.71 | 82.12 | 84.65 | 83.20 |
| Fix | 72.15 | 72.85 | 75.46 | 69.92 | 66.91 | 82.21 |
| RBRT | 88.67 | 79.32 | 45.14 | 87.27 | 87.07 | 93.20 |
| RPD$_{exc}$ | 92.41 | 96.92 | 92.26 | 97.73 | 98.15 | 97.21 |
| RPD$_{inc}$ | 90.92 | 92.99 | 81.31 | 95.26 | 95.55 | 96.17 |
| RR | 66.62 | 77.39 | 74.47 | 77.79 | 77.64 | 77.43 |
| RRT | 78.73 | 90.02 | 73.65 | 93.68 | 94.10 | 92.12 |
| SFD | 82.60 | 65.93 | 8.15 | 77.08 | 77.69 | 89.19 |
| SL$_{in}$ | 86.80 | 97.06 | 97.17 | 96.50 | 96.76 | 96.83 |
| SL$_{out}$ | 90.50 | 97.48 | 97.36 | 97.16 | 97.22 | 97.53 |
| SR | 58.73 | 69.38 | 72.21 | 68.84 | 65.69 | 80.75 |
| TFC | 93.17 | 92.81 | 94.35 | 93.14 | 93.19 | 94.04 |
| TFT | 85.69 | 83.91 | 51.63 | 90.34 | 91.26 | 92.90 |
| TRC$_{in}$ | 86.78 | 92.22 | 93.67 | 92.06 | 93.73 | 92.21 |
| TRC$_{out}$ | 85.31 | 91.75 | 93.33 | 91.67 | 93.23 | 91.54 |

Table 6: MAE for the predicted gaze features on the ZuCO dataset.

|  | Random |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  | Uniform | Saccades | E-Z reader | Eyettention | ScanDL | Human |
| FD | 0.05 | 0.01 | 0.03 | 0.11 | 0.22 | 0.59 |
| FFC | 0.07 | 0.02 | 0.50 | 0.20 | 0.34 | 0.70 |
| FFD | 0.17 | 0.02 | 0.06 | -0.02 | 0.05 | 0.62 |
| FPF | 0.16 | 0.04 | 0.33 | 0.05 | 0.13 | 0.65 |
| FPFC | 0.15 | 0.03 | 0.45 | 0.13 | 0.23 | 0.68 |
| FPRT | 0.17 | 0.02 | 0.07 | 0.05 | 0.15 | 0.68 |
| FPReg | 0.45 | 0.03 | 0.05 | 0.22 | 0.17 | 0.58 |
| FRT | 0.06 | 0.00 | 0.04 | 0.16 | 0.28 | 0.68 |
| FReg | 0.18 | 0.02 | 0.04 | 0.15 | 0.16 | 0.53 |
| Fix | 0.06 | 0.02 | 0.42 | 0.15 | 0.28 | 0.65 |
| RBRT | 0.21 | 0.03 | 0.07 | 0.08 | 0.17 | 0.69 |
| RPD$_{exc}$ | 0.49 | 0.02 | 0.00 | 0.12 | 0.07 | 0.53 |
| RPD$_{inc}$ | 0.46 | 0.03 | 0.03 | 0.05 | 0.04 | 0.61 |
| RR | 0.16 | 0.03 | -0.02 | 0.15 | 0.18 | 0.47 |
| RRT | 0.18 | 0.01 | 0.00 | 0.20 | 0.22 | 0.55 |
| SFD | 0.14 | 0.02 | 0.04 | -0.0907 | -0.04 | 0.45 |
| SL$_{in}$ | 0.24 | 0.07 | 0.24 | 0.08 | 0.11 | 0.37 |
| SL$_{out}$ | 0.31 | 0.33 | 0.31 | 0.34 | 0.25 | 0.66 |
| SR | 0.16 | 0.05 | 0.33 | 0.06 | 0.13 | 0.65 |
| TFC | 0.15 | 0.02 | 0.49 | 0.24 | 0.40 | 0.72 |
| TFT | 0.13 | 0.01 | 0.03 | 0.25 | 0.38 | 0.72 |
| TRC$_{in}$ | 0.30 | 0.04 | -0.04 | 0.15 | 0.16 | 0.54 |
| TRC$_{out}$ | 0.21 | 0.01 | 0.04 | 0.16 | 0.15 | 0.56 |

Table 7: PCC for the predicted gaze features on the ZuCO dataset.